Statistical Bioinformatics, Makerere Learning & Measures of Information

Timo Koski

TK

02.08.2018

KTH Matematik

Lecture

The lecture parallells the topics in Chapters 3 and 8 in Ewens and Grant but concentrates more on Bayesian learning from data (Ewens and Grant 3.4, 3.6−3.7). In addition we deal with information theoretic measures of distance between probability distributions, (Ewens and Grant 1.14).
See also J.S. Liu and C.E. Lawrence: Bayesian inference on biopolymer models. *Bioinformatics*, 15, 1999, pp. 38−52.

KTH Matematik

Lecture 4: Contents

1) Learning from data, Models, Bayes rule
2) Posterior densities, maximum likelihood,
3) Likelihood ratio, Model Choice
4) Entropy, Kullback Distance, and Sequence Logos

KTH Matematik

Learning from data

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.

KTH Matematik

Learning and Bayes' Rule

Since

$$p(X \mid Y) \cdot p(Y) = p(Y \mid X) \cdot p(X)$$

we have in a formal way

$$p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)}.$$

The marginal distribution $p(Y)$ is

$$p(y_j) = \sum_{i=1}^{L} p(y_j \mid x_i) p(x_i).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Learning and Bayes' Rule

$$p(x_i \mid y_j) = \frac{p(y_j \mid x_i) \cdot p(x_i)}{\sum_{i=1}^{L} p(y_j \mid x_i) p(x_i)}.$$

Bayes' rule gives a fundamental operation for *up-date of probability distributions* in response to observed information. The rule shows how knowledge about the occurrence of the event $Y = y_j$ is to be used to transform probabilities on $\mathcal{X}$.

KTH Matematik

Terminology for Bayes' Rule

- $p(X)$: A **Prior Distribution** on $\mathcal{X}$.
- $p(X \mid Y)$: A **Posterior Distribution** on $\mathcal{X}$.

If X and Y are independent, then the prior distribution and posterior distribution are identical and there is no *learning*.

KTH VETENSKAP OCH KONST
KTH Matematik

Thomas Bayes 1702-1761

KTH Matematik

Learning from data

The intended learning is done by use of 'MODELS', which serve as the language in which the constraints predicated on the data can be described. We shall in this lecture talk about parametric probabilistic models.

Modeling and Learning for Tosses of a Thumbtack

Let us consider a sequence of flips of a thumbtack (D. Heckerman). If we throw a thumbtack in the air, it will come to rest either on its point (0) or on its head (1).

KTH VETENSKAP OCH KONST
KTH Matematik

Modeling and Learning for Tosses of a Thumbtack

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.

KTH Matematik

Modeling and Learning for Tosses of a Thumbtack

We let $\mathbf{x}$ denote the sequence of outcomes of the flips, $\mathbf{x} = x_{i_1} x_{i_2} \dots x_{i_n}$, $x_{i_l} \in \{0, 1\}$. Let now $\mathbf{P}$ be a random variable, whose values are numbers, denoted by p, between zero and one, $0 \leq p \leq 1$.

KTH Matematik

MODEL FAMILY for Tosses of a Thumbtack

Conditioned on $\mathbf{P} = p$, the digits in $\mathbf{x}$ are outcomes of I.I.D $\mathrm{Be}(p)$ R.V's. Hence a model in the family is given by the probability assignment

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

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

if $\sum_{l=1}^{n} x_{i_l} = k$.

KTH VETENSKAP OCH KONST
KTH Matematik

MODEL FAMILY for Tosses of a Thumbtack

The goal is to find the model (within a preestablished family) that is best in some sense given some observed data. In the thumbtack example we understand this 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.

