KTH Matematik

sf2935 Modern Methods of Statistical Learning:

Bayesian Learning, Bayesian Reasoning (& Bayesian Perceptrons)

Timo Koski

1 Introduction

1.1 Learning a Distribution

In these lecture notes we treat parametric models and model families for I.I.D. data and the estimation of them. We present the learning method of posterior probabilities and Bayes factors.

1.2 General Summary

Bayes' theorem or Bayes' rule for events, as known from first courses in probability and statistics, can be stated mathematically as the following equation

$$P(B|A) = \frac{P(B)\,P(A|B)}{P(A)}, \tag{1.1}$$

where A and B are events, $P(A) > 0$. For a very long time, in the era until WWII, this was known as 'inverse probability', as the roles of A and B are inverted in (1.1).

By learning from data one often means the process of inferring a general law or principle from the observations of particular instances. The general law is a piece of knowledge about the mechanism of nature that generates the data. The learning can be done by use of 'MODELS', which serve as the language in which the constraints predicated on the data can be described.

We shall in these lectures first talk about *parametric statistical models*, these are in a sense only vehicles of learning, not an end in themselves.

A formal Bayesian modelling articulates the information in a (training) data set with evidence other than that of the (training) set. It is thought that there is always such evidence or that there is no such thing as the 'right analysis' if there is none. The evidence is assessed by judgement and is expressed in probability theory terms:

(1) a probability distribution specifies the probability of any sequence of data conditional on certain parameters;

(2) a prior expresses uncertainty about the parameters.

When (1) is combined with the training sequence we get the *likelihood function* of the sequence. The likelihood function is combined with (2) via *Bayes'* rule to produce a *posterior distribution* for the parameters of the model and this is the output of the formal Bayesian analysis.

A probabilist/statistician has reminisced :

> Perhaps I should start from my undergraduate studies in mathematics at the University of Rome 'La Sapienza', where Bruno De Finetti, the father of neo-Bayesianism, had been Professor in Probability for 30 years. Although I never met him, his legacy was very much alive in the probability classes, where we were taught not only about theorems, but also about the philosophy and the interpretation of probability. Once you accept the subjective (Bayesian) interpretation of probability, you see that statistical inference is all about probabilistic modeling and computation of conditional probabilities, without forgetting about model comparison and testing the goodness-of-fit of the models.

One key word in the preceding is model comparison or **model choice**. It will be seen frequently in the sequel that Bayesian statistical methods will (at least asymptotically) coincide with the frequentist statistical procedures. However, this is not the case for model choice, for which there are few standard procedures outside Bayesian learning.

1.3 Bayesian Reasoning

Suppose

$$P(B|A) > P(B).$$

In words, if A is true, B becomes more likely. Now, using Bayes' rule we can invert the roles of A and B. It follows clearly that

$$\frac{P(A)\, P(B|A)}{P(B)} > P(A)$$

and Bayes' rule (1.1) gives

$$P(A \mid B) > P(A).$$

In words: if B is true, A becomes more likely. For example, if it is raining, the sidewalk is likely to be wet. Now, the sidewalk is wet, it is likely to be raining. In fact, that if $P(B|A) > P(B)$, then we have the statements

1. If A is true, then B becomes more likely.
2. If not B is true, then A becomes less likely.
3. If B is true, A becomes more likely.
4. If not A is true, then B becomes less likely.

These look like soft versions of Boolean logic.

2 Bayes' Theorem

2.1 Bayes' rule for Random Variables: Formal Expressions

The following form of Bayes' rule is contains more of structure than (2.2). If H_i, $i = 1, \ldots, n$ are an exhaustive partition of the outcome space, one has

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(A \mid H_j)}\,. \tag{2.2}$$

Here we talk about the *posterior probability* of the hypothesis H_i given the *evidence* A. $P(A|H_i)$ is the likelihood of H_i given evidence, and $P(H_i)$ is the *prior probability* of H_i. The fact that $P(A) = \sum_{j=1}^{n} P(H_j)P(A \mid H_j)$ is known as the *law of total probability.*

Consider two random variables X and Y. In principle, Bayes' rule applies to the events $A = \{X = x\}$ and $B = \{Y = y\}$. However, to remain useful, Bayes' rule is to be formulated in terms of the pertinent probability densities or probability mass functions.

Let us recall the joint (simultaneous) probability density of (X, Y), a continuous two dimensional R.V.

$$f_{X,Y}(x, y) \quad \text{s.t.} f_{X,Y}(x, y) \geq 0, \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x, y)dxdy = 1.$$

Then we define the **conditional probability density** of Y given $X = x$ as

$$f_{Y|X}(y|x) \overset{\text{def}}{=} \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad f_X(x) > 0 \tag{2.3}$$

where

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy. \tag{2.4}$$

The conditional probability density of X given $Y = y$, $f_{Y|X}(y|x)$, is defined analogously.

By applying the definition (2.3) a few times we obtain (2.7). The formulas in (2.5) - (2.6) are obtained in similar ways.

- For X continuous and Y discrete,

$$f_{X|Y}(x|y) = \frac{P(Y = y \mid X = x)\, f_X(x)}{P(Y = y)}. \tag{2.5}$$

- For X is discrete and Y continuous,

$$P_{X|Y}(X = x \mid y) = \frac{f_{Y|X}(y \mid x)\, P(X = x)}{f_Y(y)}. \tag{2.6}$$

- If both X and Y are continuous,

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}. \tag{2.7}$$

  It is useful to express the denominator using the law of total probability. For $f_Y(y)$, we compute the integral

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, f_X(x)\, dx. \tag{2.8}$$

The equations (2.5) - (2.8) are the formal tools of the Bayesian statistical inference and learning. Already this brief glance should hint that the essential computational issue here is the mixing integral (2.8).

2.2 Parametric Statistical Models

One type of learning that we will be concerned with, is inferring, or analysing data with a model family indexed by parameters. What does this mean?

Suppose that $f(x|\theta)$ is a probability density on R^n. $f(x|\theta)$ is a known mathematical expression in x and θ. θ is an unknown parameter of the density.

> The textbooks applying no Bayesian methosd write $f\left(x|\theta\right)$, as $f\left(x;\theta\right)$, where θ is treated as a mathematical parameter in an expression for a function.

Let us regard x is an observation of a random variable X, x can be of the form $x = (x_1, \ldots, x_n)$, or x can be continuous or discrete variate or a mixed case.

We assume

$$X \sim f\left(x \mid \theta\right)$$

Suppose that $f\left(x|\theta\right)$ is a probability density on R^n. $f\left(x|\theta\right)$ is a known mathematical expression in $x \in R^n$ and θ, which is a parameter that lies in some known set of parameters $\mathbf{\Theta}$.

> [**Parametric statistical model**] $f\left(x|\theta\right)$ is a probabilistic mechanism of generating data, characterizes the behaviour of future observations conditional on θ. The parameter θ is unknown, but is assumed to lie in some known parameter space $\mathbf{\Theta}$. The expression $P(x \mid \theta)$ is for a discrete x.
>
> [**Statistical model family**]
>
> $$\mathcal{M} = \{f(x|\theta), \theta \in \mathbf{\Theta}\}$$
>
> or
>
> $$\mathcal{M} = \{P(X = x|\theta), \theta \in \mathbf{\Theta}\}.$$

Some examples illustrate these concepts.

Example 2.1 $\theta = (\mu, \sigma) \in \mathbf{\Theta} = R \times (0, \infty)$.

$$f\left(x|\theta\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{\sigma^2}(x-\mu)^2}, -\infty < x < \infty.$$

We say that x is an observation from the normal distribution $N\left(\mu, \sigma^2\right)$.

■

Example 2.2 Consider the r.v. X which is binary, its set of values is coded as $\{0,1\}$, $\boldsymbol{\Theta} = \{\theta | 0 < \theta < 1\}$ and

$$P(X = x \mid \theta) = \theta^x \cdot (1-\theta)^{1-x}. \tag{2.9}$$

Then we say X is distributed according to the Bernoulli distribution with the parameter θ.

$$X \mid \Theta = \theta \sim Be(\theta).$$

Example 2.3 Consider the r.v. X with values x in $\{0,1,2,\ldots,n\}$, $\boldsymbol{\Theta} = \{\theta | 0 < \theta < 1\}$.

$$P(x|\theta) = P(X = x|\theta) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}. \tag{2.10}$$

Then we say X is distributed according to the Binomial distribution with the parameters n and θ.

$$X \mid \Theta = \theta \sim \mathrm{Bin}(n, \theta).$$

Example 2.4 The r.v. X has the non-negative integers as values $x \in \{0,1,\ldots\}$, and $\boldsymbol{\Theta} = \{\theta \mid \theta > 0\}$ and

$$P(x|\theta) = P(X = x \mid \theta) = e^{-\theta}\frac{\theta^x}{x!} \tag{2.11}$$

Then we say that X is distributed according to the Poisson distribution with the parameter θ.

$$X \mid \Theta = \theta \sim \mathrm{Po}(\theta),$$

■

2.3 Parametric Statistical Models and Priors

> **Bayesian standpoint regards an unknown parameter as a random variable**. Uncertainty about the unknown θ is modelled by a probability distribution $\pi(\theta)$, and then $\pi(\theta \mid x)$ expresses the uncertainty about the unknown θ after the observation of x.

■

Formally, θ is an outcome of a r.v. Θ, and lies in a vector space of finite dimension $\boldsymbol{\Theta}$. For simplicity of writing $\boldsymbol{\Theta}$ is assumed to be non denumerable. We take

$$X \mid \Theta = \theta \sim f\left(x|\theta\right),$$

where $f\left(x|\theta\right)$ is a probability density conditionally on $\Theta = \theta$. The main goal is to find $\pi\left(\theta \mid x\right)$, this is where Bayes' formula comes in.

[**Bayesian statistical model family**]

$$\mathcal{M} = \{f(x|\theta), \pi(\theta), \theta \in \boldsymbol{\Theta}\}.$$

Suppose now that X is continuous. From (2.7) we get the **posterior probability density** of Θ given $X = x$ as

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{f(x)}. \tag{2.12}$$

Here $\pi(\theta)$ is the **prior density** of Θ. In (2.12)

$$f_X(x) = \int_{-\infty}^{\infty} f(x \mid \theta)\,\pi(\theta)\,d\theta. \tag{2.13}$$

In view of (2.6) we have for discrete X we have the **posterior probability density**

$$\pi(\theta|x) = \frac{P(X = x|\theta)\,\pi(\theta)}{f(x)}, \tag{2.14}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} P(X = x|\theta)\,\pi(\theta)\,d\theta. \tag{2.15}$$

Example 2.5 From Example 2.4

$$X \mid \Theta = \theta \sim \mathrm{Po}(\theta).$$

Let the parameter Θ be exponentially distributed

$$\pi(\theta) = \begin{cases} \lambda e^{-\lambda\theta}, & \theta > 0 \\ 0, & \theta < 0. \end{cases}$$

In this situation λ is called a **hyperparameter**. Then we get in (2.5)

$$f(x \mid \theta) = \begin{cases} \frac{e^{-\theta}\frac{\theta^x}{x!}\lambda e^{-\lambda\theta}}{P(X=x)}, & \theta > 0 \\ 0, & \theta < 0, \end{cases} \tag{2.16}$$

where

$$\begin{aligned} P(X = x) &= \int_0^{+\infty} e^{-\theta}\frac{\theta^x}{x!}\lambda e^{-\lambda\theta} d\theta \\ &= \frac{\lambda}{x!}\int_0^{+\infty} \theta^x e^{-\theta(1+\lambda)} d\theta. \end{aligned}$$

Here

$$\int_0^{+\infty} \theta^x e^{-\theta(1+\lambda)} d\theta = x!\frac{1}{(1+\lambda)^{x+1}}.$$

When we insert in (2.16) we get

$$\pi(\theta \mid x) = \begin{cases} (1+\lambda)^{x+1} e^{-\theta(1+\lambda)}\frac{\theta^x}{x!} & \theta > 0 \\ 0 & \theta < 0. \end{cases} \tag{2.17}$$

This is the probability density function of $\Gamma(x+1, 1/(1+\lambda))$, a Gamma distribution with parameters $x+1$ and $1/(1+\lambda)$. Hence

$$\Theta \mid X = x \sim \Gamma(x+1, 1/(1+\lambda))$$

Let us note that the prior distribution here, the exponential distribution, can be seen as $\Gamma(1, 1/\lambda)$. We have here an example of a *conjugate prior* or a *prior closed under sampling* for $f(x \mid \theta)$: the posterior and prior are members of the same family of distributions.

■

Example 2.6 $X_i \mid \Theta = \theta \sim \mathrm{N}(\theta, \sigma_0^2)$, $\Theta \sim \mathrm{N}(\mu, s^2)$. $x^{(n)} = (x_1, \ldots, x_n)$ an I.I.D. sample of X_i, respectively, $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

$$\Theta \mid (X_1, \ldots, X_n) \sim \mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right)$$

i.e., $\pi\left(\theta|x^{(n)}\right)$ is the density of this normal distribution. Here μ and s^2 are called *hyperparameters*.

■

3 Bayesian Learning/Inference of Parameters

3.1 Likelihood, Conjugate Family of Priors, MLE, MAP

The expression $f(x \mid \theta)$ regarded as a function of θ is known as the *likelihood function*

$$L_x(\theta) = f(x \mid \theta).$$

The likelihood function $L_x(\theta)$ thus compares the plausibilities of different parameter values for given x.

$$l_x(\theta) = -\log L_x(\theta)$$

is called $(-1\times)$ the log likelihood function. By Bayes formula we get the schematic formulation

posterior $\propto$ likelihood $\times$ prior

Any function $\pi(\theta)$ such that

$$\pi(\theta) \geq 0, \quad \theta \in \mathbf{\Theta}$$

and

$$\int_{\mathbf{\Theta}} \pi(\theta)\, d\theta = 1,$$

might serve as a prior probability density function. But even functions with the properties $\pi(\theta) \geq 0$, and

$$\int_{\Theta} \pi(\theta)\, d\theta = \infty,$$

are also invoked as priors, and are called *improper priors.*

The choice of $\pi(\theta)$ should, however, reflect judgements and understanding about the domain to be studied, prior to seeing data, reflected in the additional mathematical properties of the function $\pi(\theta)$. Later in section 8 there will be more about the selection of prior, but, for ease of presentation, we give one rule quantifying prior information right now.

