T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 25th of January, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. Consider a bent coin and how to estimate the probability of tails μ. The random variable $X \in \{0,1\}$ (heads=0, tails=1) is distributed accoring to the Bernoulli distribution with the parameter μ (see page 685 in Bishop, 2006).
(a) Derive a maximum likelihood estimator for μ and estimate $\hat{\mu}$ for the data set from the lecture (7 heads and 5 tails out of 12 tosses).

$$P(\{X_i\}_{i=1}^{12} \mid \mu) = \prod_{i=1}^{12} \mathrm{Bern}(X_i \mid \mu) \tag{1}$$
$$= \mu^5(1-\mu)^7 \tag{2}$$

The maximum likelihood solution is at the zero of the derivative of the likelihood:

$$\frac{\partial}{\partial \mu} P(\{X_i\}_{i=1}^{12} \mid \mu) = 5\mu^4(1-\mu)^7 - 7\mu^5(1-\mu)^6 = 0 \tag{3}$$
$$\hat{\mu} = \frac{5}{12} \approx 0.42 \tag{4}$$

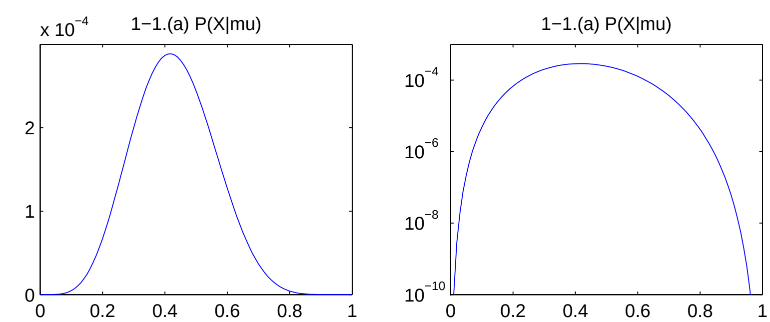

Figure 1: Problem 1.(a) The likelihood of μ as a function of μ on the absolute scale (left) and on the logarithmic scale (right).

(b) Using a fair coin, what is the probability that out of 12 tosses, strictly more than 10 are heads (see Binomial distribution, page 686).

The Binomial distribution is defined as

$$\mathrm{Bin}(m \mid N, \mu) = \frac{N!}{m!(N-m)!}\mu^m(1-\mu)^(N-m), \tag{5}$$

where m is the number of heads, $N = 12$ is the number of tosses, and $\mu = 0.5$ is the probability of heads.

$$\begin{aligned} P(m > 10) &= \mathrm{Bin}(m = 11 \mid 12, 0.5) + \mathrm{Bin}(m = 12 \mid 12, 0.5) & (6) \\ &= 13 \cdot 0.5^{12} \approx 0.0032 & (7) \end{aligned}$$

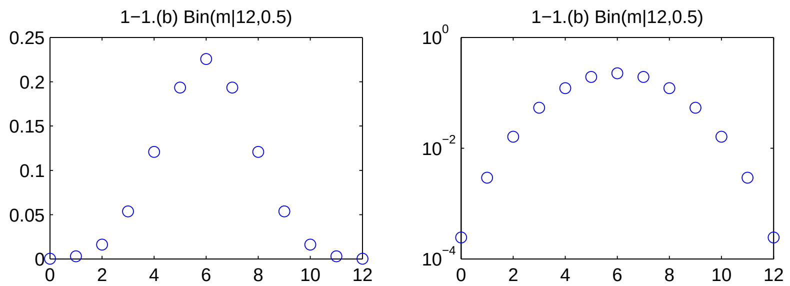

Figure 2: Problem 1.(b) The Binomial distribution. The probability is plotted as a function of $m = 0, 1, \ldots, 12$ on the absolute scale (left) and on the logarithmic scale (right).

2. Compute the probability $P(C \mid X)$ of using each coin in the guessing game from the lecture (see Bayes' theorem, p. 15). There are two bent coins ($C \in \{c_1, c_2\}$) with different properties and the player guesses which coin was used after learning whether the toss was head or tails. The properties of the coins are: $P(X = t \mid C = c_1) = \theta_1$ and $P(X = t \mid C = c_2) = \theta_2$. The used coin is chosen randomly by $P(C = c_1) = \pi_1$ and $P(C = c_2) = \pi_2$ with $\pi_1 + \pi_2 = 1$.

