T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 29th of February, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

0. Jaakko Hollmén gave a demonstration on his software package Zone for clustering zero-one data. This will be part of the project assignment.

1. Given a Naïve Bayes model with four binary variables C, X_1, X_2, X_3, that is $P(C, X_1, X_2, X_3) = P(C)P(X_1 \mid C)P(X_2 \mid C)P(X_3 \mid C)$ and a dataset with five samples $t = 1 \ldots 5$ (see table below), write the likelihood function $P(C, X_1, X_2, X_3 \mid \boldsymbol{\theta})$ of the model parameters $\boldsymbol{\theta}$ (the values in the conditional probability tables). Find $P(C)$ and $P(X_1 \mid C = 1)$ that maximize the likelihood (use the notation $\theta_1 = P(C = 1)$ and $\theta_2 = P(X_1 = 1 \mid C = 1)$).

Data:

t	C_t	X_{1t}	X_{2t}	X_{3t}
1	0	1	0	1
2	1	0	1	0
3	0	0	1	0
4	1	0	1	1
5	1	1	1	0

Solution:

Assuming the 5 samples independent of each other, the likelihood of the parameters is the product of probabilities of each data sample given the parameters, that is:

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{5} P(C_t, X_{1t}, X_{2t}, X_{3t}) = \prod_{t=1}^{5} P(C_t)P(X_{1t} \mid C_t)P(X_{2t} \mid C_t)P(X_{3t} \mid C_t) \tag{1}$$

Note that we write $P(C)$ as a shorthand of $P(C \mid \boldsymbol{\theta})$ etc. Because the logarithm function is monotonically increasing, the maximum likelihood is the same as maximum log-likelihood, and we would prefer sums over products, so let us turn to study the likelihood on the logarithmic scale.

$$\log L(\boldsymbol{\theta}) = \sum_{t=1}^{5} \left[\log P(C_t) + \sum_{i=1}^{3} \log P(X_{it} \mid C_t) \right] \tag{2}$$

The maximum of L can be found at the zero of the derivative. Most terms of L are constant w.r.t. a particular parameter, so many of them can be dropped out.

$$0 = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_1} = \frac{\partial}{\partial \theta_1} \sum_{t=1}^{5} \log P(C_t) \tag{3}$$

$$= \frac{\partial}{\partial \theta_1} 3 \log \theta_1 + 2 \log(1 - \theta_1) = \frac{3}{\theta_1} - \frac{2}{1 - \theta_1} = 0 \tag{4}$$

$$\theta_1 = 3/5 \tag{5}$$

$$P(C) = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \tag{6}$$

The solution of θ_2 is very similar:

$$0 = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_2} = \frac{\partial}{\partial \theta_2} \sum_{t=1}^{5} \log P(X_{1t} \mid C_t) \tag{7}$$

$$= \frac{\partial}{\partial \theta_1} \log P(X_{12} \mid C_2) + \log P(X_{14} \mid C_4) + \log P(X_{15} \mid C_5) \tag{8}$$

$$= \frac{\partial}{\partial \theta_1} 2 \log \theta_2 + \log(1 - \theta_2) = \frac{1}{\theta_2} - \frac{2}{1 - \theta_2} = 0 \tag{9}$$

$$\theta_2 = 1/3 \tag{10}$$

$$P(X_1 \mid C = 1) \approx \begin{pmatrix} 0.67 \\ 0.33 \end{pmatrix} \tag{11}$$

We can note that the maximum likelihood solution is basically about counting how many times each case happens, for instance $C = 1$ happens in three cases out of five so $P(C = 1) = 3/5$ for the maximum likelihood estimate of $\boldsymbol{\theta}$.

2. Given a Naïve Bayes model with three binary variables defined by the tables below, classify the data set below. Classification is defined as $C^* = \arg\max_C P(C \mid X_1, X_2)$.

P(C)	
C=0	0.7
C=1	0.3

$P(X_1 \mid C)$	C=0	C=1
X_1=0	0.5	0.8
X_1=1	0.5	0.2

$P(X_2 \mid C)$	C=0	C=1
$X_2 = 0$	0.6	0.3
$X_2 = 1$	0.4	0.7

Data:

t	X_{1t}	X_{2t}
1	1	1
2	0	1

Solution:

$P(C \mid X_1, X_2) = \frac{P(C,X_1,X_2)}{P(X_1,X_2)}$, where $P(X_1, X_2)$ is a normalization constant. We have four cases:

$$\begin{aligned}
P(C_1 = 0, X_{11}, X_{21}) &= P(C_1 = 0)P(X_{11} = 1 \mid C_1 = 0)P(X_{21} = 1 \mid C_1 = 0) \\
&= 0.7 \cdot 0.5 \cdot 0.4 = 0.14 \qquad (12)
\end{aligned}$$

$$\begin{aligned}
P(C_1 = 1, X_{11}, X_{21}) &= P(C_1 = 1)P(X_{11} = 1 \mid C_1 = 1)P(X_{21} = 1 \mid C_1 = 1) \\
&= 0.3 \cdot 0.2 \cdot 0.7 = 0.042 \qquad (13)
\end{aligned}$$

$$\begin{aligned}
P(C_2 = 0, X_{12}, X_{22}) &= P(C_2 = 0)P(X_{12} = 0 \mid C_2 = 0)P(X_{22} = 1 \mid C_2 = 0) \\
&= 0.7 \cdot 0.5 \cdot 0.4 = 0.14 \qquad (14)
\end{aligned}$$

$$\begin{aligned}
P(C_2 = 1, X_{12}, X_{22}) &= P(C_2 = 1)P(X_{12} = 0 \mid C_2 = 1)P(X_{22} = 1 \mid C_2 = 1) \\
&= 0.3 \cdot 0.8 \cdot 0.7 = 0.168 \qquad (15)
\end{aligned}$$

The normalization constants are

$$P(X_{11}, X_{21}) = P(C_1 = 0, X_{11}, X_{21}) + P(C_1 = 1, X_{11}, X_{21}) = 0.182 \qquad (16)$$

$$P(X_{12}, X_{22}) = P(C_2 = 0, X_{12}, X_{21}) + P(C_2 = 1, X_{12}, X_{21}) = 0.308 \qquad (17)$$

Now we can get the posterior probabilities for the classifications by normalizing:

$$P(C_1 \mid X_{11}, X_{21}) = \frac{P(C_1, X_{11}, X_{21})}{P(X_{11}, X_{21})} = \begin{pmatrix} 0.769 \\ 0.231 \end{pmatrix} \qquad (18)$$

$$P(C_2 \mid X_{12}, X_{22}) = \frac{P(C_2, X_{12}, X_{22})}{P(X_{12}, X_{22})} = \begin{pmatrix} 0.455 \\ 0.545 \end{pmatrix} \qquad (19)$$

The best guess or the maximum a posteriori classification is thus $C_1^* = 0$ and $C_2^* = 1$.

Problems 3 and 4 were left for the next session.