KTH Matematik

The Prior

To progress we express our uncertainty about $\mathbf{P}$ using a probability density function $f_{\mathbf{P}}(p)$, which is called the *prior*. Formally this means

$$f_{\mathbf{P}}(p) \geq 0, 0 \leq p \leq 1,$$

and $f_{\mathbf{P}}(p) = 0$ elsewhere, and

$$\int_0^1 f_{\mathbf{P}}(p)\, dp = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

The Posterior Density

By an extension of Bayes' rule to continuous random variables we get the *posterior density*

$$f_{\mathbf{P}|\mathbf{x}}(p \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \mathbf{P} = p) \cdot f_{\mathbf{P}}(p)}{\int_0^1 P(\mathbf{x} \mid \mathbf{P} = p) \cdot f_{\mathbf{P}}(p)\, dp}, 0 \leq p \leq 1$$

and zero elsewhere. Due to the standardization $f_{\mathbf{P}|\mathbf{x}}(p \mid \mathbf{x})$ is another probability density for $\mathbf{P}$.

KTH Matematik

The Posterior Density

The posterior $f_{\mathbf{P}|\mathbf{x}}(p \mid \mathbf{x})$ expresses our updated belief in the statement that p is the 'true' chance of obtaining heads given that we have observed $\mathbf{x}$.

KTH Matematik

The Posterior Density

One way to get further from here is to use an explicit form for $f_{\mathbf{P}}(p)$. There could be several choices, but some are at least analytically more advantageous. Let us consider $\mathbf{P} \in \mathrm{U}(0,1)$. i.e.,

$$f_{\mathbf{P}}(p) = \begin{cases} 1 & 0 \leq p \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

The Posterior Density: the constant

By an insertion we can now calculate

$$\int_0^1 P\left(\mathbf{x} \mid \mathbf{P} = p\right) \cdot f_{\mathbf{P}}\left(p\right) dp = \int_0^1 p^k \cdot \left(1-p\right)^{n-k} dp = \frac{k!(n-k)!}{(n+1)!}$$

by the properties of the *Beta density* (see Ewens and Grant p.32).

Recall: Beta Density

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

is a probability density. Hence

$$\int_0^1 f_{\mathbf{P}}(p)dp = 1 \Leftrightarrow \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Recall also that $\Gamma(k+1) = k!$, if k is a positive integer.

KTH VETENSKAP OCH KONST
KTH Matematik

The Posterior Density: the constant

$$\int_0^1 P\left(\mathbf{x} \mid \mathbf{P} = p\right) \cdot f_{\mathbf{P}}\left(p\right) dp = \int_0^1 p^k \cdot (1-p)^{n-k}\, dp = \frac{k!(n-k)!}{(n+1)!}.$$