The solution is the direct application of the Bayes theorem (first equation) and the marginalization principle (second equation):

$$P(C = c_1 \mid X = t) = \frac{P(X = t \mid C = c_1)P(C = c_1)}{P(X = t)} \tag{8}$$

$$= \frac{P(X = t \mid C = c_1)P(C = c_1)}{\sum_{i=1}^{2} P(X = t \mid C = c_i)P(C = c_i)} \tag{9}$$

$$= \frac{\theta_1 \pi_1}{\theta_1 \pi_1 + \theta_2 \pi_2} \tag{10}$$

and similarly

$$P(C = c_2 \mid X = t) = \frac{\theta_2 \pi_2}{\theta_1 \pi_1 + \theta_2 \pi_2} \tag{11}$$

$$P(C = c_1 \mid X = h) = \frac{(1 - \theta_1)\pi_1}{(1 - \theta_1)\pi_1 + (1 - \theta_2)\pi_2} \tag{12}$$

$$P(C = c_2 \mid X = h) = \frac{(1 - \theta_2)\pi_2}{(1 - \theta_1)\pi_1 + (1 - \theta_2)\pi_2}. \tag{13}$$

3. The Naïve Bayes model has a class label C and observations $X_1, X_2, \ldots, X_6$ such that $P(X_1, X_2, X_3, X_4, X_5, X_6, C) = P(C)P(X_1|C)P(X_2|C) \ldots P(X_6|C)$.
(a) Simplify $P(X_1 \mid C, X_2)$

First let us rewrite it without the conditional probability, using just the joint probabilities. Then we can apply the assumption of the Naïve Bayes model and finally simplify:

$$(X_1 \mid C, X_2) = \frac{P(C, X_1, X_2)}{P(C, X_2)} \tag{14}$$

$$= \frac{P(C)P(X_1 \mid C)P(X_2 \mid C)}{P(C)P(X_2 \mid C)} \tag{15}$$

$$= P(X_1 \mid C) \tag{16}$$

(b) Solve the classification problem: $P(C \mid X_1, X_2, \ldots, X_6)$

Let us apply the Bayes theorem, the marginalization principle, and finally the Naïve Bayes assumption:

$$P(C \mid X_1, X_2, \ldots, X_6) = \frac{P(X_1, X_2, \ldots, X_6 \mid C)P(C)}{P(X_1, X_2, \ldots, X_6)} \tag{17}$$

$$= \frac{P(X_1, X_2, \ldots, X_6 \mid C)P(C)}{\sum_C P(X_1, X_2, \ldots, X_6 \mid C)P(C)} \tag{18}$$

$$= \frac{P(X_1 \mid C)P(X_2 \mid C) \ldots P(X_6 \mid C)P(C)}{\sum_C P(X_1 \mid C)P(X_2 \mid C) \ldots P(X_6 \mid C)P(C)} \tag{19}$$

4. Draw a graphical representation of the models in problems 1, 2, and 3 where nodes represent random variables and arrows represent direct dependencies (see Bayesian Networks, page 360).

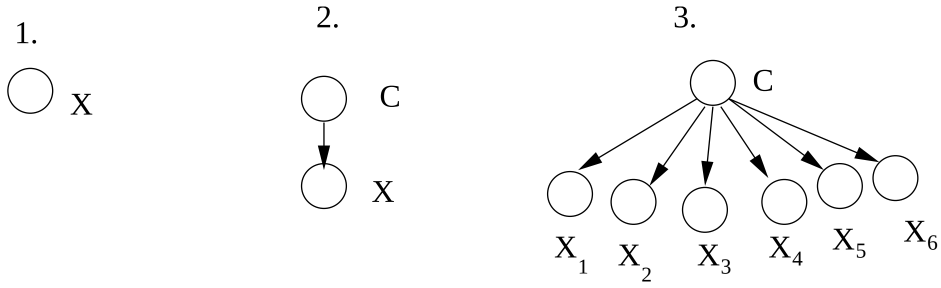

Figure 3: Problem 4.