Definition 3.1 Conjugate Family of Priors
A family $\mathcal{F}$ of probability distributions on $\boldsymbol{\Theta}$ *is said to be* **conjugate** *or* **closed under sampling** *for a likelihood function*

$$L_x(\theta) = f(x \mid \theta),$$

if for every $\pi(\theta) \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

■

We have seen examples of conjugate priors in Examples 2.5 and 2.6. **A more complete treatment is found in the sf2935 lecture/slides on exponential families and exponential deep learning**. An intuitive way of understanding conjugate priors is that with conjugate priors the prior knowledge can be translated into equivalent sample information. E.g., in Example 2.6,

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right).$$

Definition 3.2 The maximum likelihood estimate/estimator The maximum likelihood estimate, MLE, $\widehat{\theta}_{\mathrm{ML}}$ of θ, is defined by

$$\begin{aligned}\widehat{\theta}_{\mathrm{ML}} &= \mathrm{argmax}_{\theta\in\boldsymbol{\Theta}} f(x \mid \theta) \\ &= \mathrm{argmin}_{\theta\in\boldsymbol{\Theta}} l_x(\theta).\end{aligned}$$

When we regard MLE as a function of X, i.e., $\widehat{\theta}_{\mathrm{ML}}(X)$, then we talk about a *Maximum Likelihood Estimator.*

■

$\widehat{\theta}_{\mathrm{ML}}$ is a parameter value that gives the observed x the highest possible probability.

Definition 3.3 The maximum a posterior estimate/estimator The maximum a posterior estimate, MAP, $\widehat{\theta}_{\mathrm{MAP}}$ of θ is defined by

$$\widehat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\theta\in\boldsymbol{\Theta}} \pi(\theta \mid x)$$

When we regard MAP as a function of X, i.e., $\widehat{\theta}_{\mathrm{MAP}}(X)$, then we talk about a *Maximum A Posterior Estimator.*

■

We have in fact

$$\widehat{\theta}_{\text{MAP}} = \text{argmax}_{\theta \in \boldsymbol{\Theta}} f\left(x \mid \theta\right) \pi\left(\theta\right).$$

Example 3.1 $x_1 \ldots x_n$ are independent samples of $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma) \in \boldsymbol{\Theta} = \{\mu \in R, \sigma > 0\}$. Then the likelihood function is obtained by multiplication

$$f(x_1 \mid \theta) \cdots f(x_n \mid \theta) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu))^2}$$

If we set the partial derivatives of $l_x\left(\mu, \sigma^2\right)$ w.r.t. μ and σ^2 to zero and solve the system of equations we get

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j, s^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

as the maximum likelihood estimates of μ and σ^2, respectively.

■

3.2 MAP for multivariate regression & MAP for logistic regression

3.2.1 MAP for multivariate regression

$$Y = \boldsymbol{\beta}^T \mathbf{X} + \epsilon$$

$$Y \sim N_n\left(\boldsymbol{\beta}^T \mathbf{X}, \sigma^2 \mathbf{I}\right)$$

$$f\left(y \mid \boldsymbol{\beta}\right) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2}(y - \boldsymbol{\beta}^T \mathbf{X})^T (y - \boldsymbol{\beta}^T \mathbf{X})}$$

and maximum likelihood estimate

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

is nothing but the least squares estimate.

Take $\boldsymbol{\beta} \sim N_{p+1}(\mu, \psi^2 \mathbf{I})$. Then it follows that

$$\widehat{\boldsymbol{\beta}}_{MAP} = (\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{\psi^2} \mathbf{I})^{-1} \mathbf{X}^T Y$$

For $\psi \to \infty$, the MAP -estimate becomes the ML-estimate.

3.2.2 MAP for logistic regression

Logistic regression is a model for a binary r.v. Y with values y in $\{-1, +1\}$. It says that

$$P(Y = y \mid \mathbf{X}_1 \ldots, \mathbf{X}_l) = \sigma\left(y \boldsymbol{\beta}^T \mathbf{X}\right),$$

where $\sigma(t) = 1/(1 + e^{-t})$ and $\mathbf{X}$ is a vector. We have a training set

$$\mathcal{S} = \left\{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_l, y_l)\right\},$$

where $y_i = \pm 1$. We choose a prior distribution $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma)$

The logarithm of the posterior probability of the parameters $\boldsymbol{\beta}$ is

$$\ln \pi(\boldsymbol{\beta} \mid \mathbf{X}) = \sum_{i=1}^{l} \ln P(y_i \mid \mathbf{X}_i) - \frac{1}{2} \boldsymbol{\beta}^T \Sigma^{-1} \boldsymbol{\beta} + c$$

Thereby we need only to deal with (

$$\mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{l} \log \sigma\left(y_i \boldsymbol{\beta}^T \mathbf{X}_i\right) - \frac{1}{2} \boldsymbol{\beta}^T \Sigma^{-1} \boldsymbol{\beta}$$

to maximize the posterior density, i.e., to find the MAP estimate of $\boldsymbol{\beta}$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{l} y_i \mathbf{X}_i \sigma(-y_i \boldsymbol{\beta}^T \mathbf{X}_i) - \Sigma^{-1} \boldsymbol{\beta}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{l} y_i \mathbf{X}_i \sigma(-y_i \boldsymbol{\beta}^T \mathbf{X}_i) - \Sigma^{-1} \boldsymbol{\beta}$$

This follows by

$$\frac{d}{dt} \log \sigma(t) = \frac{e^{-t}}{1 + e^{-t}} = \frac{1}{1 + e^{t}} = \sigma(-t)$$

and by the derivative of a quadratic form (differential calculus for matrices)

$$\frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{2} \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} = \Sigma \boldsymbol{\beta}$$

and by

$$\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\beta}^T \mathbf{X} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{X}^T \boldsymbol{\beta} = \mathbf{X}.$$

Thus if we set the gradient $\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{F}(\boldsymbol{\beta}) = \mathbf{0}$ (= a column vector of zeros) we get

$$\sum_{i=1}^{l} y_i \mathbf{X}_i \sigma(-y_i \boldsymbol{\beta}^T \mathbf{X}_i) = \Sigma^{-1}\boldsymbol{\beta}$$

i.e. the MAP $\widehat{\boldsymbol{\beta}}$ will satisfy

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{l} y_i \sigma(-y_i \widehat{\boldsymbol{\beta}}^T \mathbf{X}_i)\Sigma \mathbf{X}_i$$

This system of equations is solved by numerical algorithms. When we insert $\widehat{\boldsymbol{\beta}}$ back to $P(y \mid \mathbf{X})$ (see Lecture 5 on Logistic Regression) we have

$$\widehat{P}(y \mid \mathbf{X}) = \sigma\left(y\widehat{\boldsymbol{\beta}}^T \mathbf{X}\right)$$

$$= \sigma\left(y \sum_{i=1}^{l} y_i \sigma(-y_i \widehat{\boldsymbol{\beta}}^T \mathbf{X}_i)\mathbf{X}_i^T \Sigma \mathbf{X}\right).$$

Here we recognize what is in machine learning and statistics called a 'valid kernel'

$$K(\mathbf{X}_i, \mathbf{X}) := \mathbf{X}_i^T \Sigma \mathbf{X}.$$

4 Issues of Bayesian Learning (1)

4.1 General Aspects of Bayesian Inference, McMC, ABC

Bayesian learning uses probability as tool for all parts of data analysis. This is part of what is being called coherence.

- Inference is based on the observed x, not on an unobserved sample space, c.f. unbiased and effective estimators.

- $\pi(\theta|x)$ is the only quantity evaluated for inference about θ.

However, evaluation of $\pi(\theta|x)$ is not in general possible by explicit means of integral calculus, due to the difficulties of computing the denominator.

This is where *Markov chain Monte Carlo* (McMC) has contributed to the widespread popularity of Bayesian learning methods. BUGS, useful software for McMC and practical Bayesian analysis is described in Lunn et.al. (2013).

We are in these lectures dealing with models, where an analytical formula for the likelihood function can typically be derived. For complex models, an analytical formula might be elusive or the likelihood function might be computationally very costly to evaluate.

Approximate Bayesian computation (ABC) methods bypass by computational means the evaluation of the likelihood function. Therefore, the ABC methods widen the realm of models for which statistical learning can be considered, c.f. Sunnåker e.t.al. (2013).

4.2 Estimators

This example is due to Dennis Lindley.

$$x = (x_1, \ldots, x_n), \theta = (\theta_1, \ldots, \theta_n)$$

$$X_i | \Theta_i = \theta_i \sim N(\theta_i, 1), \qquad \text{I.I.D.}$$

$$\Theta_i \sim N(0, 1), \qquad \text{I.I.D.}$$

One can check that $f_X(X_i)$ is $N(0,2)$, and $\Theta_i | X_i = x_i \sim N\left(\frac{x_i}{2}, \frac{1}{2}\right)$. We use $\widehat{\theta}_i = \frac{x_i}{2}$ as a *point estimator* of θ_i. Then

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_i^{\,2} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^2}{4} \to \frac{1}{2}$$

by the law of large numbers as $n \to \infty$, but

$$\frac{1}{n}\sum_{i=1}^{n} \theta_i^2 \to 1$$

by the law of large numbers as $n \to \infty$.

Bayesian analysis works with the distribution of $\frac{1}{n}\sum_{i=1}^{n} \theta_i^2$ given x. The expectation w.r.t. this distribution is

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \theta_i^2 \mid x\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^2}{4} + \frac{1}{2}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i^2}{4}+\frac{1}{2}\to 1.$$

by law of large numbers as $n\to\infty$. Hence the posterior distribution has the right properties, not the estimator.

4.3 Confidence intervals

$X_i \mid \Theta=\theta \sim \mathrm{N}\left(\theta,\sigma_0^2\right)$, $\Theta\sim \mathrm{N}\left(\mu,s^2\right)$. $x^{(n)}=(x_1,\ldots,x_n)$ an I.I.D. sample of X_i (conditionally on $\Theta=\theta$, as explained later), respectively, and $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

$$\Theta \mid x^{(n)} \sim \mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)$$

i.e., $\pi\left(\theta|x^{(n)}\right)$ is the density of this normal distribution. Here μ and s^2 are again hyperparameters.

- $P\left(a(x)\leq\Theta\leq b(x)\right)=\int_{a(x)}^{b(x)}f\left(\theta|x\right)d\theta$
- $$\underbrace{P\left(a(x)\leq\Theta\leq b(x)\right)}_{\text{This is a probability, not a degree of confidence}}$$

Let $s\to\infty$ (the prior becomes improper). Then

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)\to\mathrm{N}\left(\overline{x},\frac{\sigma_0^2}{n}\right).$$

But in this limit the Bayesian confidence interval, e.g., $P\left(a(x)\leq\Theta\leq b(x)\right)=0.95$ becomes the familiar confidence interval for the mean of a normal distribution with known variance, or $\overline{x}\pm\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}}$. This confidence interval is taught as having a confidence level equalling the the probability of the interval $[\overline{x}-\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}},\overline{x}+\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}})]$ covering the unknown θ.

But from the point of the limit above the common erroneous but natural interpretation of

$$P\left(\overline{x}-\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}}\leq\Theta\leq\overline{x}+\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}}\right)$$

as the probability of Θ lying in $[a(x),b(x)]$, looks like being a correct one !

My friend and colleague Gunnar Englund, retired senior lecturer of mathematical statistics at KTH, has recently sent this to Editors of *Scientific American*:

> As a statistician I regrettably noted that again (for approximately the umpteenth time) an otherwise excellent article is stained by an incorrect interpretation of a so-called hypothesis testing. In "Neutrinos at the ends of the earth" Francis Halzen states *the probability is greater than 99.9999 percent that these events truely come from deep space.* This is again a case of *fallacy of the transposed conditional*, i.e. the claim by Halzen is that the probability of the null hypothesis (the events do not come from deep space) given the data is 0.0001 percent, while the correct interpretation is that this is the probability of the data given the null hypothesis. In fact, using a Bayesian approach a statement like Halzen's could be made, but that would presuppose an (arbitrary) a priori probability that the theory is correct.

5 Probabilistic Models with Conditional Independence

Next we study a family of models that explicitly considers a sequence of data. Here we introduce the *conditional independence* as a part of the model. We give a somewhat generic definition: X and Y are conditionally independent given Z if and only if

$$P((X,Y) \mid Z) = P(X \mid Z)P(Y \mid Z).$$

It follows from this definition that if X and Y are conditionally independent given Z, then

$$P(Y \mid X, Z) = P(Y \mid Z).$$

We shall consider two examples of how and why Bayesian thinking uses this.

5.1 Modelling and Learning by Tosses of a Thumbtack

5.1.1 The Model Family

The mathematics involved here is found in greater detail in (v. Mises and Geiringer 1964) or in (Heckerman 2008).

We consider a sequence of flips of a thumbtack. If we throw a thumbtack in the air, it will come to rest either on its point (0) or on its head (1). Suppose we flip the thumbtack n times (fixing n in advance), making sure that that the physical properties of the thumbtack and the conditions under which it is flipped remain stable over time. We let $\mathbf{x}$ denote the sequence of outcomes of the flips, $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$, $x_{i_l} \in \{0, 1\}$. Let now Θ be a random variable (quantity), whose values are numbers, denoted by θ, between zero and one, $0 \leq \theta \leq 1$. These values θ correspond to the possible values of the *chance of obtaining heads* in tossing thumbtack .

MODEL FAMILY:

CONDITIONED ON $\Theta = \theta$, THE SYMBOLS IN $\mathbf{x}$ ARE INDEPENDENT.

Or more completely:

> $\mathcal{M}$ = the symbols in $\mathbf{x}$ are outcomes of conditionally independent Bernoulli random variables with the parameter $\Theta = \theta$, $\boldsymbol{\Theta} = \{0 \leq \theta \leq 1\}$.

Note that this is different from what you may have gathered from earlier teaching, where θ is fixed but unknown, has no probability distribution and $X_1 \ldots X_n$ are independent random variables.

Hence, in view of (2.9), for any a model in the family the observed $\mathbf{x}$ is given the probability assignment

$$P(\mathbf{x} \mid \Theta = \theta) = \prod_{l=1}^{n} \theta^{x_{i_l}} \cdot (1-\theta)^{1-x_{i_l}} =$$

$$\theta^{\sum_{l=1}^{n} x_{i_l}} \cdot (1-\theta)^{n-\sum_{l=1}^{n} x_{i_l}} = \theta^{k} \cdot (1-\theta)^{n-k}, \tag{5.1}$$

if $\sum_{l=1}^{n} x_{i_l} = k$. In the thumbtack example we can understand learning as follows. We have observed n outcomes of flips of a thumbtack $\mathbf{x}$ and wish to determine which of the models in the family that best describes **this** set of flips.

5.1.2 The Posterior Density

To progress with this we express our uncertainty about $\theta \in \boldsymbol{\Theta}$ using a prior $\pi(\theta)$, By (2.14) we get the posterior

$$\pi(\theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta = \theta) \cdot \pi(\theta)}{\int_0^1 P(\mathbf{x} \mid \Theta = \theta) \cdot \pi(\theta)\, d\theta}, 0 \leq \theta \leq 1 \tag{5.2}$$

and zero elsewhere.