KTH Matematik

The Posterior Density for n tosses of a Thumbtack

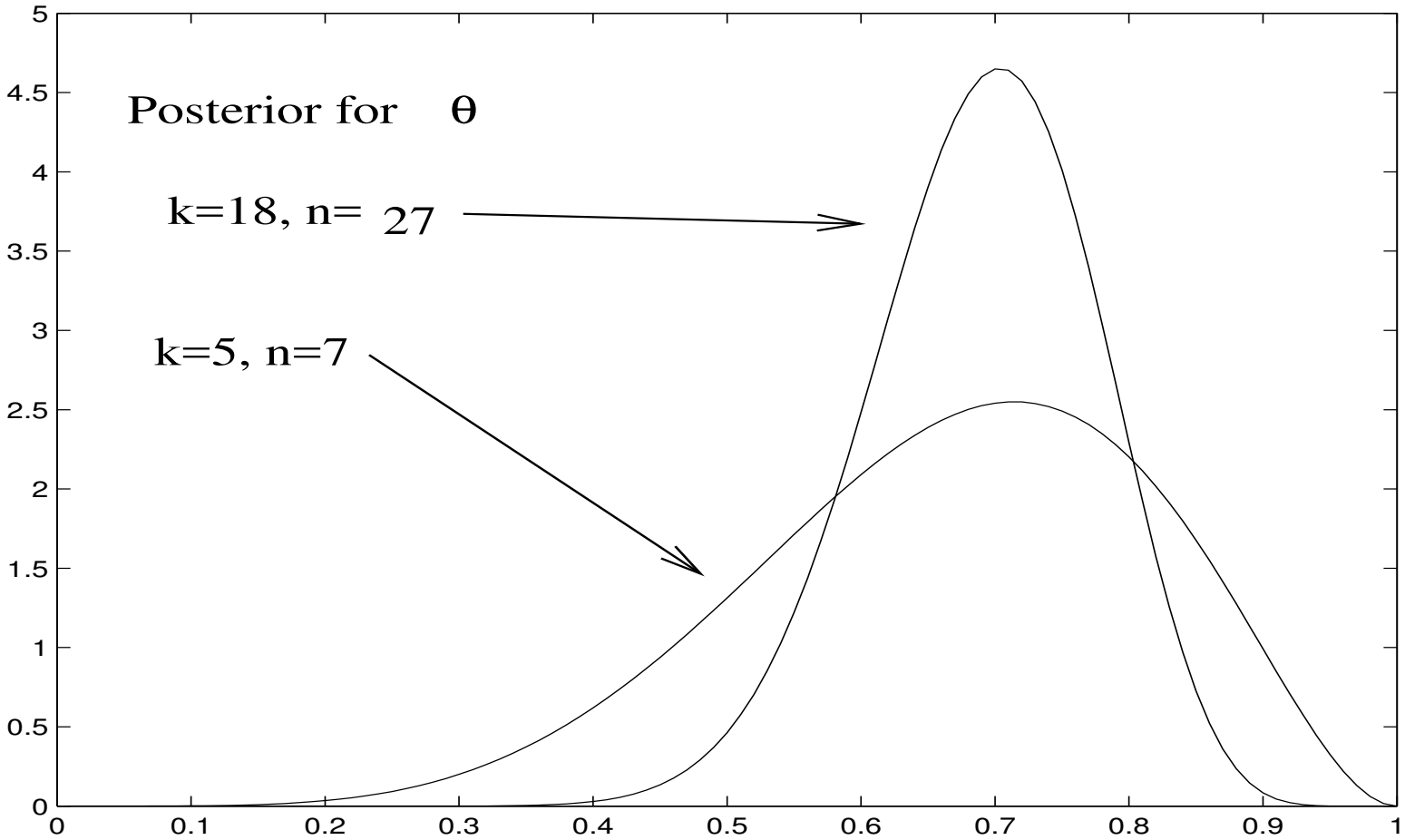
Then we have

$$f_{\mathbf{P}|\mathbf{x}}(p \mid \mathbf{x}) = \begin{cases} \frac{(n+1)!}{k!(n-k)!} \cdot p^k (1-p)^{n-k} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is again recognized as a Beta density.

KTH Matematik

Posterior Densities for p in Be(p)



KTH Matematik

The Maximum Likelihood Estimate

We introduce the *maximum likelihood estimate* $\widehat{p}_{ML}$ *of* p by

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

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

KTH Matematik

The Maximum Likelihood Estimate

The probability $P(\mathbf{x} \mid \mathbf{P} = p)$ regarded as a function of p is known as the *likelihood function*

$$L(p) = P(\mathbf{x} \mid \mathbf{P} = p).$$

The likelihood function $L(p)$ thus compares the plausibilities of different models for given $\mathbf{x}$.

KTH VETENSKAP OCH KONST
KTH Matematik

The Maximum Likelihood Estimate

A straightforward maximization of the likelihood function gives

$$\widehat{p}_{ML} = \frac{k}{n}.$$

KTH Matematik

The Predictive Estimate of 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 modeling the next toss, given n flips of the thumbtack as recorded in $\mathbf{x}$.

KTH Matematik

The Predictive Estimate of the Outcome of the Next Toss

Using the model family above we have

$$P\left(X_{n+1} = \text{head} | \mathbf{x}\right) = \int_0^1 p f_{\mathbf{P}|\mathbf{x}}\left(p \mid \mathbf{x}\right) dp$$

$$= \frac{(n+1)!}{k!(n-k)!} \int_0^1 p \cdot p^k \left(1-p\right)^{n-k} dp$$

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

KTH VETENSKAP OCH KONST
KTH Matematik

Pierre-Simon Laplace, 1749 – 1827

The probability

$$P\left(X_{n+1} = \text{head}|\mathbf{x}\right) = \frac{k+1}{n+2}$$

is known as Laplace's rule of succession.

