# Introduction to Machine Learning, Spring 2022
# Homework 2

(Due Friday, Apr. 8 at 11:59pm (CST))

He Haoyu

April 5, 2022

## 1 Problem1

Given a set of data $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, where $y_i \in \{0, 1\}$. We want to conduct a binary classification, and the decision boundary is $\beta_0 + x^T\beta = 0$. When $\beta_0 + x^T\beta > 0$, the sample will be classified as 1, and 0 otherwise.

(a) Define a function which enables to map the range of an arbitrary linear function to the range of a probability [2 points]

(b) Derive the posterior probability of $P(y_i = 1|x_i)$ and $P(y_i = 0|x_i)$ [3 points]

(c) Write the log-likelihood for N observations, which means:

$$l(\theta) = logP(Y|X) = \sum_{i=1}^{N} log(P(y_i|x_i))$$

(Using the expression of $P(y_i|x_i)$ in (b) and eliminate redundant items) [5 points]

**Solution:**

**(a)** We can define a map $f(x) = \frac{e^x}{e^x+1}$ from the range of an arbitrary linear function to the range of a probability

**(b)** Let

$$\log \frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)} = \log \frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} = \beta_0 + x^T\beta$$

By the properties of possibility, we know

$$P(y_i = 1|x_i) + P(y_i = 0|x_i) = 1$$

Thus,

$$P(y_i = 1|x_i) = \frac{\exp(\beta_0 + x^T\beta)}{1 + \exp(\beta_0 + x^T\beta)}$$

$$P(y_i = 0|x_i) = \frac{1}{1 + \exp(\beta_0 + x^T\beta)}$$

**(c)** Let $\theta = \{\beta_0, \beta\}$

$$\begin{aligned} l(\theta) &= \log P(Y|X;\theta) \\ &= \sum_{i=1}^{N} \log(P(y_i|x_i;\theta)) \\ &= \sum_{i=1}^{N} (y_i \log(P(y_i = 1|x_i;\theta)) + (1 - y_i) \log(P(y_i = 0|x_i;\theta))) \\ &= \sum_{i=1}^{N} (y_i(\beta_0 + x^T\beta) - \log(1 + \exp(\beta_0 + x^T\beta))) \end{aligned}$$

Table 1: probability distribution for $X$

| $X$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P$ | $\theta^2$ | $2\theta(1-\theta)$ | $\theta^2$ | $1-2\theta$ |

## 2 Problem2

(a) Given a random variable $X$ and its probability distribution is shown in Table 1. Now, we sample 8 times and get the results $\{3, 1, 3, 0, 3, 1, 2, 3\}$. Please derive the MLE estimate for $\theta (0 < \theta < \frac{1}{2})$. [4 points]

(b) Now we discuss Bayesian inference in coin flipping. Let's denote the number of heads and the total number of trials by $N_1$ and $N$, respectively. Please derive the MAP estimate based on the following prior:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.3 \\ 0 & \text{otherwise,} \end{cases}$$

which believes the coin is fair, or is slightly biased towards tails. [4 points]

(c) Suppose the true parameter is $\theta = 0.31$. Please compare the prior in (b) with the Beta prior distribution(You can review this part in Lecture 07). Which prior leads to a better estimate when $N$ is small? Which prior leads to a better estimate when N is large? [2 points]

**Solution:**

**(a)** $D = \{3, 1, 3, 0, 3, 1, 2, 3\}$,$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)$

$$\begin{aligned} P(D|\theta) &= P^4(X = 3|\theta)P^2(X = 1|\theta)P(X = 0|\theta)P(X = 2|\theta) \\ &= (1 - 2\theta)^4(2\theta(1 - \theta))^2\theta^4 \end{aligned}$$

$$\begin{aligned} &\frac{dP(D|\theta)}{\mathrm{d}\theta} = 0 \\ &\Rightarrow 12\theta^2 - 14\theta + 3 = 0 \\ &\Rightarrow \theta = \frac{7 \pm \sqrt{13}}{12} \end{aligned}$$

Because of the range of $\theta$, we know $\hat{\theta} = \frac{7-\sqrt{13}}{12}$

**(b)** Let X be the number of heads,by MAP,$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(\theta|X = N_1)$

$$\begin{aligned} P(\theta|X = N_1) &= \frac{P(X = N_1|\theta)P(\theta)}{P(X = N_1)} \\ &\propto P(X = N_1|\theta)P(\theta) \\ &\propto \theta^{N_1}(1 - \theta)^{N-N_1}p(\theta) \\ &= \begin{cases} 0.5 \times 0.5^{N_1} \times 0.5^{N-N_1} & \text{if } \theta = 0.5 \\ 0.5 \times 0.3^{N_1} \times 0.7^{N-N_1} & \text{if } \theta = 0.3 \\ 0 & otherwise \end{cases} \\ &\propto \begin{cases} 0.5^N & \text{if } \theta = 0.5 \\ 0.3^{N_1} \times 0.7^{N-N_1} & \text{if } \theta = 0.3 \\ 0 & otherwise \end{cases} \end{aligned}$$

Thus,

$$\hat{\theta} = \begin{cases} 0.5 & \text{if } N_1 > N \log_{\frac{3}{7}} \frac{5}{7} \\ 0.3 & \text{if } N_1 < N \log_{\frac{3}{7}} \frac{5}{7} \end{cases}$$