The posterior $\pi\left(\theta \mid \mathbf{x}\right)$ expresses our updated belief in the statement that a certain θ is the true chance of obtaining heads given that we have observed $\mathbf{x}$.

One way to get further from here is to use a conjugate prior $\pi\left(\theta\right)$. This is not the only useful choice, but at least analytically quite advantageous. Let us consider the *uniform prior* (which is a Beta probability density, see appendix 11.3) given by

$$\pi\left(\theta\right) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The uniform prior is often interpreted as a representation of complete ignorance about θ.

By an insertion we can calculate

$$\int_0^1 P\left(\mathbf{x} \mid \Theta = \theta\right) \cdot \pi\left(\theta\right) d\theta = \int_0^1 \theta^k \cdot \left(1-\theta\right)^{n-k} d\theta = \frac{k!(n-k)!}{(n+1)!}$$

by the *Beta integral*, see (A.10). Then we have

$$\pi\left(\theta \mid \mathbf{x}\right) = \begin{cases} \frac{(n+1)!}{k!(n-k)!} \cdot \theta^k \left(1-\theta\right)^{n-k} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{5.3}$$

This is a *Beta density*, see appendix 11.3. Hence the family of Beta densities is conjugate priors to the likelihood in (5.1).

5.1.3 $\pi\left(\theta \mid \mathbf{x}\right)$ evolves for $\mathbf{x} = 1101101110$

Suppose we see a success ($1 \leftrightarrow S$). What is posteriorin $\pi(\theta|S)$? We write

$$\begin{array}{c|cc} x & \text{F} & \text{S} \\ p_X(x|\theta) & 1-\theta & \theta \end{array}$$

so that likelihood × prior becomes

$$p_X(S \mid \theta)\pi(\theta) = \theta \cdot 1.$$

Then the marginal data likelihood is

$$m(S) = \int_0^1 p_X(S \mid \theta)\pi(\theta) d\theta = \int_0^1 \theta d\theta = \left[\frac{\theta^2}{2}\right]_0^1 = \frac{1}{2}.$$

Hence Bayes' theorem gives

$$\pi(\theta \mid S) = \begin{cases} 2\theta & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The Figure (from Berry, Donald A: *Bayesian clinical trials*, **Nature reviews Drug discovery**,5, pp. 27-36, 2006) is in twelve fields. In the topmost left field one plots the prior $\pi(\theta)$ and in the following (After S) to the right $\pi(\theta \mid S) = 2\theta$ is plotted.

Suppose we get again (S). What will be $\pi(\theta|SS)$? We assume conditional independence. Then we get

$$p_X(SS \mid \theta) = p_X(S \mid \theta) \cdot p_X(S \mid \theta) = \theta^2$$

and

$$m(SS) = \int_0^1 p_X(SS \mid \theta)\pi(\theta)d\theta = \int_0^1 \theta^2 d\theta = \left[\frac{\theta^3}{3}\right]_0^1 = \frac{1}{3}$$

And Bayes' theorem gives

$$\pi(\theta \mid SS) = \begin{cases} 3\theta^2 & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is the curve in the second field (Another S). We have a posterior skewed to the right. Next outcome is F.

$$p_X(SSF \mid \theta) = p_X(S \mid \theta) \cdot p_X(S \mid \theta) \cdot \underbrace{p_X(F \mid \theta)}_{=(1-\theta)} = \theta^2(1-\theta).$$

Then

$$m(SSF) = \int_0^1 p_X(SSF \mid \theta)\pi(\theta)d\theta = \int_0^1 \theta^2(1-\theta)d\theta =$$

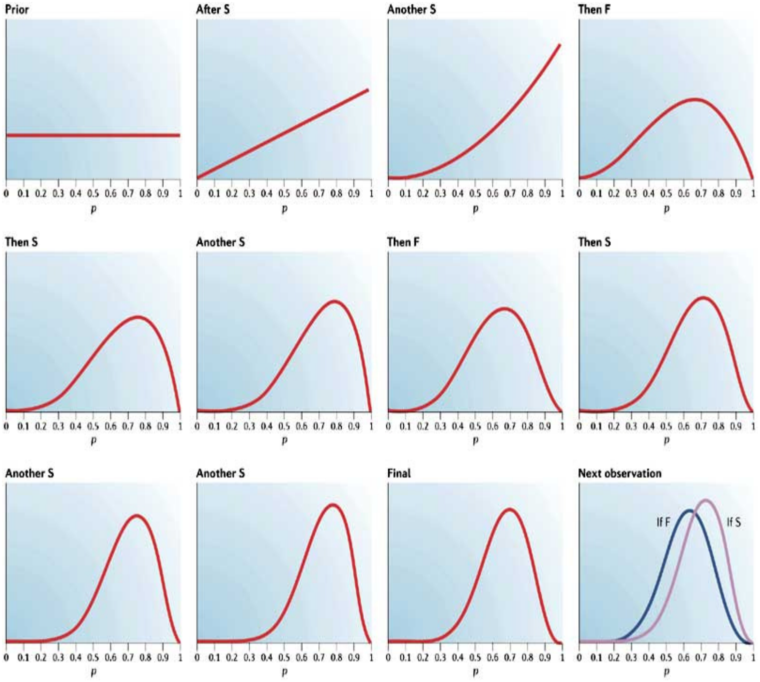
$$= \int_0^1 \theta^2 d\theta - \int_0^1 \theta^3 d\theta = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

and by Bayes' theorem we find

$$\pi(\theta \mid SSF) = \begin{cases} 12\theta^2(1-\theta) & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

We can clearly continue like this to find $\pi\left(\theta \mid \mathbf{x}\right)$ for $\mathbf{x} = 1101101110$ (coded as $SSFSSFSSSF$).

The figure shows how $\pi\left(\theta \mid \mathbf{x}\right)$ evolves for read from left to right and starting from the flat prior.



Copyright © 2005 Nature Publishing Group
Nature Reviews | Drug Discovery

5.1.4 The Maximum Likelihood Estimate

To understand better the alluded properties of $f_{\Theta|\mathbf{x}}\left(\theta \mid \mathbf{x}\right)$ we consider the maximum likelihood estimate $\widehat{\theta}_{ML}$ of θ from

$$\widehat{\theta}_{ML} = \mathrm{argmax}_{0\leq\theta\leq 1} P\left(\mathbf{x} \mid \Theta = \theta\right) = \mathrm{argmax}_{0\leq\theta\leq 1} \theta^{k} \cdot \left(1-\theta\right)^{n-k}.$$

The rationale for this is that we try to find the model within the family that gives the (training) sequence $\mathbf{x}$ the highest possible probability. The likelihood function is

$$L\left(\theta\right) = P\left(\mathbf{x} \mid \Theta = \theta\right).$$

The likelihood function $L\left(\theta\right)$, as said above, compares the plausibilities of different models for given $\mathbf{x}$. A maximization of the likelihood function gives

$$\widehat{\theta}_{ML} = \frac{k}{n}. \qquad (5.4)$$

5.2 The Bernstein – von Mises theorem

According to the Bernstein - von Mises theorem, the asymptotic distribution of the MAP estimator depends for large data sets on the *Fisher information* and not on the prior. We need the definition of the Fisher information, denoted by $I(\theta)$. We assume $\mathbf{\Theta}$ is one dimensional. The integral

$$I(\theta) \overset{\text{def}}{=} -\int_{R^n} \frac{d^2}{d\theta^2} \log f(x|\theta)\, f(x \mid \theta)\, dx \tag{5.5}$$

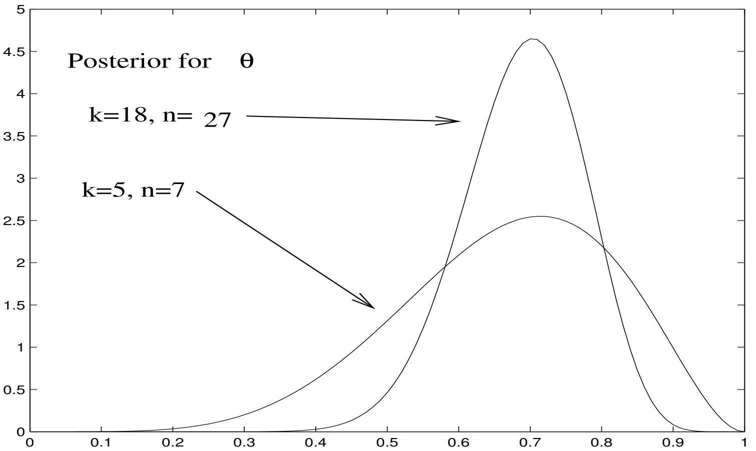
is called the Fisher information. We can establish for the thumbtack toss model family in this section that

$$I(\theta) = \frac{1}{\theta \cdot (1-\theta)}.$$

By a Taylor expansion of $\log P(\mathbf{x} \mid \Theta = \theta)$, see section 14 for further heuristic details around $\widehat{\theta}_{ML}$ we obtain something like the statement of the Bernstein – von Mises theorem, or

$$\pi(\theta \mid \mathbf{x}) \approx e^{-\frac{1}{2} n I\left(\widehat{\theta}_{ML}\right) \cdot \left(\theta - \widehat{\theta}_{ML}\right)^2} = e^{-\frac{1}{2}\frac{n}{\widehat{\theta}_{ML} \cdot (1-\widehat{\theta}_{ML})} \cdot \left(\theta - \widehat{\theta}_{ML}\right)^2}. \tag{5.6}$$

We can empirically, say for $\mathbf{x}$ drawn from a pseudo random number generator, plot the posterior density (5.3) as a function of θ and observe the property (5.6), when the length of a string $\mathbf{x}$ increases. In fact this holds independently of the prior density. This behaviour is clearly present in a typical simulation, the posterior densities in the Figure are not properly standardized.



5.3 Three MAPs for $\mathrm{Bin}(n, \theta)$

In all cases in this subsection $X \sim \mathrm{Bin}(n, \theta)$ and thus from Example 2.3 for $k = \{0, 1, \ldots, n\}$,

$$P(X = k|\theta) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}.$$

- **Jeffreys' Prior** (c.f. Example 8.1 below) on $\boldsymbol{\Theta}$ is

  $$\pi(\theta) = \frac{\theta^{-1/2} (1-\theta)^{-1/2}}{\mathrm{B}(1/2, 1/2)}.$$

  Then

  $$\widehat{\theta}_{\mathrm{MAP}} = \arg \max_{\theta \in [0,1]} \left(\theta^{k-1/2} (1-\theta)^{n-k-1/2} \right)$$

  $$\Leftrightarrow$$

  $$\widehat{\theta}_{\mathrm{MAP}} = \arg \max_{\theta \in [0,1]} \left[(k-1/2) \log \theta + (n-k-1/2) \log(1-\theta) \right]$$

  We set

  $$\phi(\theta) = (k-1/2) \log \theta + (n-k-1/2) \log(1-\theta).$$

  Then by some simple calculus we see that

  $$\frac{d}{d\theta} \phi(\theta) = 0 \Leftrightarrow \theta = \frac{k-1/2}{n-1},$$

  which clearly requires $n \geq 2$ and $k > 0$. If $n \geq 2$ and $k = 0$, the solution is negative, and cannot be used. $\phi(\theta)$ decreases to $-\infty$ as θ approaches 0, and $\phi(\theta)$ decreases to $-\infty$ as θ approaches 1 . Hence

  $$\widehat{\theta}_{\mathrm{MAP}}^{(1)} = \max \left(\frac{k-1/2}{n-1}, 0 \right)$$

  The special cases $n = 2$ and $n = 1$ can be treated by study of

  $$\phi(\theta) = (k-1/2) \log \theta + (n-k-1/2) \log(1-\theta)$$

- **Haldane's prior** on $\mathbf{\Theta}$ is

$$\pi(\theta) = \theta^{-1}(1-\theta)^{-1}$$

$$\phi(\theta) = (k-1)\log\theta + (n-k-1)\log(1-\theta)$$

  Same kind of calculations as in the first case yield

$$\frac{d}{d\theta}\phi(\theta) = 0 \Leftrightarrow$$

$$\theta = \frac{k-1}{n-2},$$

  which clearly requires $n \geq 3$ and $k > 0$. By reasoning similar to the one used above we get for $n \geq 3$ and $k \geq 0$

$$\widehat{\theta}^{(2)}_{\text{MAP}} = \max\left(\frac{k-1}{n-2}, 0\right)$$

- **Laplace's prior** is the uniform density on $[0,1]$, which we already have studied.

$$\phi(\theta) = k\log\theta + (n-k)\log(1-\theta)$$

$$\frac{d}{d\theta}\phi(\theta) = 0 \Leftrightarrow$$

$$\theta = \frac{k}{n}$$

  which, as it should, equals also the maximum likelihood estimate, already stated above.

$$\widehat{\theta}^{(3)}_{\text{MAP}} = \frac{k}{n}.$$

We summarize: If $n > 2$

$$\widehat{\theta}^{(1)}_{\text{MAP}} = \max\left(\frac{k-1/2}{n-1}, 0\right),$$

$$\widehat{\theta}^{(2)}_{\text{MAP}} = \max\left(\frac{k-1}{n-2}, 0\right),$$

$$\widehat{\theta}^{(3)}_{\text{MAP}} = \frac{k}{n}.$$

When n is large, then all three estimates are equivalent, as is expected by the Bernstein -von Mises Theorem.

6 More on Modelling and Learning: Dirichlet Priors and Posteriors

6.1 The Model Family

Let $X_1, X_2, \ldots, X_n$ be independent random variables assuming values in

$$\mathcal{X} = \{x_1, \cdots, x_L\}$$

with the common distribution

$$\theta_l = P\left(X_i = x_l\right), l = 1, 2, \ldots, L.$$

Hence $\theta_1 + \theta_2 + \ldots + \theta_L = 1$. Let $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$ be a string of symbols from $\mathcal{X}$ and let for $l = 1, 2, \ldots, L$

$$n_l = \text{ the number of times the symbol } x_l \text{ is found in } x_{i_1} x_{i_2} \ldots x_{i_n}.$$

We set

$$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_L)$$

and consider $\underline{\Theta}$ as a multivariate random variable that assumes values in the simplex

$$\mathbf{\Theta} = \{\underline{\theta} \mid \theta_1 + \theta_2 + \ldots + \theta_L = 1, \theta_l \geq 0, l = 1, \ldots, L\}.$$

THE MODEL FAMILY M:

CONDITIONED ON $\underline{\Theta} = \underline{\theta} \in \mathbf{\Theta}$, THE SYMBOLS IN $\mathbf{x}$ ARE INDEPENDENT.

