Bayesian Networks & Statistical Genetics Lecture 2.

Timo Koski

12.04.2010

KTH Matematik

X is a (**discrete**) random variable that assumes values in $\mathcal{X}$ and Y is a (**discrete**) random variable that assumes values in $\mathcal{Y}$.

Random Variables $\mathcal{X}$ and $\mathcal{Y}$ are two discrete *state spaces*, whose generic elements are called *values* or *instantiations* and denoted by x_i and y_j, respectively.

$$\mathcal{X} = \{x_1, \cdots, x_L\}, \mathcal{Y} = \{y_1, \cdots, y_J\}.$$

$\mid \mathcal{X} \mid$ (:= the number of elements in $\mathcal{X}) = L \leq \infty$, $\mid \mathcal{Y} \mid = J \leq \infty$. Unless otherwise stated the alphabets considered here are finite.

Joint Probability Distributions

A two dimensional *joint (simultaneous) probability distribution* (simultan sannolikhetsfrdelning) is a probability defined on the alphabet $\mathcal{X} \times \mathcal{Y}$

$$p(x_i, y_j) := P(X = x_i, Y = y_j). \tag{1}$$

$$p(x_i, y_j) \geq 0, \tag{2}$$

$$\sum_{i=1}^{L} \sum_{j=1}^{J} p(x_i, y_j) = 1. \tag{3}$$

KTH Matematik

Marginal distribution

Marginal distribution for X:

$$p(x_i) = \sum_{j=1}^{J} p(x_i, y_j). \tag{4}$$

Marginal distribution for Y:

$$p(y_j) = \sum_{i=1}^{L} p(x_i, y_j). \tag{5}$$

Marginal distribution

These notions can be extended to define the joint (simultaneous) probability distribution of n random variables and the marginal distributions of any subset thereof.

Simultaneous distirbution as a table

$X \,/\, Y$	y_1	y_2	y_3
x_1	0.05	0.10	0.05
x_2	0.15	0.00	0.25
x_3	0.10	0.20	0.10

For example

$$p\left(X = x_2, Y = y_3\right) = 0.25$$

KTH Matematik

Marginal distribution

X/Y	y_1	y_2	y_3
x_1	0.05	0.10	0.05
x_2	0.15	0.00	0.25
x_3	0.10	0.20	0.10

$$p(X = x_1) = 0.05 + 0.10 + 0.05 = 0.20$$

$$p(X = x_2) = 0.15 + 0.00 + 0.25 = 0.40$$

$$p(X = x_3) = 0.10 + 0.20 + 0.10 = 0.40$$

Conditional Probability Distributions

The conditional probability for $X = x_i$ given $Y = y_j$ is

$$p(x_i \mid y_j) := \frac{p(x_i, y_j)}{p(y_j)}. \qquad (6)$$

The conditional probability for $Y = y_j$ given $X = x_i$ is

$$p(y_j \mid x_i) := \frac{p(x_i, y_j)}{p(x_i)}. \qquad (7)$$

Here we assume $p(y_j) > 0$ and $p(x_i) > 0$.

KTH Matematik

In other words

$$p(y_j \mid x_i) = \frac{\text{prob. for the event } \{X = x_i, Y = y_j\}}{\text{prob. for the event } \{X = x_i\}}.$$

Hence

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

for every j and i, respectively.

Conditional Probability Distributions

In the table above

$$p(y_1|x_1) = \frac{p(x_1, y_1)}{p(x_1)} = \frac{0.05}{0.20} = \frac{5}{20}$$

$$p(y_2|x_1) = \frac{p(x_1, y_2)}{p(x_1)} = \frac{0.10}{0.20} = \frac{1}{2}$$

$$p(y_3|x_1) = \frac{p(x_1, y_3)}{p(x_1)} = \frac{0.05}{0.20} = \frac{5}{20}$$

$$\frac{5}{20} + \frac{1}{2} + \frac{5}{20} = 1$$

Next

$$P_X(A) := \sum_{x_i \in A} p(x_i) \tag{8}$$

is the probability of the event that X assumes a value in A, a subset of $\mathcal{X}$. Then one easily establishes the complement rule

$$P_X\left(A^c\right) = 1 - P_X(A), \tag{9}$$

where A^c is the complement of A, i.e., those outcomes which do not lie in A.

$$P_X(A \cup B) = P_X(A) + P_X(B) - P_X(A \cap B), \tag{10}$$

is immediate. If $A \cap B = \emptyset$, then $P_X(A \cap B) = 0$.