KTH VETENSKAP OCH KONST
KTH Matematik

General Summary

A formal Bayesian modeling articulates the information in a sequence with evidence other than that of the sequence. The evidence is assessed by judgement and is expressed in probability theory terms:

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

(2) a prior expresses uncertainty about the parameters.

When (1) is combined with the data 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.

KTH Matematik

Likelihood Ratio

We wish to compare two different models, H_1 and H_0 , proposed for a given sequence $\mathbf{x}$.

H_1: the sequence is generated as I.I.D samples $X_1, \ldots, X_n$ of $\mathbf{f}$ (a density or a pmf).

H_0: the sequence is generated as I.I.D samples of $\mathbf{g}$ (a density or a pmf).

Then we can choose between H_1 and H_0 by using the likelihood ratio LR between the two models:

$$\mathrm{LR} = \frac{\mathbf{f}(X_1) \cdot \mathbf{f}(X_2) \cdot \ldots \cdot \mathbf{f}(X_n)}{\mathbf{g}(X_1) \cdot \mathbf{g}(X_2) \cdot \ldots \cdot \mathbf{g}(X_n)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Likelihood Ratio: Example

$\mathbf{x}$ is a binary sequence with k ones and $n-k$ zeros.

H_1: $\mathbf{f} \leftrightarrow Be(p)$

H_0: $\mathbf{g} \leftrightarrow Be(g)$

$$\mathrm{LR} = \frac{p^k(1-p)^{n-k}}{g^k(1-g)^{n-k}} = \left(\frac{p}{g}\right)^k \left(\frac{1-p}{1-g}\right)^{n-k}$$

KTH Matematik

Likelihood Ratio: Rule of Decision

Choose a threshold $c > 0$.

- If $LR > c$, then decide for H_1 as model for $\mathbf{x}$.
- If $LR = c$, no decision.
- If $LR < c$, then decide for H_0 as model $\mathbf{x}$.

c can be chosen so as to give some desirable level of error probability.

KTH VETENSKAP OCH KONST
KTH Matematik

Log Likelihood Ratio

$$\log \mathrm{LR} = \log \frac{\mathbf{f}(X_1)}{\mathbf{g}(X_1)} + \log \frac{\mathbf{f}(X_2)}{\mathbf{g}(X_2)} + \ldots + \log \frac{\mathbf{f}(X_n)}{\mathbf{g}(X_n)}$$

$$= \sum_{i=1}^{n} \log \frac{\mathbf{f}(X_i)}{\mathbf{g}(X_i)}$$

Or set $Z_i = \log \frac{\mathbf{f}(X_i)}{\mathbf{g}(X_i)}$ so that

$$\log \mathrm{LR} = \sum_{i=1}^{n} Z_i$$

KTH VETENSKAP OCH KONST
KTH Matematik

Likelihood Ratio: Example Continued

$$\mathrm{LR} = \frac{p^k(1-p)^{n-k}}{g^k(1-g)^{n-k}} = \left(\frac{p}{g}\right)^k \left(\frac{1-p}{1-g}\right)^{n-k}$$

$$\log \mathrm{LR} = k \log \log \frac{p}{g} + (n-k) \log \frac{1-p}{1-g}$$

KTH Matematik

Log Likelihood Ratio

$$\frac{1}{n} \log \mathrm{LR} = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

is the arichtmetic mean of I.I.D random variables $Z_1, \ldots, Z_n$.

Log Likelihood Ratio: Law of Large numbers

If

$$\mathbf{f} \stackrel{def}{=} (f(x_1), \cdots, f(x_L)), \quad \mathbf{g} \stackrel{def}{=} (g(x_1), \cdots, g(x_L))$$

and assume $\mathbf{f}$ is the 'true model', then the Law of Large Numbers entails

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \to E_{\mathbf{f}}(Z_1)$$

$$= \sum_{i=1}^{L} f(x_i) \log \frac{f(x_i)}{g(x_i)},$$

which will reappear below.