Thus

$$P\left(\mathbf{x} \mid \underline{\theta}\right) = \theta_{i_1} \cdot \theta_{i_2} \cdots \theta_{i_n} = \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

Again we find a prior $\pi\left(\underline{\theta}\right)$ for Θ. Let us consider the *Dirichlet prior* given by

$$\pi\left(\underline{\theta}\right) = \begin{cases} \frac{\Gamma(\alpha)}{\prod_{j=1}^{L} \Gamma(\alpha q_j)} \prod_{j=1}^{L} \theta_j^{\alpha q_j - 1} & \underline{\theta} \in \mathbf{\Theta} \\ 0 & \text{elsewhere,} \end{cases}$$

where *the hyperparameters* are $\alpha > 0$, $q_j \geq 0$, $\sum_{j=1}^{L} q_j = 1$, $\Gamma\left(z\right)$ is the Euler gamma function as given in the appendix. The prior $\pi(\theta)$ is in (A.7) in the appendix given the symbol

$$Dir\left(\alpha q_1, \ldots, \alpha q_L\right).$$

By extension of Bayes' rule we get the *posterior*

$$\pi\left(\underline{\theta}|\mathbf{x};\underline{\alpha}\right)=\frac{P\left(\mathbf{x}\mid\underline{\Theta}=\underline{\theta}\right)\cdot\phi\left(\underline{\theta}\right)}{\int_{\boldsymbol{\Theta}}P\left(\mathbf{x}\mid\underline{\Theta}=\underline{\theta}\right)\cdot\pi\left(\underline{\theta}\right)d\underline{\theta}},\underline{\theta}\in\boldsymbol{\Theta}\tag{6.7}$$

and zero elsewhere. Using the Dirichlet integral expounded in the appendix we get

Proposition 6.1 *The posterior density* $\phi_{\underline{\Theta}|\mathbf{x}}\left(\underline{\theta}|\mathbf{x};\underline{\alpha}\right)$ *is a Dirichlet density*

$$Dir\left(n_1+\alpha q_1,\ldots,n_L+\alpha q_L\right)$$

or

$$\pi\left(\underline{\theta}|\mathbf{x};\underline{\alpha}\right)=\frac{\Gamma\left(n+\alpha\right)}{\prod_{i=1}^{L}\Gamma\left(\alpha q_i+n_i\right)}\prod_{i=1}^{L}\theta_i^{n_i+\alpha q_i-1}.\tag{6.8}$$

■

The posterior density is in the same family of densities as the prior. Hence the prior is closed under sampling or a conjugate prior.

6.2 Mean Posterior Estimate and a Rule of Succession

One useful property of the Dirichlet density is that we can compute explicitly the expectation of any θ_i with respect to the posterior density. In fact this expectation is by (A.9) and (6.8)

$$\widehat{\theta}_i=\int_{\boldsymbol{\Theta}}\theta_i\phi\left(\theta_1,\ldots,\theta_L|\mathbf{x};\underline{\alpha}\right)d\theta_1\ldots d\theta_L=\frac{n_i+\alpha q_i}{n+\alpha}.\tag{6.9}$$

This result can be seen as a *regularization* adding pseudocounts αq_i to the vector of observed counts $\underline{n}$ and then normalising so that $\sum_{i=1}^{L}\widehat{\theta}_i=1$. If we have $n=0$, the estimate is simply q_i.

Here we may note that a way of referring to α in (6.9) is to talk about the *flattening constant* (Bender 1996, pp. 554 - 555). The flattening constant determines a linear interpolation between the maximum likelihood estimate (see proposition 12.1 in section 12) $\frac{n_i}{n}$ of θ_i and the prior estimate q_i. Hence α has the interpretation as the degree of confidence we distribute between the data and the prior. The probability in (6.9) is known as a *rule of succession.*

7 Issues of Bayesian Learning (2): Bayes Factor & Bayesian Model Comparison/Selection & Occam's Razor

7.1 Bayes Factor

The **Bayes factor** compares posterior odds ratios to prior odds ratios of model families without supposing that any model is true or false. This deals with questions of the kind treated in hypothesis testing, but in particular with model comparison.

Bayes factor was developed in statistics for the purpose of evaluating the evidence in favor of a scientific theory i.e. hypothesis testing. Bayes factors offer a way of evaluating evidence in favour of a null hypothesis.

An example of model selection is determining the optimum level of complexity (= number of hidden layers and the width of the layers) required to develop an artificial neural network (ANN) for a given problem. This is a difficult task, and there is perhaps no generally accepted formal systematic model selection method for ANNs, but Bayes factors or the evidence defined below have been tested, when augmented by additional checks of complexity.

Bayesian model comparison is a method of model selection based on Bayes factors. Given that we are asked to choose between two parametric model families $\mathcal{M}_1$ and $\mathcal{M}_0$ on the basis of observed data x, the plausibility of the two different model families, parametrised by model parameter vectors θ_1 and θ_0 is assessed by the Bayes factor B_{01}^{π}.

We shall compare models based on the *marginal likelihood* a.k.a. (weight of) the evidence for each model family, which is the *probability the model family assigns to the observed data.*

We might choose the model that gives higher probability to the data, or average predictions from both models with weights based on their marginal likelihood. In the sequel x represents an observed data set $x = \{x_1, \ldots, x_n\}$.

If the likelihood corresponding to the maximum likelihood estimate of the parameter for each model is used, then we get a **likelihood-ratio test**.

The Bayes factor is defined as the ratio of the posterior probabilities of two model families $\mathcal{M}_1$ and $\mathcal{M}_0$ over the ratio of the prior probabilities of the model families, i.e.,

$$B_{01}^{\pi} := \frac{\frac{\Pi(\mathcal{M}_0|x)}{\Pi(\mathcal{M}_1|x)}}{\frac{\Pi(\mathcal{M}_0)}{\Pi(\mathcal{M}_1)}} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}$$

We compute

$$\Pi\left(\mathcal{M}_i|x\right) = \frac{p\left(x|\mathcal{M}_i\right) \cdot \Pi\left(\mathcal{M}_i\right)}{p\left(x\right)},$$

where

$$p\left(x\right) = \sum_{i=0}^{1} p\left(x|\mathcal{M}_i\right) \cdot \Pi\left(\mathcal{M}_i\right).$$

The Bayes' factor becomes

$$B_{01}^{\pi} := \frac{p\left(x|\mathcal{M}_0\right)}{p\left(x|\mathcal{M}_1\right)},$$

where the *marginal likelihood* or *evidence* is

$$p\left(x \mid \mathcal{M}_i\right) = \int_{\Theta_i} f_i\left(x|\theta_i\right) \pi_i\left(\theta\right) d\theta.$$

In section 13.4 it is shown that the marginal likelihood has a minimizing property of Kullback distances averaged over the prior. The following qualitative interpretation of a Bayes Factor B_{01}^{π} is known. For the unit dHart, see section 13.1. dHart is a change in a weight of evidence about as fine as humans can reasonably be expected to quantify their degree of belief in a model.

$$10 \cdot \log_{10} B_{01}^{\pi} = \text{ Bayes factor in dHart.}$$

- < 0 [dHart]: negative
- $0 - 5$ [dHart]: is barely worth a mention
- $5 - 10$ [dHart]: is substantial
- $10 - 15$ [dHart]: is strong
- $15 - 20$ [dHart]: is very strong
- over a 20 [dHart]: is decisive evidence in favour of model family $\mathcal{M}_0$.

Values in dHart below 0 take the inverted interpretation in favour of model family $\mathcal{M}_1$.

Example 7.1 We compute the Bayes factor (Bernardo and Smith (1994, pp. 392−393) under two parametric model families

$$\mathcal{M}_0 : \quad \text{Poisson distribution with } \Theta_0 \sim \Gamma\left(\alpha_0, \beta_0\right)$$

$$\mathcal{M}_1 : \quad \text{Geometric Distribution with } \Theta_1 \sim \mathcal{B}\mathrm{e}\left(\alpha_1, \beta_1\right)$$

Geometric Distribution is the parametric model

$$f_1\left(x \mid \theta_1\right) = \theta_1 \cdot \left(1-\theta_1\right)^x, x = 0, 1, 2, \ldots,$$

$$\mathbf{Theta}_1 = \{\theta \mid 0 \leq \theta \leq 1\}.$$

$$x_i | \Theta_1 = \theta_1 \sim f\left(x | \theta_1\right), \text{ I.I.D. },$$

$$x^{(n)} = \left(x_1, x_2, \ldots, x_n\right)$$

$$f_1\left(x^{(n)} \mid \theta_1\right) = \theta_1^n \cdot \left(1-\theta_1\right)^{\sum_{i=1}^n x_i}$$

Poisson distribution is the parametric model

$$f_0\left(x \mid \theta_0\right) = e^{-\theta_0} \frac{\theta_0^x}{x!}, x = 0, 1, 2, \ldots,$$

$$\mathbf{Theta}_0 = \{\theta \mid 0 < \theta\}.$$

$$x_i | \Theta_0 = \theta_0 \sim f_0\left(x | \theta_0\right), \text{ I.I.D. },$$

$$x^{(n)} = \left(x_1, x_2, \ldots, x_n\right)$$

$$f_0\left(x^{(n)} \mid \theta_0\right) = e^{-n\theta_0} \frac{\theta_0^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

If we consider single parameter values the Bayes ratio becomes a **likelihood ratio**:

$$\log \frac{f_0\left(x^{(n)} | \theta_0\right)}{f_1\left(x^{(n)} | \theta_1\right)} = \sum_{i=1}^n x_i \log\left(\frac{\theta_0}{1-\theta_1}\right) - n \log \theta_1 - n\theta_0 - \sum_{i=1}^n \log x_i$$

Then the marginal likelihood of $x^{(n)}$ under $\mathcal{M}_0$ is

$$\int_{\Theta_0} f_0\left(x^{(n)} \mid \theta_0\right) \pi_0\left(\theta_0\right) d\theta_0 =$$

$$= \int_0^\infty e^{-n\theta_0} \frac{\theta_0^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta_0^{\alpha_0 - 1} e^{-\beta_0\theta_0} d\theta_0$$

$$= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\prod_{i=1}^n x_i!} \underbrace{\int_0^\infty e^{-(\beta_0+n)\theta_0} \theta_0^{\sum_{i=1}^n x_i + \alpha_0 - 1} d\theta_0}_{= \frac{\Gamma(\alpha_0 + \sum_{i=1}^n x_i)}{(n+\beta_0)^{\alpha_0 + \sum_{i=1}^n x_i}}}$$

$$= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\prod_{i=1}^n x_i!} \frac{\Gamma(\alpha_0 + \sum_{i=1}^n x_i)}{(n+\beta_0)^{\alpha_0 + \sum_{i=1}^n x_i}}.$$

The marginal likelihood of $x^{(n)}$ under $\mathcal{M}_1$ is

$$\int_{\Theta_1} f_1\left(x^{(n)} \mid \theta_1\right) \pi_1\left(\theta_1\right) d\theta_1 =$$

$$= \int_0^1 \theta_1^n \cdot (1-\theta_1)^{\sum_{i=1}^n x_i} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \theta_1^{\alpha_1 - 1}(1-\theta_1)^{\beta_1 - 1} d\theta_1$$

$$= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \underbrace{\int_0^1 \theta_1^{n+\alpha_1-1}(1-\theta_1)^{\sum_{i=1}^n x_i + \beta_1 - 1} d\theta_1}_{= \frac{\Gamma(n+\alpha_1)\Gamma(\sum_{i=1}^n x_i + \beta_1)}{\Gamma(n + \sum_{i=1}^n x_i + \alpha_1 + \beta_1)}}$$

$$= \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n+\alpha_1)\Gamma(\sum_{i=1}^n x_i + \beta_1)}{\Gamma(n + \sum_{i=1}^n x_i + \alpha_1 + \beta_1)}.$$

Then Bayes' factor becomes

$$B_{01}^{\pi} = \frac{\int_{\Theta_0} f\left(x|\theta_0\right) \pi_0\left(\theta_0\right) d\theta_0}{\int_{\Theta_1} f\left(x|\theta_1\right) \pi_1\left(\theta_1\right) d\theta_1}$$

$$= \frac{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\prod_{i=1}^n x_i!} \frac{\Gamma(\alpha_0 + \sum_{i=1}^n x_i)}{(n+\beta_0)^{\alpha_0 + \sum_{i=1}^n x_i}}}{\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n+\alpha_1)\Gamma(\sum_{i=1}^n x_i + \beta_1)}{\Gamma(n + \sum_{i=1}^n x_i + \alpha_1 + \beta_1)}}.$$

The Bayes factor in favour of Poisson distribution relative to Geometric distribution established above depends only on the data and the hyperparameters (contrary to the ratio of likelihoods). This is because all unknown parameters are integrated out. All constants of integration are to be kept. In likelihood ratio tests only those parts of the densities that depend on the parameters are kept.

■

7.2 Bayesian Model Comparison & Occam's Razor

In the fourteenth century, William of Ockham proposed:

> Pluralitas non est ponenda sine neccesitate

which is by Latinists translated as *entities should not be multiplied unnecessarily.* Its original historic context was theological, but the (heuristic) principle remains relevant for machine learning today, where it is called **Occam's Razor**, and it might be rephrased as *models should be no more complex than is sufficient to explain the data*, Tipping (2004) or *Given multiple hypotheses that are consistent with the data, the simplest should be preferred.*

We quote Legg, Shane and Hutter, Marcus: Universal intelligence: A definition of machine intelligence, *Minds and Machines*, 17, 391−444, 2007:

> Consider the following type of question which commonly appears in intelligence tests. There is a sequence such as 2, 4, 6, 8, and the test subject needs to predict the next number. Of course the pattern is immediately clear: the numbers are increasing by 2 each time, or more mathematically, the kth item is given by 2k. An intelligent person would easily identify this pattern and predict the next digit to be 10. However, the polynomial $2k^4 - 20k^3 + 70k^2 - 98k + 48$ is also consistent with the data, in which case the next number in the sequence would be 58. Why then, even if we are aware of the larger polynomial, do we consider the first answer to be the most likely one? It is because we apply, perhaps unconsciously, the principle of Occam's razor. The fact that intelligence tests define this as the "correct" answer, shows us that using Occam's razor is considered the intelligent thing to do. Thus, although we do not usually mention Occam's razor when defining intelligence, the ability to effectively use it is an important facet of intelligent behaviour.