The conditional probability for $X = x_i$ given $X \in A$ is denoted by $P_X(x_i \mid A)$ and given by

$$P_X(x_i \mid A) = \begin{cases} \frac{P_X(x_i)}{P_X(A)} & \text{if } x_i \in A \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

KTH Matematik

Law of Total Probability

$$P(X \in A) = \sum_{j=1}^{J} P(X \in A \mid Y = y_j)\, p(Y = y_j) \qquad (*)$$

$$P(Y \in B) = \sum_{i=1}^{L} P(Y \in B \mid X = x_i)\, p(X = x_i)$$

Independence

X and Y are *independent* random variables if and only if

$$p(x_i, y_j) = p(x_i) \cdot p(y_j) \tag{12}$$

for *all pairs* (x_i, y_j) in $\mathcal{X} \times \mathcal{Y}$. In other words all events $\{X = x_i\}$ and $\{Y = y_j\}$ are to be independent.

Independence

$X\,/\,Y$	y_1	y_2	y_3
x_1	0.05	0.10	0.05
x_2	0.15	0.00	0.25
x_3	0.10	0.20	0.10

$$p\left(X = x_1\right) = 0.05 + 0.10 + 0.05 = 0.20$$

$$p\left(Y = y_3\right) = 0.05 + 0.25 + 0.10 = 0.40$$

$$p\left(X = x_1\right) \cdot p\left(Y = y_3\right) = 0.08 \neq 0.05 = p\left(X = x_1, Y = y_3\right)$$

Independence

We say that $X_1, X_2, \ldots, X_n$ are **independent** random variables if and only if the joint distribution

$$p(x_{i_1}, x_{i_2} \ldots, x_{i_n}) = P\left(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}\right) \tag{13}$$

equals

$$= p_{X_1}(x_{i_1}) \cdot p_{X_2}(x_{i_2}) \cdots p_{X_n}(x_{i_n}) \tag{14}$$

for every $x_{i_1}, x_{i_2} \ldots, x_{i_n} \in \mathcal{X}^n$.

KTH Matematik

Chain Rule

Let Z be a (discrete) random variable that assumes values in $\mathcal{Z} = \{z_k\}_{k=1}^{K}$. If $p(z_k) > 0$,

$$p(x_i, y_j \mid z_k) = \frac{p(x_i, y_j, z_k)}{p(z_k)}.$$

Then we obtain as an identity

$$p(x_i, y_j \mid z_k) = \frac{p(x_i, y_j, z_k)}{p(y_j, z_k)} \cdot \frac{p(y_j, z_k)}{p(z_k)},$$

Chain Rule

and again by definition of conditional probability

$$p(x_i \mid y_j, z_k) \cdot p(y_j \mid z_k).$$

Chain Rule So,

$$p(x_i, y_j \mid z_k) = \frac{p(x_i, y_j, z_k)}{p(y_j, z_k)} \cdot \frac{p(y_j, z_k)}{p(z_k)}$$

In other words,

$$p(x_i, y_j \mid z_k) = p(x_i \mid y_j, z_k) \cdot p(y_j \mid z_k). \tag{15}$$

KTH Matematik

A generalization

$$p\left(X_1,\ldots,X_n\right)=\prod_{i=1}^{n} p\left(X_i \mid X_1,\ldots,X_{i-1}\right)$$

$p\left(X_1 \mid X_0\right)=p\left(X_0\right)$.

Conditional Independence (IRRELEVANCE) The random variables X and Y are called *conditionally independent* given Z if

$$p(x_i, y_j | z_k) = p(x_i | z_k) \cdot p(y_j | z_k) \tag{16}$$

for all triples $(z_k, x_i, y_j) \in \mathcal{Z} \times \mathcal{X} \times \mathcal{Y}$. We write this as

$$X \perp Y | Z. \tag{17}$$

Y is irrelevant for X given Z, and X is irrelevant for Y given Z.

KTH Matematik

There are several equivalent ways of expressing conditional independence. We have for instance

$$X \perp Y|Z \Leftrightarrow p(x_i|y_j, z_k) = p(x_i|z_k).$$

To see this equivalence in one direction we write

$$p(x_i|y_j, z_k) = \frac{p(x_i, y_j, z_k)}{p(y_j, z_k)}$$

and assume $p(z_k) > 0$, so

$$= \frac{p(x_i, y_j, z_k)}{p(z_k)} \frac{p(z_k)}{p(y_j, z_k)}$$

$$= \frac{p(x_i, y_j \mid z_k)}{p(y_j \mid z_k)},$$

and assuming $X \perp Y|Z$ we get

$$= \frac{p(x_i|z_k) \cdot p(y_j|z_k)}{p(y_j \mid z_k)} = p(x_i|z_k),$$

as claimed.