Likelihood Ratio: Example Continued

$$\frac{1}{n}\log \mathrm{LR} = \frac{k}{n}\log\frac{p}{g} + \frac{n-k}{n}\log\frac{1-p}{1-g}$$

$$= \widehat{p}_{ML}\log\frac{p}{g} + (1-\widehat{p}_{ML})\log\frac{1-p}{1-g}$$

$$\to p\log\frac{p}{g} + (1-p)\log\frac{1-p}{1-g}$$

KTH Matematik

Comparison of Model Families as Hypothesis Testing

We wish to compare two different model families proposed for a given sequence $\mathbf{x}$. Under the model family M_i the sequence is related to the parameters $\underline{\theta}_i$ by a distribution $P_i\left(\mathbf{x} \mid \underline{\theta}_i\right)$ and the prior densities for the parameters are $\phi_i\left(\underline{\theta}_i\right)$ for $i = 1, 2$. Then we can compare the two families by computing the ratio of the posterior probabilities $P\left(M_i \mid \mathbf{x}\right)$ of the two model families.

KTH Matematik

Comparison of Model Families as Hypothesis Testing

We do not want to compare individual models within the respective families. To avoid this we do **model averaging** by

$$q_i(\mathbf{x}) = \int P_i(\mathbf{x} \mid \underline{\theta}_i) \cdot \phi_i(\underline{\theta}_i)\, d\underline{\theta}_i, i = 1, 2.$$

The quantity $q_i(\mathbf{x})$ is sometimes referred to as the *evidence* for M_i.

KTH Matematik

Comparison of Model Families as Hypothesis Testing

Then Bayes' rule gives the ratio (*posterior odds*)

$$\frac{P\left(M_1 \mid \mathbf{x}\right)}{P\left(M_2 \mid \mathbf{x}\right)} = \frac{P\left(M_1\right)}{P\left(M_2\right)} \cdot \frac{q_1\left(\mathbf{x}\right)}{q_2\left(\mathbf{x}\right)},$$

where $P\left(M_i\right)$ is the prior for M_i for $i = 1, 2$.

KTH Matematik

Comparison of Model Families as Hypothesis Testing

The ratio

$$B(\mathbf{x}) = \frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}$$

is known as the *Bayes factor* and can be seen representing the weight of evidence in the sequence $\mathbf{x}$ in favour of model family M_1 against model family M_2.