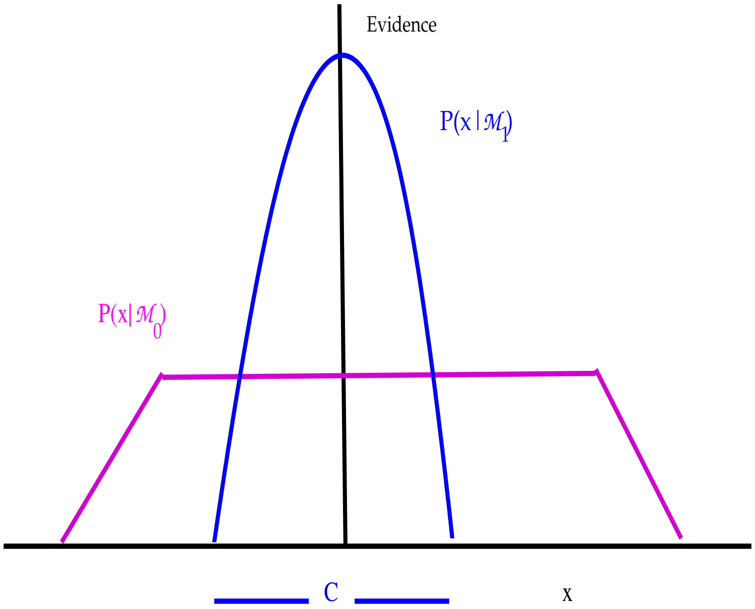
A more generally operational form of the razor is next derived from Bayesian model comparison. This is based on Bayes factors and can be used to compare models that do not fit the observations equally well. These methods can sometimes optimally balance the complexity and power of a model. The exact Occam factor is analytically intractable, but approximations such as *BIC= Bayesian information criterion*, and others, can be used without difficulty.

7.3 Bayes factor has a built-in Occam's Razor

7.3.1 A Picture

Suppose $\mathcal{M}_0$ is a more complex model than $\mathcal{M}_1$, in the sense that $\mathcal{M}_0$ has more parameters. Then $\mathcal{M}_0$ will model the data more closely. The Bayes factor is capable of making a trade-off between good fit and model complexity (elaborated in, e.g., MacKay (1992)).

Bayes ratio measures, intuitively speaking, the proportion of how much the model family predicted the data that occurred. A simple model family makes a limited range of predictions, a more complex family $\mathcal{M}_0$ will be able to over-fit a small number of cases to a huge number of models, or fit a greater variety of data sets but with low probability $p\,(x|\mathcal{M}_0)$, hence the less complex model family will make $p\,(x|\mathcal{M}_1)$ high in a small region of data and is there the more probable model.



The figure, from MacKay (1992), another is found in Tipping (2004), gives an intuition of how complex models are penalized. The horizontal axis represents the space of all possible data sets x. Bayes factor rewards models in proportion to how much they predicted the data that occurred, which is quantized by the evidence $p\,(x|\mathcal{M}_0)$. A simple model family $\mathcal{M}_1$ makes only a limited range of predictions. A more powerful model family $\mathcal{M}_0$ has more free parameters than $\mathcal{M}_1$ and is able to predict a greater variety of data sets. This means, however, that $\mathcal{M}_0$ does not predict the data sets in region C with as strong evidence as $\mathcal{M}_1$. In the region C a complex model family $\mathcal{M}_0$ with many parameters, each of which is free to vary over a large range $\triangle\theta_0^0$, will be penalized with a small Bayes factor w.r.t. a simpler model family.

7.3.2 Mathematical Approximation, BIC

Some piece of mathematical approximation can be added. The integration method of Laplace is a technique used to approximate integrals of the form

$$\int_a^b e^{Mf(x)}\,dx,$$

when we assume that the function $f(x)$ has a unique global maximum at x_0 and is twice differentiable. Main contributions to the integral of $f(x)$ will come only from points x in a neighbourhood of x_0, the size of which can be estimated. The formula obtained is

$$\int_a^b e^{Mf(x)}\,dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)} \text{ as } M \to \infty.$$

To apply this idea we write

$$\int_{\Theta_i} f\,(x|\theta_i)\,\pi_i\,(\theta)\,d\theta = \int_{\Theta_i} e^{\ln f(x|\theta_i)\pi_i(\theta)} d\theta$$

and make series expansions of $\ln f\,(x|\theta_i)\,\pi_i\,(\theta)$, c.f. Appendix 14.

Thus, the integrand in the evidence can be approximated by the height of the peak of the integral at the most probable parameters, $\widehat{\theta}_{\text{MAP}}$, times its width, $\triangle\theta_i$.

$$\int_{\Theta_i} f\,(x|\theta_i)\,\pi_i\,(\theta)\,d\theta \approx f_i\left(x|\widehat{\theta}_{i,\text{MAP}}\right)\pi_i\left(\widehat{\theta}_{i,\text{MAP}}\right)\triangle\theta_i.$$

Or

$$\underbrace{p\,(x|\mathcal{M}_i)}_{\text{Evidence}} \approx \underbrace{f_i\left(x|\widehat{\theta}_{i,\text{MAP}}\right)}_{\text{Best fit likelihood}} \underbrace{\pi_i\left(\widehat{\theta}_{i,\text{MAP}}\right)\triangle\theta_i}_{\text{Occam factor}}.$$

The quantity $\triangle\theta_i$ is the posterior uncertainty in θ. Suppose that $\pi_i\left(\widehat{\theta}_{i,\mathrm{MAP}}\right) = \frac{1}{\triangle\theta_i^0}$. Here $\triangle\theta_i^0$ represents the range of values of θ that $\mathcal{M}_i$ thought possible before data arrived. Then

$$\text{Occam factor} = \frac{\triangle\theta_i}{\triangle\theta_i^0}.$$

This is seen as the ratio of the posterior accessible volume of $\mathbf{\Theta}_i$ to the prior accessible volume or as the factor by which $\mathbf{\Theta}$ collapses when the data arrive. The model family $\mathcal{M}_i$ is composed of a certain number of equivalent submodels, of which only one survives after the data arrive. The Occam factor is the inverse of that number. The log of the Occam factor can be interpreted as the amount of information we gain about the model family when the data arrive.

The Bayes' factor is now written as

$$B_{01}^{\pi} = \frac{p\left(x|\mathcal{M}_0\right)}{p\left(x|\mathcal{M}_1\right)} \approx \frac{f\left(x|\widehat{\theta}_{0,\mathrm{MAP}}\right)}{f\left(x|\widehat{\theta}_{1,\mathrm{MAP}}\right)} \cdot \frac{\triangle\theta_0}{\triangle\theta_0^0}\frac{\triangle\theta_1^0}{\triangle\theta_1}.$$

BIC=Bayesian Information Criterion

By some tedious multivariable calculus it can be demonstrated that the logarithm of evidence can be well approximated by

$$\ln p\left(x \mid \mathcal{M}_i\right) = \ln f\left(x|\widehat{\theta}_i\right) - \frac{|\theta_i|}{2}\ln n, \qquad (7.1)$$

where $\widehat{\theta}_i$ is the maximum likelihoood estimate of θ_i and $|\theta_i|$ is the dimension of θ_i, i.e. the number of parameters in the vector θ_i and n is got from $x = \{x_1, \ldots, x_n\}$.

The expression in the right hand side of (7.2) is the BIC=Bayesian Information Criterion,

$$\mathrm{BIC}_i \overset{\mathrm{def}}{=} \ln f\left(x|\widehat{\theta}_i\right) - \frac{|\theta_i|}{2}\ln n. \qquad (7.2)$$

The criterion BIC_i is maximized to find the best model family amongst a finite number of model families $\mathcal{M}_i$. Here we see that a complex model giving good fit to data will yield a high value of $\ln f\left(x|\widehat{\theta}_i\right)$, but this is punished/counterbalanced by the term $\frac{|\theta_i|}{2}\ln n$.

8 Issues of Bayesian Learning (3):Quantification of the Prior

The angry controversies around the Bayesian procedures concern amongst other things the nature of the prior probability density $\pi(\theta)$. The difficulty is a subjective judgement being merged into data analysis that is expected to be an objective science without arbitrary elements. Recall Gunnar Englund's words above

> ...In fact, using a Bayesian approach a statement like Halzen's could be made, but that would presuppose an (arbitrary) a priori probability that the theory is correct.

8.1 How do we choose $\pi(\theta)$?

- Conjugate prior
- Non-informative or reference prior:
  Reference priors produce *objective Bayesian inference*, in the sense that inferential statements depend only on the assumed model and the available data, and the prior distribution used to make an inference is least informative (in a certain information-theoretic sense). Reference priors have been rigorously defined in specific contexts and heuristically defined in general.

  – Laplace's prior

  – Jeffreys' prior: Let $I(\theta)$ be the Fisher information (5.5) of a parametric model. Take the prior density as

  $$\pi(\theta) \stackrel{def}{=} \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}d\theta},$$

  assuming the integral exists. This prior is invariant to monotonous transformations of θ.

Example 8.1 It turns out that Jeffreys' prior for θ in binomial likelihood is obtained by taking $\alpha = 1/2$ and $\beta = 1/2$ in

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

and with $\Gamma(1/2 + 1/2) = 1$, $\Gamma(1/2) = \sqrt{\pi}$,

$$\pi_{Jeffreys}(p) = \begin{cases} \frac{1}{\pi} p^{-1/2}(1-p)^{-1/2} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

- Maximum entropy prior (Jaynes 2003). E.T. Jaynes (2003) regards the Bayesian reasoning as the very logic of science.

One form of Bayesian learning relies upon a **personalistic theory of probability** for quantification of prior knowledge. In such a theory

- probability measures the confidence that a particular individual (assessor) has in the truth of a particular proposition
- no attempt is made to specify which assessments are correct
- personal probabilities should satisfy certain **postulates of coherence**. Personal probabilities do follow the ordinary rules for probability calculus. The validity of these rules expresses to the self-consistency in terms of responses to various betting propositions. Consistent decisions can be obtained from these probabilities.

There has been a development of diverse methods of assessing personalistic probability:

1. Choice of prior distributions: questionnaires R.L.Winkler (work published in

    - Robert L. Winkler: The Assessment of Prior Distributions in Bayesian Analysis
      Journal of the American Statistical Association, Vol. 62, No. 319. (Sep., 1967), pp. 776-800.)

devises questionnaires (or interviews) to elicit information to write down a prior distribution. Students of University of Chicago were asked to, e.g., assess the uncertainty about the probability of a randomly chosen student of University of Chicago being Roman Catholic using a probability distribution. The assessment was done by four different methods, like giving fractiles, making bets, assessing impact of additional data, drawing graphs. One interesting finding is that the assessments by the same person using different methods may be conflicting.

2. Assessing Priors: Conjugate Prior The interviews by Winkler were mathematically speaking all concerned with assessing the prior of θ in a Bernoulli $\mathrm{Be}\,(\theta)$ $-$ I.I.D. process. Winkler claims a sensitivity analysis (loc.cit p. 791) showing that the prior distributions assessed by the interviews yielded posterior distributions that were 'only little' different (by a test of goodness-of-fit) from those obtained from Beta densities on θ.

3. Choice of prior distributions by elicitation

   - A. O$'$Hagan: Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* , 47, pp. 21$-$35, 1998.

   Not only priors are elicited in

   - L. Gingnell, U. Franke, R. Lagerström, E. Ericsson, J. Lilliesköld: Quantifying Success Factors for IT Projects - An Expert-Based Bayesian Model. *Information Systems Management*, pp. 21$--$36, 2014
   - R.L. Keeney & D. von Winterfeldt: Eliciting Probabilities in Complex Technical Problems. *IEEE Transactions on Engineering Management*, 38, pp.191$-$201, 1991.

4. Choice of prior distributions by assessment can evolve rapidly to a topic of research in cognitive science and/or economic behaviour, c.f.,

   - C-A. S. Stael von Holstein: *Assessment and Evaluation of Subjective Probability Distributions.* 1970, Stockholm School of Economics.

- A. G. Wilson: *Cognitive factors affecting subjective probability assessments.*
  ISDS Discussion Paper # 94-02,
  `http://www.isds.duke.edu/`

9 On the fundamental task of statistical learning

9.1 General Outline

It has been said that the fundamental task of statistical learning is that of passing from one set of data or observations $\mathcal{S} = x$ to express an opinion about another, as yet unobserved set $\mathcal{T} = y$. We shall handle this topic in terms of the *predictive probability density* or *predictive probability mass function*

$$g\left(y|x\right) = \int_{\mathbf{\Theta}} f\left(y|\theta\right) \pi\left(\theta \mid x\right) d\theta. \tag{9.1}$$

This section will give an extensive derivation and argument to justify $g\left(y|x\right)$ as the expression we desire for. We illustrate by an example the meaning of the issue to be treated.

Example 9.1 (The Probability for the Outcome of the Next Toss) In the thumbtack model we may be concerned with

$$P\left(X_{n+1} = \text{head}|\mathbf{x}\right),$$

if X_{n+1} is a random variable modelling the next toss $\mathcal{T}$, given n flips of the thumbtack as recorded in $\mathcal{S} = \mathbf{x}$. In other words, $P\left(X_{n+1} = \text{head}|\mathbf{x}\right)$ can be found by an integral of the form (9.1). For this we shall rerun some of the machinery in the thumbtack model family in a context given by Th. Bayes himself.

■

The book Bishop (2006) explains how (9.1) is used in training neural networks.

9.2 Predictive Distribution

9.2.1 Formal manipulations

We assume a parametric model with continuous variables and joint distribution $(x, y, \theta) \sim \phi(x, y, \theta)$ so that:

$$\phi(x, y) = \int_{\boldsymbol{\Theta}} \phi(x, y, \theta) \, d\theta.$$

The aim is to derive (9.1) from this. We set

$$m(x) = \int \int_{\boldsymbol{\Theta}} \phi(x, y, \theta) \, d\theta dy.$$

The conditional distribution $g(y|x)$ is then

$$g(y|x) = \frac{\phi(x, y)}{m(x)} = \frac{\int_{\boldsymbol{\Theta}} \phi(x, y, \theta) \, d\theta}{m(x)}.$$

By successive applications of the definition of conditional density

$$\frac{\phi(x, y)}{m(x)} = \frac{\int_{\boldsymbol{\Theta}} \phi(x, y, \theta) \, d\theta}{m(x)} = \frac{\int_{\boldsymbol{\Theta}} g(y|x, \theta) \, \psi(x, \theta) \, d\theta}{m(x)} = \frac{\int_{\boldsymbol{\Theta}} g(y|x, \theta) \, f(x|\theta) \, \pi(\theta) \, d\theta}{m(x)}.$$

Here

$$m(x) = \int \int_{\boldsymbol{\Theta}} \phi(x, y, \theta) \, d\theta dy =$$

$$= \int_{\boldsymbol{\Theta}} \int g(y|x, \theta) \, dy f(x|\theta) \, \pi(\theta) \, d\theta = \int_{\boldsymbol{\Theta}} f(x|\theta) \, \pi(\theta) \, d\theta,$$

since $g(y|x, \theta)$ is a probability density. Thus

$$g(y|x) = \frac{\int_{\boldsymbol{\Theta}} g(y|x, \theta) \, f(x|\theta) \, \pi(\theta) \, d\theta}{m(x)} = \frac{\int_{\boldsymbol{\Theta}} g(y|x, \theta) \, f(x|\theta) \, \pi(\theta) \, d\theta}{\int_{\boldsymbol{\Theta}} f(x \mid \theta) \cdot \pi(\theta) \, d\theta}$$

$$= \int_{\boldsymbol{\Theta}} g(y|x, \theta) \, \frac{f(x|\theta) \, \pi(\theta)}{\int_{\boldsymbol{\Theta}} f(x \mid \theta) \cdot \pi(\theta) \, d\theta} d\theta.$$

Here we recognize Bayes' rule. Thus

$$g(y|x) = \int_{\boldsymbol{\Theta}} g(y|x, \theta) \, \pi(\theta \mid x) \, d\theta.$$

If y and x are conditionally independent given $\Theta = \theta$, then

$$g\left(y|x,\theta\right) = g\left(y|\theta\right)$$

and

$$g\left(y|x\right) = \int_{\boldsymbol{\Theta}} g\left(y|\theta\right)\pi\left(\theta \mid x\right)d\theta.$$

If we assume that $g\left(y|\theta\right) = f\left(y|\theta\right)$, then we get the final formula as expected in (9.1). We give a name to what we have dealt with.