Learning and Bayes' Rule

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

we have in a formal way

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

But the marginal distribution $p(Y)$ is by the law of total probability (see $(*)$ above) written as

$$p(y_j) = \sum_{i=1}^{L} p(y_j \mid x_i) p(x_i). \tag{18}$$

KTH Matematik

Bayes' Rule

$$p(x_i \mid y_j) = \frac{p(y_j \mid x_i) \cdot p(x_i)}{\sum_{i=1}^{L} p(y_j \mid x_i) p(x_i)}. \tag{19}$$

KTH Matematik

$p(X)$: 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*. Bayes' rule can be seen as just a formal identity derived from certain premises and definitions.

Learning and Bayes' Rule 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 $\mathcal{X}$. Probability is a degree of rational belief, Bayes' rule is a rule for reasoning.

KTH Matematik

If we learn that the event $Y = y_j$ is true, then we change $p(X)$ to a new probability distribution $p^*(X)$ according to Bayes' Rule.

$$p(X) \mapsto p^*(X) = p(X \mid Y = y_j)$$

So the posterior becomes the new prior.

Terminology for Bayes' Rule

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

$$\textbf{Posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

Jeffrey's Rule: kinematics of probability (1)

Suppose you change your probabilities on $\mathcal{Y}$ from the distribution $p(Y)$ to the distribution $p^*(Y)$. How should this change be propagated to the distribution on $\mathcal{X}$. R. Jeffrey thinks that Bayes' rule is not the only way. He suggests that $p(X)$ is updated to $p^*(X)$ defined by the rule

$$p^*(x_i) = \sum_{j=1}^{J} p(x_i \mid y_j) p^*(y_j), \tag{20}$$

where the assumption is that

$$p(x_i \mid y_j) = p^*(x_i \mid y_j).$$

Jeffrey's Rule: kinematics of probability (2)

$$p^*(x_i) = \sum_{j=1}^{J} p(x_i \mid y_j) p^*(y_j)$$

The argument is that if the event $X = x_i$ is 'not directly affected' by the flow of experience that was involved in $p(Y) \mapsto p^*(Y)$, then we should not use Bayes' rule. What does this mean ?

KTH Matematik

Jeffrey's Rule: kinematics of probability (3)

Let us say that e is the evidence that made us do $p(Y) \mapsto p^*(Y)$. Then we set

$$p^*(x_i) = p(x_i \mid e)$$

and get by Bayes rule and law of total probability

$$p(x_i \mid e) = \sum_{j=1}^{J} p(x_i \mid y_j, e) p(y_j \mid e)$$

$$= \sum_{j=1}^{J} p(x_i \mid y_j) p^*(y_j),$$

if X and e are conditionally independent given Y.

KTH Matematik

But, are we permitted to write

$$p^*(x_i) = p(x_i \mid e)$$

$$= \frac{p(x_i, e)}{p(e)},$$

as $p(x_i, e)$ was not specified, if e was not a part of our knowledge base. E.g., e may not have been anticipated. Hence Jeffrey's rule seems more generally valid than Bayes' rule.

But even if $p(x_i, e)$ was not specified as a numerical quantity, we may still be permitted to apply conditional independence of X and e given Y by qualitative judgement.
Lesson: We shall specify conditional independencies instead of numerical joint distributions.

BERNOULLI R.V'S

Consider X with values $\mathcal{X} = \{0, 1\}$ and $0 < \theta < 1$ with the probability table

p	$x = 1$	$x = 0$
$p(x)$	θ	$1 - \theta$

then we call X a Bernoulli random variable with the *probability of success* θ. We write

$$X \in Be(\theta).$$

We refer to θ as the *parameter* of the Bernoulli distribution p.

If $X_1, X_2, \ldots, X_n$ are independent and $X_i \in Be(\theta)$, then

$$p(1,1,0,1,0,1,1) =$$

$$= \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$= \theta^5 \cdot (1-\theta)^2.$$

KTH Matematik

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 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

$$\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}, x_{i_l} \in \{0, 1\}.$$

As our model we take the bits in $\mathbf{x}$ to be outcomes of $X_i \in \mathrm{Be}(\theta)$ conditionally independent given $\Theta = \theta$.

$$X_i \perp X_j \mid \Theta \quad \text{for all } i \neq j$$

Not only are pairs independent, but all subsets of $X_{i_1}, \ldots, X_{i_k}$. In subjective probability the parameters of a probability model are regarded as random variables.