KTH VETENSKAP OCH KONST
KTH Matematik

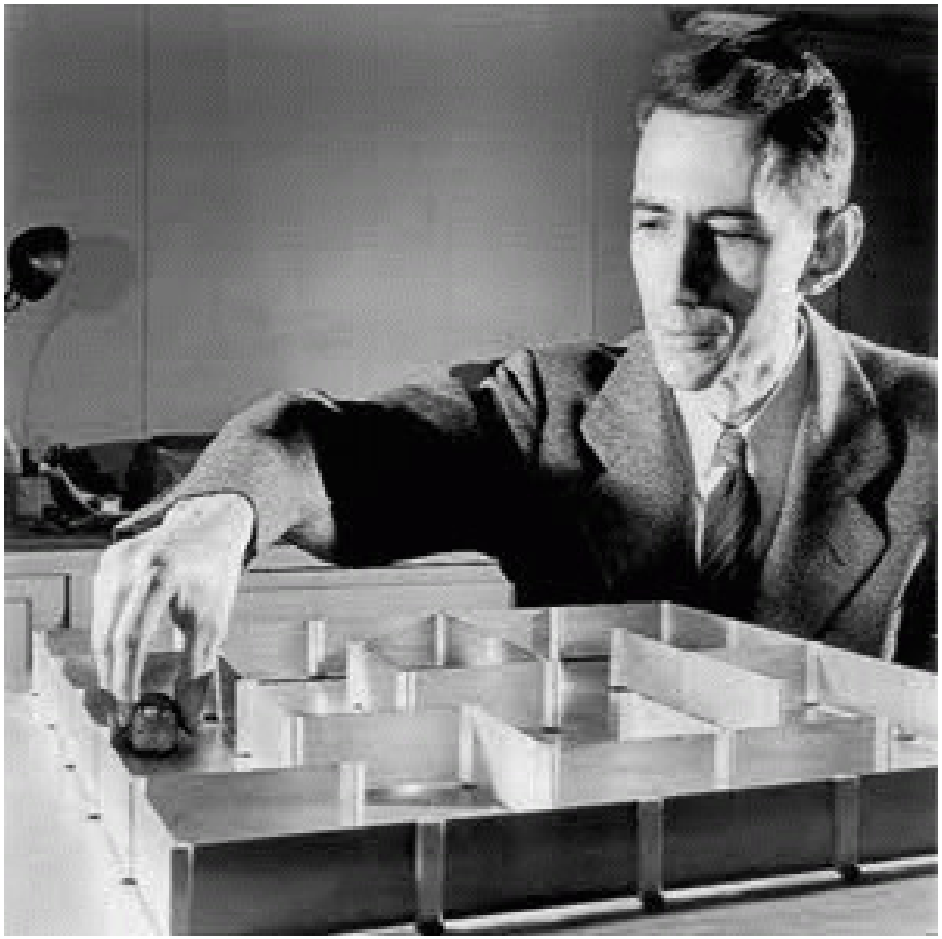
$-\log f_X$

There is a connection between learning and information (or surprise). This emerges by study of

$$-\log f_X$$

(-log likelihood)

Claude E. Shannon, 1916-2001



KTH Matematik

Entropy

We define the *entropy* of X as

$$H(X) \stackrel{def}{=} -\sum_{i=1}^{L} f_X(x_i) \log(f_X(x_i))$$

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

KTH VETENSKAP OCH KONST
KTH Matematik

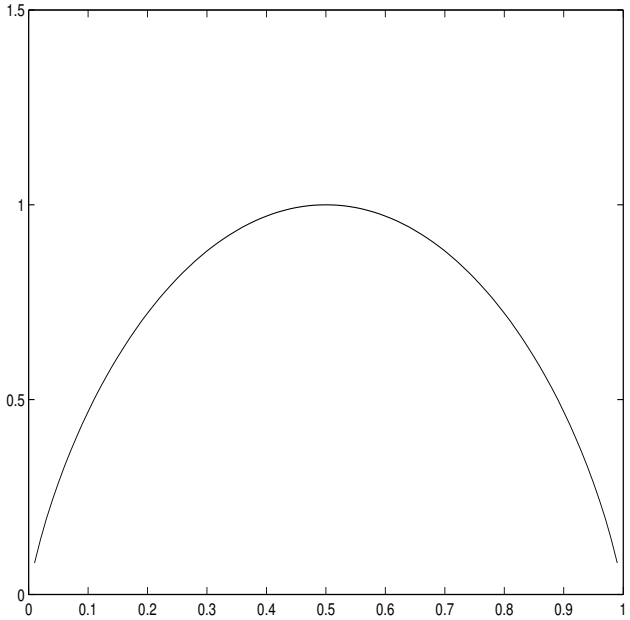
Binary entropy function

For the special case Bernoulli random variable $X \in Be(p)$ with $\mathcal{X} = \{0, 1\}$) $p = f_X(1)$,

$$h(p) \stackrel{def}{=} -p \log_2(p) - (1-p) \log_2(1-p)$$

is the *(binary) entropy function*.

Binary entropy function



$$h(p) = -p\log_2(p) - (1-p)\log_2(1-p)$$

$h\left(\frac{1}{2}\right) = 1$ defines the binary information unit **BIT**.

KTH VETENSKAP OCH KONST
KTH Matematik

Entropy of random DNA

Let X assume values in $\mathcal{X} = \{A, T, C, G\}$ and let $X \in U(1, 4)$. X is a nucleotide chosen at random. Then

$$H(X) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$= \log_2 4 = 2.$$

We think in this case of $H(X)$ as the number of yes/no questions needed to identify an outcome of X. The first question is "purine or pyriminide" followed by the second question "A or G " or " C or T" depending on the answer to the first question.

KTH VETENSKAP OCH KONST
KTH Matematik

Entropy

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.