[**(Posterior) Predictive Distribution**]

$$g\left(y|x\right) = \int_{\boldsymbol{\Theta}} f\left(y|\theta\right)\pi\left(\theta \mid x\right)d\theta. \tag{9.2}$$

■

For some of us, this is one of the best illustrations of the importance of Bayes' rule. To claim this in a more acute manner, we consider updates.

9.2.2 The probability of the next data point

We introduce an ordering of data, like, e.g., in a sequence. $x = x^{(n)} = (x_1, \ldots, x_n)$ and $y = x_{n+1}$.

$$g\left(x_{n+1}|x^{(n)}\right) = \int g\left(x_{n+1}|x^{(n)},\theta\right)\pi\left(\theta \mid x^{(n)}\right)d\theta.$$

We assume conditional independence w.r.t. $\Theta = \theta$

$$\begin{aligned}\phi\left(x,y|\theta\right) = \phi\left(x^{(n)},x_{n+1}|\theta\right) &= \left(f\left(x_1|\theta\right)\cdots f\left(x_n|\theta\right)\right)\cdot f\left(x_{n+1}|\theta\right)\\ &= \phi\left(x^{(n)}|\theta\right)\cdot f\left(x_{n+1}|\theta\right)\\ \phi\left(x^{(n)},x_{n+1} \mid \theta\right) &= \phi\left(x^{(n)}|\theta\right)\cdot f\left(x_{n+1}|\theta\right).\end{aligned}$$

Then

$$g\left(x_{n+1}|x^{(n)},\theta\right) = \frac{\phi\left(x^{(n)},x_{n+1} \mid \theta\right)}{\phi\left(x^{(n)}|\theta\right)} = f\left(x_{n+1}|\theta\right)$$

and

$$g\left(x_{n+1}|x^{(n)}\right) = \int_{\boldsymbol{\Theta}} f\left(x_{n+1}|\theta\right)\pi\left(\theta \mid x^{(n)}\right)d\theta.$$

Note that there was conditional independence w.r.t. $\Theta = \theta$, but x_{n+1} and $x^{(n)}$ are no longer independent in $g\left(x_{n+1}|x^{(n)}\right)$. We can learn something about x_{n+1} from $x^{(n)}$.

9.2.3 Updates of Posterior

In order to update $g\left(x_{n+1}|x^{(n)}\right)$ for the next data point, i.e. to get $g\left(x_{n+2}|x^{(n+1)}\right)$, it suffices to update $\pi\left(\theta \mid x^{(n)}\right)$. The first expression for up-date of posterior distribution is

$$\begin{aligned}
\pi\left(\theta \mid x^{(n+1)}\right) &= \frac{f\left(x^{(n+1)}|\theta\right)\pi\left(\theta\right)}{\int_{\boldsymbol{\Theta}} f\left(x^{(n+1)} \mid \theta\right)\cdot\pi\left(\theta\right)d\theta} \\
&= \frac{f\left(x_{n+1}|\theta\right)f\left(x_1|\theta\right)\cdots f\left(x_n|\theta\right)\pi\left(\theta\right)}{\int f\left(x_{n+1}|\theta\right)f\left(x_1|\theta\right)\cdots f\left(x_n|\theta\right)\cdot\pi\left(\theta\right)d\theta} \\
&= \frac{f\left(x_{n+1}|\theta\right)\frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{m\left(x^{(n)}\right)}}{\int_{\boldsymbol{\Theta}} f\left(x_{n+1}|\theta\right)\frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{m\left(x^{(n)}\right)}d\theta} \\
&= \frac{f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)}{\int_{\boldsymbol{\Theta}} f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)d\theta}.
\end{aligned}$$

The final up-date of posterior distribution is thus

$$\pi\left(\theta \mid x^{(n+1)}\right) = \frac{f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)}{\int_{\boldsymbol{\Theta}} f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)d\theta}$$

Hence, under the assumptions made, we can update posterior distribution in a sequential manner when new data points accrue. Or, we can use the posterior of $\pi\left(\theta|x^{(n)}\right)$ as a new prior for computing $\pi\left(\theta \mid x^{(n+1)}\right)$.

Example 9.2 $f(x \mid \theta)$ is the density of $\mathrm{N}\left(\mu,\sigma^2\right)$, the mean μ and variance σ^2 are unknown, i.e., $\theta = (\mu,\sigma^2)$. The (improper) prior density is taken as

$$\pi\left(\theta\right) \propto d\mu\frac{1}{\sigma}d\sigma.$$

Let $x = \underline{x}_n = \left(x^{(1)},\ldots,x^{(n)}\right)$, $x^{(i)} \sim \mathrm{N}\left(\mu,\sigma^2\right)$. We have the estimates $\widehat{\mu}$ $= \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^{(i)}-\widehat{\mu}\right)^2$. The predictive density is

$$g\left(x^{(n+1)}|\underline{x}_n\right) = \sqrt{\frac{n}{(n^2-1)\pi}}\cdot\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}(n-1)s\right)}\cdot\left(1+\frac{n\left(x^{(n+1)}-\widehat{\mu}\right)^2}{(n^2-1)s^2}\right)^{-n/2}$$

is known in Bayesian statistics as 't-like distribution'[1] .

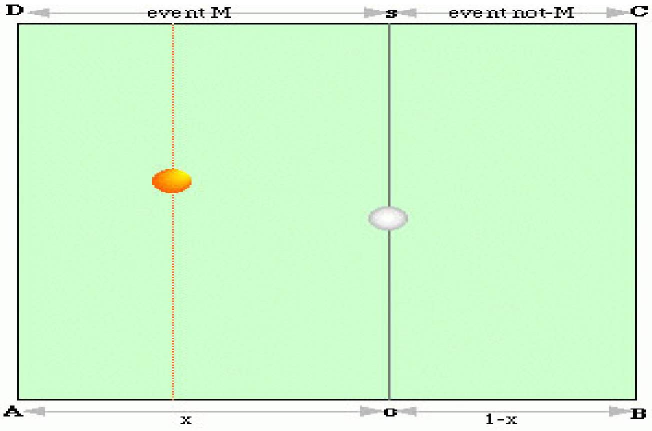
■

[1] J.M. Dickey (1968): Three Multidimensional-Integral Identities with Bayesian Applications. *The Annals of Mathematical Statistics*, 39, pp. 1615−1628.

9.3 Bayes' Billiard Ball

9.3.1 White Ball and Orange Balls

A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at θ, not disclosed to us. A second ball O is rolled n times under the same assumptions and X denotes the number of times O stops to the left of W. Given $X = x$, what can we learn about θ? (In the figure $x \leftrightarrow \theta$.) It is, however, not perfectly clear what questions the good reverent was set to study by this arrangement.



The square table is made in such a way that, if the White or Orange ball be thrown on it, the probability that it rests on any part of the plane is the same. First, the White ball is thrown, and suppose it rests on line **os**; then the Orange ball is thrown n times. We define event M as any event the Orange ball rests between **A** and **o**, and not-M as its resting between **o** and **B**. What this means is this:

The first throw of the White ball determines the value of probability x (i.e. the probability of an unknown event) from a uniform distribution between 0 and 1; and then a series of trials, with probability x of success (i.e., M) is generated (this provides the data on which to infer the correct value of x).

Then, thanks to the geometrical representation of the problem in the Figure, we can obtain the solution to the initial problem, by calculating integration. Although we have omitted mathematical formulas, the preceding is the central idea.

We let Θ be a random variable, with values in $0 \leq \theta \leq 1$. Conditionally on $\Theta = \theta$, the rolls are taken as outcomes of I.I.D Be(θ) R.V's. Hence for $x = 0, 1, 2, \ldots, n$,

$$\begin{aligned} f(x|\theta) &= P\left(X = x \mid \Theta = \theta\right) \\ &= \left(\begin{array}{c} n \\ x \end{array}\right) \theta^{x} \cdot (1-\theta)^{n-x}, \end{aligned}$$

which is the Binomial distribution. We introduce the Beta prior $\mathcal{B}\text{eta}(\alpha,\beta)$

$$\pi(\theta)=\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} & 0<\theta<1\\ 0 & \text{elsewhere,}\end{cases}$$

and

$$B\left(\alpha,\beta\right):=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

$\alpha>0$ and $\beta>0$ are the hyperparameters. $\alpha=\beta=1$ yields the uniform distribution on $(0,1)$. Then it follows that

$$\pi\left(\theta\mid x\right)=\begin{cases}\frac{1}{B(x+\alpha,n-x+\beta)}\cdot\theta^{x+\alpha-1}\left(1-\theta\right)^{\beta+n-x-1} & 0\leq\theta\leq 1\\ 0 & \text{elsewhere.}\end{cases}$$

This is the Beta density $\mathcal{B}\text{eta}(\alpha+x,\beta+n-x)$. Note that we get the posterior in (5.2) when $\alpha=\beta=1$.

9.3.2 What did Bayes want to find out?

There is an interpretation of Bayes' work claiming that the problem really attacked and solved by Bayes was: What should $\pi(\theta)$ be so that

$$m(x)=\int_0^1 f\left(x\mid\theta\right)\cdot\pi(\theta)d\theta=\frac{1}{(n+1)} \tag{9.3}$$

holds for the billiard balls. One can check that if the prior is $\mathcal{B}\text{eta}(1,1)$, the uniform distribution on $[0,1]$, then

$$m(x)=\int_0^1 f\left(x\mid\theta\right)\cdot d\theta=\left(\begin{array}{c}n\\ x\end{array}\right)\frac{x!(n-x)!}{(n+1)!}$$

$$=\frac{n!}{x!(n-x)!}\frac{x!(n-x)!}{(n+1)!}=\frac{1}{(n+1)}.$$

The more difficult mathematical question is the show that if (9.3) is true, then the prior must be $\mathcal{B}\text{eta}(1,1)$. If this was actually what Th. Bayes proved, then his approach to finding priors was quite different from the approaches of his followers mentioned in section 8.1 above.

9.4 Next rolls of the Bayes Orange Ball

The predictive distribution of y positions of O left of W in r additional rolls is with uniform prior density $\pi(\theta)$ given by

$$g\left(y|x\right) = \int_0^1 f\left(y|\theta\right)\pi\left(\theta \mid x\right)dp$$

$$= \left(\begin{array}{c} r \\ y \end{array}\right)\int_0^1 \theta^y \cdot \left(1-\theta\right)^{r-y}\pi\left(\theta \mid x\right)d\theta$$

$y = 0, 1, \ldots, r$

$$\int_0^1 \theta^y \cdot \left(1-\theta\right)^{r-y}\pi\left(\theta \mid x\right)d\theta =$$

$$\int_0^1 \theta^y \cdot \left(1-p\right)^{r-y}\frac{1}{B(x+1,n-x+1)}\cdot p^x\left(1-\theta\right)^{n-x}dp =$$

$$= \frac{1}{B(x+1,n-x+1)}\int_0^1 \theta^{y+x}\cdot\left(1-\theta\right)^{r+n-y-x}d\theta =$$

$$= \frac{B(y+x+1,r+n-x-y+1)}{B(x+1,n-x+1)}.$$

by the Beta integral. In other words

$$g(y|x;r) = \left(\begin{array}{c} r \\ y \end{array}\right)\int_0^1 \theta^y \cdot \left(1-\theta\right)^{r-y}\pi\left(\theta \mid x\right)d\theta$$

$$= \left(\begin{array}{c} r \\ y \end{array}\right)\frac{B(y+x+1,r+n-x-y+1)}{B(x+1,n-x+1)}.$$

$$= \frac{r!}{(r-y)!y!}\frac{\Gamma(n+2)\Gamma(y+x+1)\Gamma(r+n-x-y+1)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(r+n+2)}.$$

Hence the **predictive distribution for $y=1$ in the next $(r=1)$ roll of Bayes' Ball** is nothing but

$$g(1|x;1) = \frac{\Gamma(x+2)\Gamma(n-x+1)\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(n+3)}$$

$$= \frac{(x+1)!(n-x)!(n+1)!}{x!(n-x)!(n+2)!} = \frac{x+1}{n+2}$$

This is (notoriously) famous predictive probability, known as *Laplace′s rule of succession*,

$$\frac{x+1}{n+2} = \frac{x+1}{(n+1)+1}$$

The prior knowledge can be translated into equivalent sample information. We know that

$$\widehat{\theta}_{\mathrm{ML}} = \frac{x}{n}$$

If you observed $x = 0$, would you believe in the estimate $\widehat{\theta} = 0$ for all future purposes ? The predictive probability found above

$$\frac{x+1}{n+2} = \frac{x+1}{(n+1)+1}$$

is a maximum likelihood estimate of θ when $n+1$ rolls of the ball O and the first roll of the ball W are included in the data.

Wilson (1927) suggests for the thumbtack model a different rule of succession

$$\widehat{\theta}^{W} = \frac{k + \alpha^2/2}{n + \alpha^2}. \tag{9.4}$$

Wilson says that the value of α depends on 'our readiness to to gamble on the typicalness of our experience'.

10 Priors on Perceptrons

10.1 Bayesian Version of the Computational Learning Theory (Buntine 1992)

We are given a training set $\mathcal{S} = \{z_i = (\mathbf{x}^{(i)}, y^{(i)}), i = 1, ..., n\}$, where $\mathbf{x} \in \mathcal{X}$ is an input and $y \in \mathcal{Y}$. This is drawn from an unknown probability distribution $P_Z = P_{X,Y}$.