Hence

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

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

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

Find the model that is in some sense best for **x**. In the thumbtack example we understand this as follows. We have observed n outcomes of flips of a thumbtack and wish to determine which of the values θ that best describes this set of flips.

Learning about probilities: Bayes' Rule for parameters

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

$p(\Theta \mid X)$ and $p(\Theta)$ are probability densities.

Learning about probilities: Bayes' Rule for parameters

Θ is given a probability density function $f_\Theta(\theta)$, called the *prior density* .

$$f_\Theta(\theta) \geq 0, 0 \leq \theta \leq 1$$

and $f_\Theta(\theta) = 0$ elsewhere, and

$$\int_0^1 f_\Theta(\theta)\, d\theta = 1.$$

Also $P(a < \Theta \leq b) = \int_a^b f_\Theta(\theta)\, d\theta$.

$$f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta = \theta) \cdot f_{\Theta}(\theta)}{\int_0^1 P(\mathbf{x} \mid \Theta = \theta) \cdot f_{\Theta}(\theta)\, d\theta}, 0 \leq \theta \leq 1 \qquad (21)$$

and zero elsewhere. Due to the standardization $f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x})$ is a probability density for Θ.

Posterior density

The posterior $f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x})$ expresses our updated belief in the statement that θ is the probability of success given that we have observed $\mathbf{x}$.

Prior density

Let us consider the *uniform prior* given by

$$f_\Theta(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The uniform prior is often interpreted as a representation of complete ignorance. This is a special case of a *Beta density*.

Prior density

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

by the *Beta integral*.

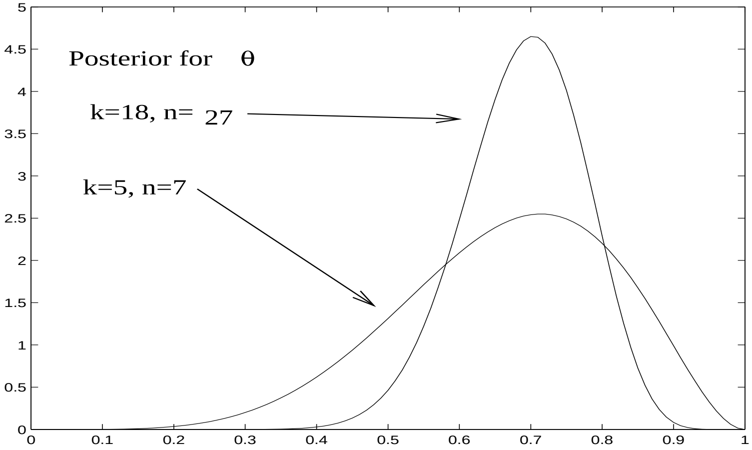
KTH Matematik

Posterior is a Beta density

$$f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x}) = \begin{cases} \frac{(n+1)!}{k!(n-k)!} \cdot \theta^k (1-\theta)^{n-k} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{22}$$

KTH Matematik

Posterior Densities for θ in Be(θ)



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

$$\widehat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{0 \leq \theta \leq 1} P\left(\mathbf{x} \mid \Theta = \theta\right)$$

$$= \mathrm{argmax}_{0 \leq \theta \leq 1} \theta^{k} \cdot (1-\theta)^{n-k}.$$

The Maximum A Posterior Estimate

The **maximum a posterior estimate** MAP $\widehat{\theta}_{\mathrm{MAP}}$ of θ is defined by

$$\widehat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{0 \leq \theta \leq 1} f_{\Theta \mid \mathbf{x}} \left(\theta \mid \mathbf{x} \right)$$

KTH Matematik

Interpretation of MLE

Find the parameter value within the model that gives the (training) sequence $\mathbf{x}$ *the highest possible probability*. The probability $P(\mathbf{x} \mid \Theta = \theta)$ regarded as a function of θ is known as the *likelihood function*

$$L_{\mathbf{x}}(\theta) = P(\mathbf{x} \mid \Theta = \theta).$$

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

$$-\log L_{\mathbf{x}}(\theta) = -\log P(\mathbf{x} \mid \Theta = \theta).$$

is called the log likelihood function.

Maximization of the likelihood function or the log likelihood function by calculus gives

$$\widehat{\theta}_{\mathrm{ML}} = \frac{k}{n}. \tag{23}$$

What is $\widehat{\theta}_{\mathrm{MAP}}$ in this case ?

KTH Matematik