KTH Matematik

Kullback Distance (Relative Entropy)

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

$$\mathbf{f} \stackrel{def}{=} (f(x_1), \cdots, f(x_L)), \quad \mathbf{g} \stackrel{def}{=} (g(x_1), \cdots, g(x_L))$$

be two probability distributions defined on $\mathcal{X}$. The *Kullback distance* between $\mathbf{f}$ and $\mathbf{g}$ is defined by

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

KTH VETENSKAP OCH KONST
KTH Matematik

Kullback Distance

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

Here we use the conventions $0 \cdot \log \frac{0}{g(x_i)} = 0$ and $f\left(x_i\right) \log \frac{f(x_i)}{0} = \infty$. The logarithm is the natural logarithm unless otherwise stated.

Kullback Distance

We have seen above that

$$\frac{1}{n} \log \mathrm{LR} = \frac{1}{n} \sum_{i=1}^{n} Z_i \to D\left(\mathbf{f} \mid \mathbf{g}\right).$$

Hence we can think of the Kullback distance as the expected number of bits per sample we learn about $\mathbf{f}$ if we erroneously think that $\mathbf{g}$ is the 'true' model.

KTH Matematik

Kullback Distance & Chernoff's Inequality

$$P_{\mathbf{g}}\left(\sum_{i=1}^{n} Z_i \geq nD\left(\mathbf{f} \mid \mathbf{g}\right)\right) \leq e^{-nD(\mathbf{f}|\mathbf{g})}$$

Note that here $P_{\mathbf{g}}$ means computing probability using $\mathbf{g}$ as the 'true model'.

KTH Matematik

Kullback Distance: Information Content

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, $\mathbf{g} = (1/L, \cdots, 1/L)$, then

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

This case of $D(\mathbf{f} \mid \mathbf{g})$ is in some quarters of bioinformatics known as the **information content**. We shall soon see why this may make sense and how it is used.

KTH Matematik

Sequence Logos

Sequence Logos display the frequencies of bases at each position in aligned sequence. The characters presenting the sequence are stacked on top of each other in each position of in the aligned sequence. The height of the letter is made proportional to the relative frequency. The most common one is on the top. The height of the entire stack is adjusted to signify the information content of the sequences at that position.

`http://www.lecb.ncifcrf.gov/~toms/glossary.html#sequencewalker`

KTH VETENSKAP OCH KONST
KTH Matematik

Sequence Logos

$$\text{height of base } b \text{ at position I} = f(b) \cdot R_{\text{seq}},$$

where

$$R_{\text{seq}} = 2 - (H(X_l) + e(n)).$$

Note that $2 - H(X_l)$ is the information content for the base at position l, since 2 is the entropy of the random DNA. For the biological rational and details see T.D. Schneider and R.M. Stephens: Sequence Logos (at url above).

KTH VETENSKAP OCH KONST
KTH Matematik

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 $\mathrm{Be}(p)$ and $\mathrm{Be}(g)$, respectively. Then

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

KTH VETENSKAP OCH KONST
KTH Matematik

Properties

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

- Recalling information content $D\left(\mathbf{f} \mid \mathbf{g}\right) = \log L - H(X)$ we have

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

Hence uncertainty is maximized by the uniform distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Complexity

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

The entropy $H(X)$ is thus also a measure of complexity.

KTH Matematik

Proof of $D(\mathbf{f} \mid \mathbf{g}) \geq 0$:

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

which equals

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

$\phi(x) = -\log x$ is a convex function, Jensen's inequality gives,

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

KTH Matematik

Proof: Jensen's inequality

ϕ is a convex function:

$$E\left[\phi\left(X\right)\right] \geq \phi\left(E\left[X\right]\right).$$

For a discrete random variable this is nothing more than the definition of convexity.

KTH VETENSKAP OCH KONST
KTH Matematik

Convexity

ϕ is a convex function if and only if

$$\phi\left(\pi_1 x_1 + \ldots + \pi_n x_n\right) \leq \pi_1 \phi\left(x_1\right) + \ldots + \pi_n \phi\left(x_n\right)$$

for all $x_1, \ldots, x_n$ and $\pi_1 + \ldots + \pi_n = 1$, $\pi_i \geq 0$.

KTH Matematik

Proof: continued

But

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

and since $\log 1 = 0$, we have proved our assertion, as claimed above.

KTH Matematik

End of Lecture 4

KTH Matematik