We are given a *fixed* set $\mathcal{H}$ of functions h, $\mathcal{X} \overset{h}{\mapsto} \mathcal{Y}$. $\mathcal{H}$ is called the hypothesis space. The task of learning is to find the function h^*, that performs best on new, yet unseen patterns $z = (\mathbf{x}, y)$ drawn from $P_Z = P_{X,Y}$ in the sense of predicting y from $\mathbf{x}$. The loss function penalizing the prediction error is $l(h(\mathbf{x}), y)$. (Here l does not necessarily refer to loglikelihood). The

- The **training error** of h is

$$R_{\mathrm{emp}}(h) \overset{\mathrm{def}}{=} \frac{1}{m}\sum_{i=1}^{n} l(h\left(\mathbf{x}^{(i)}\right), y^{(i)}).$$

- The **generalization error** of h is

$$R(h) \overset{\mathrm{def}}{=} E\left[l(h(\mathbf{X}), Y)\right].$$

where the expectation is w.r.t. $P_Z = P_{X,Y}$.

$R_{\mathrm{emp}}(h)$ provides an estimate of $R(h)$. The posterior probability of the hypothesis h given the training set is

$$P(H = h \mid \mathcal{S}) = \frac{P(Y^{(n)} = y^{(n)} \mid X^{(n)} = \mathbf{x}^{(n)}, H = h)P(X^{(n)} = \mathbf{x}^{(n)})P(H = h)}{C(\mathcal{S})}.$$

The *Bayes classification strategy* is to minimize the expected loss computed according to $P(H = h \mid z^n)$. Or,

$$\mathrm{Bayes}_{\mathcal{S}}(\mathbf{x}) = \mathrm{argmin}_{y\in\mathcal{Y}} E_{H|\mathcal{S}}\left[l(H(\mathbf{x}), y)\right]$$

Note that this strategy does not correspond to any fixed h, and is therefore computationally demanding. Let $\mathcal{A}$ denote a generic algorithm for learning h. It can be proven, Herbrich et.al. (2001), that the Bayes classification strategy minimizes the average generalization error

$$R_m(\mathcal{A}) \overset{\mathrm{def}}{=} E_H\left[E_{Z^m|H=h}\left[E_X\left[E_{Y|X=x,H=h}\left[l\left(\mathcal{A}(\mathcal{S})(\mathbf{x}), Y\right)\right]\right]\right]\right].$$

under some not so hard conditions. *Bayes point machine* minimizes

$$\mathcal{A}_{\mathrm{Bp}}(\mathcal{S})(z) = \mathrm{argmin}_{h\in\mathcal{H}} E_X\left[E_{H|\mathcal{S}}\left[l(h(X), H(X))\right]\right]$$

and is, under some conditions, a good approximation of the Bayes classification strategy.

10.2 Bayesian Learning of Perceptrons

We recall from Lecture 1.

$$f\left(\mathbf{x}\right) = \ll \mathbf{w}, \mathbf{x} \gg +b.$$

The function f determines a half plane in the Euclidean space R^p

$$D = \{\mathbf{x} \in R^p \mid f(\mathbf{x}) = 0\}$$

which divides the space into two half spaces, so that $\mathcal{R}_1$ and $\mathcal{R}_2$, as

$$\mathcal{R}_{+1} = \{\mathbf{x} \in R^p \mid f(\mathbf{x}) > 0\}$$

$$\mathcal{R}_{-1} = \{\mathbf{x} \in R^p \mid f(\mathbf{x}) < 0\}$$

The perceptron rule of classification is to assign $y = +1$ to $\mathbf{x} \in \mathcal{R}_1$ and $y = -1$ to $\mathbf{x} \in \mathcal{R}_{-1}$.

Next we follow an idea due to Radford M. Neal. It is known that the vector $\mathbf{c}$ in the hyperlane D that is *closest to* the origin, $\mathbf{0}$, is given by

$$\mathbf{c} = \frac{(-b)}{\|\mathbf{w}\|^2}\mathbf{w}.$$

The vector $\mathbf{c}$ determines the orientation of the hyperplane uniquely. Thus we can write

$$f(\mathbf{x}) = \ll \frac{(-b)}{\|\mathbf{w}\|^2}\mathbf{w}, \mathbf{x} \gg +b.$$

But then the separating hyperplane is

$$D = \{\mathbf{x} \in X \mid \ll \mathbf{w}, \mathbf{x} \gg = 1\}$$

We can additionally take $\|\mathbf{w}\| = 1$ by the scaling freedom (Lecture 1). Here

$$\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\ll \mathbf{w}, \mathbf{x} \gg -1) \mid \ll \mathbf{w}, \mathbf{x} \gg = 1\}.$$

We believe that there exists a a hyperplane so that $\mathcal{S}$ is linearly separable (Lecture 1). The Bayesian standpoint asks for a prior probability distribution on $\mathcal{H}$. By the nature of the situation it suffices to put a prior on vectors $\mathbf{w}$ such that $\ll \mathbf{w}, \mathbf{w} \gg = 1$, or, equivalently a probability distribution on the vector $\mathbf{c}$ closest to origin. We can take a random vector $\mathbf{u}$ in the unit circle, and r uniformly distributed in $[0, a]$, or $\mathbf{c} = r\mathbf{u}$. Thus our prior expresses the judgement that the separating hyperplane is no further than the distance a from the origin.

The parametric model family is given by

$$P(y^{(i)} = y|\mathbf{x}^{(i)}, u, \mathbf{w}) = \begin{cases} 1 & \text{if } yu(\ll \mathbf{w}, \mathbf{x}^{(i)} \gg -1) > 0 \\ 0 & \text{if } yu(\ll \mathbf{w}, \mathbf{x}^{(i)} \gg -1) < 0 \end{cases}$$

where $u \in \{-1,+1\}$ and $\mathbf{w}$ are unknown parameters of the model. The likelihood function is

$$\prod_{i=1}^{n} P(y^{(i)} = y|\mathbf{x}^{(i)}, u, \mathbf{w})$$

$$= \begin{cases} 1 & \text{if } yu(\ll \mathbf{w}, \mathbf{x}^{(i)} \gg -1) > 0 \text{ for } i = 1, \ldots, n \\ 0 & \quad \text{if } yu(\ll \mathbf{w}, \mathbf{x}^{(i)} \gg -1) < 0 \text{ otherwise.} \end{cases}$$

The posteriori for $u \in \{-1,+1\}$ and $\mathbf{w}$ is the same as their prior except that the parameter values incompatible with the data are eliminated. After normalizing the posterior probabilities, the parameter values compatible with the data will have higher probability than they did in the prior.

The predictive probability of y^* with $\mathbf{x}^*$ is

$$P_{\mathbf{x}^{ast}}(y^* = +1 \mid \mathcal{S}) = \int \sum_{u=\pm 1} P_{\mathbf{x}^{ast}}(y^* = +1 \mid u, \mathbf{w}) P(u, \mathbf{w} \mid \mathcal{S}) d\mathbf{w}.$$

Using a sample of m values from the posterior $(\mathbf{w}^{(i)}, u^{(i)}), i = 1, ..., m\}$ we can approximate this as

$$P_{\mathbf{x}^{ast}}(y^* = +1 \mid \mathcal{S}) \approx \frac{1}{m} \sum_{j=1}^{m} P_{\mathbf{x}^{ast}}(y^* = +1 \mid u^{(i)}, \mathbf{w}^{(i)}).$$

The average is just the fraction of lines drawn from the posterior that would put (x^{ast}, y^*) in the class $+1$.

The Bayes point machine $\mathbf{w}_{\mathrm{Bp}}$ would seem to correspond the vector maximizing the posterior on $\mathbf{w}$.

Some simulations....(in progress).

Appendices for Technical Details

11 About Dirichlet Densities

11.1 Euler's gamma function

The *gamma* function $\Gamma(z)$ is defined for complex numbers z, whose real part is positive, by the definite integral

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx. \tag{A.1}$$

A special case, obtained by the substitution $x = u^2/2$ is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

The recursion formula is

$$\Gamma(z) = (z-1)\Gamma(z-1). \tag{A.2}$$

Hence, if $z = n$, where n is a positive integer, we have the factorial

$$\Gamma(n) = (n-1)!. \tag{A.3}$$

11.2 The Dirichlet density

Let $\mathbf{\Theta} \subset R^L$ be the *simplex*

$$\mathbf{\Theta} = \left\{ (\theta_1, \ldots, \theta_L) \,|\theta_i \geq 0, i = 1, \ldots, L, \sum_{i=1}^{L} \theta_i = 1 \right\}. \tag{A.4}$$

Let for $\alpha_i > 0$

$$\phi\left(\theta_1, \ldots, \theta_L\right) = \begin{cases} \frac{\prod_{i=1}^{L} \theta_i^{\alpha_i - 1}}{Z}, & \text{if } \theta_1, \ldots, \theta_L \in \mathbf{\Theta} \\ 0 & \text{otherwise.} \end{cases} \tag{A.5}$$

Here

$$\frac{1}{Z} = \frac{\Gamma\left(\sum_{i=1}^{L} \alpha_i\right)}{\prod_{i=1}^{L} \Gamma\left(\alpha_i\right)}. \tag{A.6}$$

The density $\phi(\theta_1, \ldots, \theta_L)$ is called a *Dirichlet density.* We designate it symbolically by

$$Dir(\alpha_1, \ldots, \alpha_L). \tag{A.7}$$

If $\alpha_1 = \alpha_2 = \ldots = \alpha_L = \kappa$, then we talk about a *symmetric Dirichlet density.* For the proof of the fact that

$$\int_{\mathbf{\Theta}} \phi(\theta_1, \ldots, \theta_L)\, d\theta_1 \ldots d\theta_L = 1. \tag{A.8}$$

This means also that

$$\int_{\mathbf{\Theta}} \prod_{i=1}^{L} \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_L = \frac{\prod_{i=1}^{L} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{L} \alpha_i\right)}. \tag{A.9}$$

(Gupta and Richards 1987) is a concise compendium of knowledge about the Dirichlet distributions.

11.3 Beta density

As a special case for $L = 2$ we obtain in (A.9) the *Beta integral*

$$\int_0^1 \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = \frac{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}. \tag{A.10}$$

Thus

$$f(\theta) = \begin{cases} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{A.11}$$

is a probability density called the *Beta density* and denoted by

$$\mathcal{B}e(\theta; \alpha_1, \alpha_2).$$

Note the difference in the heuristic notation between Beta and Bernoulli $Be(p)$. If $\theta = (\theta_1, \ldots, \theta_L)$ is a random variable that assumes values in $\mathbf{\Theta}$ in (A.4) and has the symmetric $Dir(\alpha, \ldots, \alpha)$ distribution, then the marginal density of any θ_i is given by

$$\theta_i \in \mathcal{B}e(\theta; \alpha, (L-1)\alpha). \tag{A.12}$$

12 Maximum Likelihood for a Finite Table of Probabilities

The maximum likelihood estimate of $\underline{\theta}$ (a finite table of probabilities) is by a familiar principle given by

$$\widehat{\underline{\theta}}_{ML} = \mathrm{argmax}_{\underline{\theta}\in\mathbf{\Theta}} P\left(\mathbf{x} \mid \underline{\theta}\right) = \mathrm{argmax}_{\underline{\theta}\in\mathbf{\Theta}} \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

The presence of $\mathbf{\Theta}$ imposes a constrained problem of maximization. We take the natural logarithm of $P\left(\mathbf{x} \mid \underline{\theta}\right)$, which gives us the *loglikelihood function*

$$l\left(\theta_1, \theta_2, \ldots, \theta_L\right) = \log P\left(\mathbf{x} \mid \underline{\theta}\right)$$

We may equivalently seek the maximum of $l\left(\theta_1, \theta_2, \ldots, \theta_L\right)$. Since the constraint $\theta_1 + \theta_2 + \ldots + \theta_L = 1$ must be met, we consider the new auxiliary function in $L-1$ free variables

$$\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = l\left(\theta_1, \theta_2, \ldots, 1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)\right).$$

This gives

$$\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = n_1 \cdot \log\theta_1 + n_2 \cdot \log\theta_2 + \ldots + n_L \cdot \log\left(1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)\right).$$

Vi differentiate partially $\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right)$ with respect to θ_1, $\theta_2, \ldots,$ θ_{L-1} and set the partial derivatives equal to zero. This gives us the system of equations

$$\frac{\partial}{\partial\theta_1}\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = \frac{n_1}{\theta_1} - \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)} = 0,$$

$$\vdots$$

$$\frac{\partial}{\partial\theta_{L-1}}\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = \frac{n_{L-1}}{\theta_{L-1}} - \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)} = 0.$$

This leads to the equalities

$$\frac{n_1}{\theta_1} = \frac{n_2}{\theta_2} = \ldots = \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)}.$$

Let us denote the common value of these ratios as λ so that

$$\theta_1 = \frac{n_1}{\lambda}, \theta_2 = \frac{n_2}{\lambda}, \ldots, \theta_L = \frac{n_L}{\lambda}.$$

We determine λ from the constraint $\theta_1 + \theta_2 + \ldots + \theta_L = 1$, which gives

$$1 = \theta_1 + \theta_2 + \ldots + \theta_L = \frac{n_1}{\lambda} + \frac{n_2}{\lambda} + \ldots + \frac{n_L}{\lambda}$$

or

$$\lambda = n_1 + n_2 + \ldots + n_L = n.$$

Hence we have obtained the solution to $\nabla \widetilde{l}\,(\theta_1, \theta_2, \ldots, \theta_{L-1}) = 0$ written in a component wise form as

$$\widehat{\theta}_i = \frac{n_i}{n}, i = 1, \ldots, L.$$

Strictly taken we have yet to prove that this yields a maximum. For this we could check the matrix of second order partial derivatives of $\widetilde{l}$, (Khuri 1993 p. 283). There is a more instructive way to prove that the estimate found above actually gives the maximum. In fact the proof of the next proposition shows that it is not even necessary to differentiate to prove that we have found the maximum likelihood estimate.

Proposition 12.1 *The maximum likelihood estimate $\widehat{\theta}_{ML}$ of $\underline{\theta}$ is*

$$\widehat{\underline{\theta}}_{ML} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_L}{n} \right).$$

Proof: Clearly the candidate solution $\widehat{\underline{\theta}}_{ML}$ belongs to $\mathbf{\Theta}$ and is thus admissible. Since $P\left(\mathbf{x} \mid \underline{\theta}\right) = \prod_{i=1}^{L} \theta_i^{n_i}$, the following identity is evident

$$H\left(\widehat{\underline{\theta}}_{ML}\right) = -\frac{1}{n} \log P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right), \tag{B.1}$$

where

$$H\left(\widehat{\underline{\theta}}_{ML}\right) = -\sum_{i=1}^{L} \widehat{\theta}_i \log \widehat{\theta}_i \tag{B.2}$$

is the (empirical) Shannon's entropy in nats.