**(c)** By Beta prior, we know $\theta \sim Beta(N_1, N - N_1)$, so,

$$P(\theta) = \frac{1}{\beta(N_1, N - N_1)}\theta^{N_1-1}(1-\theta)^{N-N_1-1}$$

By MAP,

$$\begin{aligned} P(\theta|X = N_1) &= \frac{P(X = N_1|\theta)P(\theta)}{P(X = N_1)} \\ &\propto P(X = N_1|\theta)P(\theta) \\ &\propto \theta^{N_1}(1-\theta)^{N-N_1}p(\theta) \\ &= \theta^{N_1}(1-\theta)^{N-N_1}\frac{1}{\beta(N_1, N - N_1)}\theta^{N_1-1}(1-\theta)^{N-N_1-1} \\ &\propto \theta^{2N_1-1}(1-\theta)^{2N-2N_1-1} \end{aligned}$$

$\hat{\theta} = \underset{\theta}{\text{argmax}} P(\theta|X = N_1)$,so

$$\begin{aligned} \frac{d\theta^{2N_1-1}(1-\theta)^{2N-2N_1-1}}{\mathrm{d}\theta} &= 0 \\ \Rightarrow \theta &= \frac{2N_1 - 1}{2N - 2} \end{aligned}$$

When N is large, the Beta prior is better. Because if N is large, $\lim_{x\to\infty} \frac{2N_1-1}{2N-2} = \frac{N_1}{N} = \theta$, has less gap. When N is small, the prior in (b) is better.Because if N is small, $\hat{\theta} = 0.3$ is great enough. In fact, with small N, $\hat{\theta} = \frac{2N_1-1}{2N-2}$ by Beta prior is probably not closed to 0.31.

## 3 Problem3

According to the following Fig. 3, answer the following questions:

(a) use the D-separation to discus whether the following statements are true or not:

(1) Given $x_4$, $\{x_1, x_2\}$ and $\{x_6, x_7\}$ are conditionally independent.[1(reason)+1(conclusion) points]

(2) Given $x_6$, $x_3$ and $x_2$ are conditionally independent.[1(reason)+1(conclusion) points]

(b) if all the nodes are observed and boolean variables, please complete the process of learning the parameter $\theta_{x_6|i,j}$ by using **MLE**, where $\theta_{x_6|i,j} = p(x_6 = 1 \mid x_3 = i, x_4 = j), i, j \in \{0, 1\}$.[6 points]

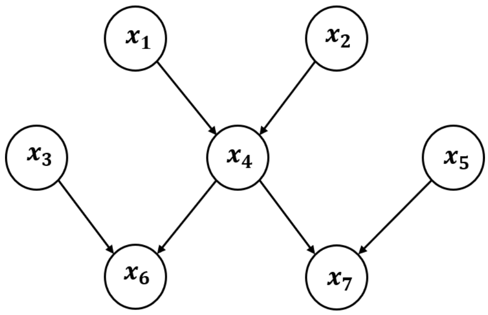


Figure 1: The Bayesian network for questions 3.

**Solution:**

**(a)** (1) True. Let A = $\{x_1, x_2\}$, B = $\{x_6, x_7\}$, the path is like $A \to x_4 \to B$ head-to-tail. For given $x_4$, $x_4$ is blocked, which means $\{x_1, x_2\}$ and $\{x_6, x_7\}$ are conditionally independent given $x_4$.

(2) False. First of all, $x_3 \to x_6 \leftarrow x_4$ is head-to-head path,so for given $x_6$, $x_6$ is not blocked. Then, $x_2 \to x_4 \to x_6$ is head-to-tail path, $x_4$ is not blocked for not given $x_4$. Thus, Given $x_6$, $x_3$ and $x_2$ are not conditionally independent.

**(b)** Suppose we observed N points. Let $\theta = \{\theta_{x_1}, \theta_{x_2}, \theta_{x_3}, \theta_{x_4}, \theta_{x_5}, \theta_{x_6|i,j}, \theta_{x_7|j}\}$

$$\begin{aligned}
\log P(D|\theta) &= \log \prod_{k=1}^{N} P(x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k}, x_{7k}|\theta) \\
&= \sum_{k=1}^{N} \log(P(x_{1k}|\theta) + P(x_{2k}|\theta) + P(x_{3k}|\theta) + P(x_{4k}|\theta) + P(x_{5k}|\theta) + P(x_{6k}|x_3, x_4, \theta) + P(x_{7k}|x_4, \theta))
\end{aligned}$$

$$\frac{\partial \log P(D|\theta)}{\partial \theta_{x_6}|i,j} = \sum_{k=1}^{N} \frac{\partial \log P(x_{6k}|x_3, x_4, \theta)s}{\partial \theta_{x_6}|i,j}$$

Now define $I(\cdot)$ be the indicator variable, $I(\cdot) = 1$ if and only if is $(\cdot)$ is right, and 0 for all other cases. Thus,

$$\hat{\theta}_{x_6}|i,j = \frac{\sum_{k=1}^{N} I(x_{6k} = 1, x_{3k} = i, x_{4k} = j)}{I(x_{3k} = i, x_{4k} = j)}$$