Next we take an arbitrary $\underline{\theta}$ in $\mathbf{\Theta}$. Then we have in view of (B.2) for an arbitrary $\underline{\theta}$ in $\mathbf{\Theta}$ another evident identity

$$P\left(\mathbf{x} \mid \underline{\theta}\right) = \prod_{i=1}^{L} \theta_i^{n_i} = e^{-n\left(D\left(\widehat{\underline{\theta}}_{ML} | \underline{\theta}\right) + H\left(\widehat{\underline{\theta}}_{ML}\right)\right)}. \tag{B.3}$$

Here we have used the Kullback distance between two discrete probability distributions defined as

$$D\left(f|g\right) = \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}.$$

Thus from (B.1) and (B.3)

$$\frac{P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right)}{P\left(\mathbf{x} \mid \underline{\theta}\right)} = e^{-nH\left(\widehat{\underline{\theta}}_{ML}\right)} \cdot e^{n\left(D\left(\widehat{\underline{\theta}}_{ML}|\underline{\theta}\right)+nH\left(\widehat{\underline{\theta}}_{ML}\right)\right)} =$$

$$= e^{nD\left(\widehat{\underline{\theta}}_{ML}|\underline{\theta}\right)} \geq 1,$$

where the last inequality follows due to the fact that $D\left(\widehat{\underline{\theta}}_{ML}|\underline{\theta}\right)$ is the Kullback distance, which is known to be nonnegative. Equality holds if and only if $\widehat{\underline{\theta}}_{ML} = \underline{\theta}$. Thus

$$P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right) \geq P\left(\mathbf{x} \mid \underline{\theta}\right)$$

for every $\underline{\theta}$ in $\mathbf{\Theta}$ and the assertion is proved. ■

13 Kullback Distance

13.1 Entropy, shannon and dHart

We define the *entropy* of a discrete r.v. X as

$$H(X) := -\sum_{i=1}^{L} f_X(x_i) \log(f_X(x_i)) \tag{C.1}$$

This has the dimension [bits/symbol] if logarithms to the base 2 are used.

Example 13.1 (Binary entropy function) For the special case $\mathcal{X} = \{x_1, x_2\}$, with $p := f_X(x_1)$,

$$h(p) := -p \log_2(p) - (1-p) \log_2(1-p) \tag{C.2}$$

is the *(binary) entropy function*. This is also the entropy of a Bernoulli random variable $X \in Be(p)$ with $\mathcal{X} = \{0, 1\}$). The unit of information is *shannon* and corresponds to the binary entropy of $X \in Be(1/2)$. The shannon (symbol Sh) is a unit defined by IEC 80000-13[2].

■

Example 13.2 (Uniform Distribution on the Ten Digits) Let X assume the ten digits as values with equal probabilities. Then

$$H(X) = -10\left(\frac{1}{10}\log_2\frac{1}{10}\right) = \log_2 10 = 3.32[\mathrm{Sh}] \tag{C.3}$$

■

dHart is a measure information or entropy, using logarithms on the base 10. Hence, when X assumes the ten digits as values with equal probabilities, i.e. $f_X(x_i) = \frac{1}{10}$ for $i = 1, \ldots, 10$, the dHart is

$$\mathrm{dHart}(X) = -\sum_{i=1}^{L} f_X(x_i)\log_{10}(f_X(x_i)) = \log_{10} 10 = 1.$$

Thus one dHart equals 3.32 [Sh].

13.2 Definition and Examples

The entropy $H(X)$ is a measure of the uncertainty in bits (=binary information units) of the random variable X. It is also a lower bound for the number of bits (binary digits) needed on the average to describe the random variable.

Now we introduce the *relative entropy* based on the distribution $g(x)$ when the "true" distribution is $f(x)$. The relative entropy is then defined as follows.

Definition 13.1 Let $\mathcal{X} = \{x_1, \cdots, x_L\}$ be an alphabet and let

$$\mathbf{f} := (f(x_1), \cdots, f(x_L))$$

and

$$\mathbf{g} := (g(x_1), \cdots, g(x_L))$$

[2] IEC 80000-13 is an international standard that defines quantities and units used in information science, and specifies names and symbols for these quantities and units.

be two probability distributions defined on $\mathcal{X}$. Then their *relative entropy* or the *Kullback distance* between $\mathbf{f}$ and $\mathbf{g}$ is defined by

$$D(\mathbf{f} \mid \mathbf{g}) = \sum_{i=1}^{L} f(x_i) \log \frac{f(x_i)}{g(x_i)}. \tag{C.4}$$

■

Here we use the conventions $0 \cdot \log \frac{0}{g(x_i)} = 0$ and $f(x_i) \log \frac{f(x_i)}{0} = \infty$. The logarithm is the natural logarithm unless otherwise stated. Note that this is not a distance in the sense of being a metric, since e.g. the symmetry property of a metric obviously need not hold.

Example 13.3 (Information Content) If X is a random variable with the distribution $\mathbf{f} = (f(x_1), \cdots, f(x_L))$, any probability distribution on an alphabet of L symbols, and if $\mathbf{g} = (1/L, \cdots, 1/L)$ is the uniform distribution, then

$$D(\mathbf{f} \mid \mathbf{g}) = \log L - H(X). \tag{C.5}$$

This quantity is sometimes known as the *information content (of* $\mathbf{f}$*)*.

■

Example 13.4 (Two Bernoulli distributions) Let $\mathcal{X} = \{0, 1\}$ and $0 \leq p \leq 1$ and $0 \leq g \leq 1$. Let $\mathbf{f} = (1-p, p)$ and $\mathbf{g} = (1-g, g)$ be the two Bernoulli distributions $Be(p)$ and $Be(g)$, respectively. Then

$$D(\mathbf{f} \mid \mathbf{g}) = (1-p) \cdot \log \frac{1-p}{1-g} + p \cdot \log \frac{p}{g}. \tag{C.6}$$

We can also rewrite this as

$$D(\mathbf{f} \mid \mathbf{g}) = -(1-p) \cdot \log(1-g) - p \cdot \log g - h(p),$$

where $h(p)$ is the entropy function (C.2) in natural logarithm.

■

13.3 Properties

Next we give a general proof of the important fact that the relative entropy is nonnegative.

Proposition 13.5 *For any probability distributions* **f** *and* **g** *on the same alphabet*

$$D\left(\mathbf{f} \mid \mathbf{g}\right) \geq 0. \tag{C.7}$$

Proof: Let X be a random variable that has the distribution $\mathbf{f}$. We write $p(x)$ and $g(x)$ for the generic values of the probability of $x \in \mathcal{X}$ in the two distributions. Then we have

$$D\left(\mathbf{f} \mid \mathbf{g}\right) = E\left[\log \frac{p\left(X\right)}{g\left(X\right)}\right]$$

and this equals

$$D\left(\mathbf{f} \mid \mathbf{g}\right) = -E\left[\log \frac{g\left(X\right)}{p\left(X\right)}\right].$$

Since $\phi(x) = -\log x$ is a convex function we have that

$$-E\left[\log \frac{g\left(X\right)}{p\left(X\right)}\right] \geq -\log E\left[\frac{g\left(X\right)}{p\left(X\right)}\right],$$

where we have used Jensen's inequality. But

$$E\left[\frac{g\left(X\right)}{p\left(X\right)}\right] = \sum_{i=1}^{L} f(x_i)\frac{g(x_i)}{f(x_i)} = 1$$

and since $\log 1 = 0$, we have proved our assertion. ■

Since $D\left(\mathbf{f} \mid \mathbf{g}\right) \geq 0$ we get by (C.5) that

$$H(X) \leq \log L.$$

13.4 The marginal data likelihood and the Kullback Distance, Aitchison (1975)

Let

$$R_{\pi}\left(q\right) = \int_{\Theta} D\left(\prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) \mid q\right) d\pi(\theta)$$

and define

$$\mathbf{m}^* = \arg\min_q R_w\left(q\right). \tag{C.8}$$

Then the marginal data likelihood is the minimizer, or,

$$m^*\left(x^{(1)}, \ldots, x^{(N)}\right) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) d\pi(\theta). \tag{C.9}$$

To prove this we write

$$R_\pi\left(q\right) = \int_{\boldsymbol{\Theta}} D\left(\prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) \mid \mathbf{m}^*\right) d\pi(\theta)$$

$$+ \int_{\boldsymbol{\Theta}} \int_{x \in \mathcal{X}^N} \prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) \ln\left[\frac{\mathbf{m}^*(x)}{q(x)}\right] dx d\pi(\theta),$$

$$= \int_{\boldsymbol{\Theta}} D\left(\prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) \mid \mathbf{m}^*\right) d\pi(\theta) + \int_{x \in \mathcal{X}^N} \left[\int_{\boldsymbol{\Theta}} \prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) d\pi(\theta)\right] \ln\left[\frac{\mathbf{m}^*(x)}{q(x)}\right] dx$$

$$= \int_{\boldsymbol{\Theta}} D\left(\prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right) \mid \mathbf{m}^*\right) d\pi(\theta) + \int_{x \in \mathcal{X}^N} \mathbf{m}^*(x) \ln\left[\frac{\mathbf{m}^*(x)}{q(x)}\right] dx.$$

Thus the conclusion follows. ■

14 Asymptotic Shape of the Likelihood

Parametric Statistical Model, n I.I.D. $|\theta$ rv's

$$x_i|\theta \in f\left(x|\theta\right), \text{ I.I.D. },$$

or independent, identically, distributed conditional on θ

$$x^{(n)} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$$

$f\left(x|\theta\right)$ is a probability density on R^p. $f\left(x|\theta\right)$ is a known function of x and θ. Let us assume that $f(x|\theta)$ is a density with a scalar parameter (for simplicity of notation), and that $f(x|\theta)$ is some $k \geq 2$ times differentiable in θ. We let $\widehat{\theta}_{ML}$ be the maximum likelihood estimate of θ. We expand the log likelihood function around $\widehat{\theta}_{ML}$

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + \left(\theta - \widehat{\theta}_{ML}\right) \frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right)$$

$$+\frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n(\theta)$$

But here $\widehat{\theta}_{ML}$ is a solution of the equation

$$\frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = 0$$

Hence

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n(\theta).$$

We have by assumption of conditionally I.I.D. data

$$\frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\theta\right) = \sum_{l=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_l|\theta\right).$$

We set $Y_l = \frac{d^2}{d\theta^2} \log f\left(x_l|\theta\right)$. Then the Law of Large Numbers says that

$$\frac{1}{n}\sum_{l=1}^{n} Y_l \to E\left[Y\right], \quad \text{as } n \to \infty,$$

where

$$E\left[Y\right] = \int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log f\left(x|\theta\right) f\left(x \mid \theta\right) dx.$$

The integral

$$I\left(\theta\right) = -\int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log f\left(x|\theta\right) f\left(x \mid \theta\right) dx$$

is called the *Fisher information.* Then we may feel inclined to take that

$$\frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = \sum_{l=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_l|\widehat{\theta}_{ML}\right) \approx -n \cdot I\left(\widehat{\theta}_{ML}\right).$$

Note that even $\widehat{\theta}_{ML}$ depends on n. This gives

$$\log f\left(x^{(n)}|\theta\right) \approx \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) - \frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 n \cdot I\left(\widehat{\theta}_{ML}\right).$$

The first term does not involve θ. Then

$$f\left(x^{(n)}|\theta\right) \approx e^{-\frac{n}{2}\left(\theta-\widehat{\theta}_{ML}\right)^2 \cdot I\left(\widehat{\theta}_{ML}\right)}.$$

The interpretation of the relation is that the likelihood function can be for large n be approximated by a normal density for which the mean is $\widehat{\theta}_{ML}$ and the variance is $\frac{1}{nI\left(\widehat{\theta}_{ML}\right)}$.

15 References and further reading:

1 Journal articles and technical reports on Bayesian learning/machine learning and Dirichlet distribution:

- J. Aitchison (1975): Goodness of prediction fit. *Biometrika*,62, pp. 547−181.
- W.L. Buntine (1992) *A theory of learning classification rules*, PhD Thesis. School of Computing Science, Sydney University of Technology, Sydney.
- P. Cheeseman (1988): An Inquiry into Computer Understanding. *Computational Intelligence*, **4**, 58−66.
- J.M. Dickey (1983): Multiple Hypergeometric Functions: Probabilistic Interpretation and Statistical Uses. *Journal of the American Statistical Association*, 78, pp. 628−637.
- R.D. Gupta and D.St.P. Richards (1987): Multivariate Liouville Distributions. *Journal of Multivariate Analysis*, 23, pp. 232−256.
- D. Heckerman (2008): A tutorial on learning with Bayesian networks, in *Innovations in Bayesian Networks*, pp. 33−82, Springer, Berlin.
- R. Herbrich, T. Graepel and C. Campbell (2001):Bayes point machines. *The Journal of Machine Learning Research*, 1, 245−279,
- D.J.C. MacKay (1992): *Bayesian methods for adaptive models*, PhD Thesis, California Institute of Technology.
- D.V. Lindley (1970): A non-frequentist view of probability and statistics. *The Teaching of Probability & Statistics*, L. Råde editor, Almqvist & Wicksell, Uppsala, pp. 209−222.
- J. Rissanen (1997): Stochastic complexity and learning. *Journal of Computer and System Sciences*, 55, pp. 89−95.

- H.V. Roberts (1965): Probabilistic Prediction. *Journal of the American Statistical Association*, 60, pp. 50−62.

- M. Sunnåker, A. Busetto, G. Alberto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, (2013): Approximate bayesian computation. *PLoS Computational Biology*, 9, pp. e1002803

- M. E. Tipping: Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. pp. 49−70, in O. Bousquet, U. von Luxburg, G. Rätsch: (ed.s) *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence Vol. 3176, 2004.

- E.B. Wilson (1927): Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, pp. 209−212.

2 Books:

- D. Barber (2012): *Bayesian Reasoning and Machine Learning.* Cambridge University Press, Cambridge, U.K.

- J.M. Bernardo and A.F.M. Smith (1994): *Bayesian Theory.* John Wiley and Sons, Chichester, New York, Brisbane, Toronto and Singapore.

- C.M. Bishop (2006): *Pattern recognition and machine learning*, 2006, Springer Berlin.

- P. Gregory (2005): *Bayesian Logical Data Analysis for the Physical Sciences.* Cambridge University Press, Cambridge, U.K..

- E.T. Jaynes (2003): *Probability theory: the logic of science*, Cambridge University Press, Cambridge, U.K.

- D. Lunn, C. Jackson, N. Best, A. Thomas and D. Spiegelhalter (2013): *The BUGS book: A practical introduction to Bayesian analysis*, CRC Press, Boca Raton, London, New York.

- D.J. MacKay (2003): *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, U.K..

- R. v. Mises with H. Geiringer (1964): *Mathematical Theory of Probability and Statistics.* Academic Press, New York and London.

3 Software:

- BUGS (Bayesian inference Using Gibbs Sampling), software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (McMC) methods. `http://www.mrc-bsu.cam.ac.uk/software/bugs/`