Optimization and Machine Learning, Spring 2021
Homework 2 Solution

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

April 24, 2022

1 Problem1

[10 points]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]

Solution:

$$f(t) = \frac{e^t}{1+e^t}$$

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

Solution: Let

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

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

$$P(y_i = 1|x_i) = \frac{e^t}{1+e^t}$$

$$P(y_i = 0|x_i) = \frac{1}{1+e^t}$$

where $t = \beta_0 + x^T\beta$

(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]

Solutions:

$$P(y|x) = P(y = 1|x)^y(1 - P(y = 1|x))^{(1-y)}$$

Let $t = \beta_0 + x_i^T\beta$

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

Table 1: probability distribution for X

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

Solution:

The likelihood function is:

$$\begin{aligned} L(\theta) &= (\theta^2)^1[2\theta(1-\theta)]^2(\theta^2)^1(1-2\theta)^4 \\ &= 4\theta^6(1-\theta)^2(1-2\theta)^4 \end{aligned}$$

The log-likelihood function is:

$$l(\theta) = lnL(\theta) = ln4 + 6ln\theta + 2ln(1-\theta) + 4\ln(1-2\theta)$$

Let

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \frac{6}{\theta} - \frac{2}{1-\theta} - \frac{8}{1-2\theta} \\ &= \frac{24\theta^2 - 28\theta + 6}{\theta(1-\theta)(1-2\theta)} = 0 \\ &\Rightarrow 12\theta^2 - 14\theta + 3 = 0 \\ \theta_{1,2} &= \frac{7 \pm \sqrt{13}}{12} \end{aligned}$$

Note that $0<\theta<\frac{1}{2}$, so $\hat{\theta}^{MLE} = \frac{7-\sqrt{13}}{12}$.

(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]

Solution:

With the prior, the posterior becomes

$$\begin{aligned} P(D|\theta)P(\theta) &= \begin{cases} 0.5 \cdot 0.5^{N_1}(1-0.5)^{N-N_1} & \theta = 0.5 \\ 0.5 \cdot 0.3^{N_1}(1-0.3)^{N-N_1} & \theta = 0.3 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 0.5^{N+1} & \theta = 0.5 \\ 0.5 \cdot 0.3^{N_1} 0.7^{N-N_1} & \theta = 0.3 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Since the value of θ only can be taken 0.5 or 0.3, we just need to compare two posteriors as follows:

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \begin{cases} 0.5 & \text{if } 0.5^{N+1} > 0.5 \cdot 0.3^{N_1} 0.7^{N-N_1}, \\ 0.3 & \text{if } 0.5^{N+1} < 0.5 \cdot 0.3^{N_1} 0.7^{N-N_1}. \end{cases}$$

Here, we don't consider the case of $0.5^{N+1} = 0.5 \cdot 0.5^{N_1} 0.7^{N-N_1}$. After some simple computations, we have the solution:

$$\hat{\theta}^{MAP} = \begin{cases} 0.5 & \text{if } N < \frac{\ln 7 - \ln 3}{\ln 7 - \ln 5} N_1, \\ 0.3 & \text{if } N > \frac{\ln 7 - \ln 3}{\ln 7 - \ln 5} N_1. \end{cases}$$

(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:

When N is small, the prior in (b) leads a better estimate since the prior is a summary of our subjective beliefs about the data. When N is large, the Beta prior distribution is better according to the law of large number.

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]

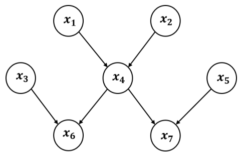

Figure 1: The Bayesian network for questions 3.

Solution:

(a) (1) The statement is True. According to D-separation, $\{x_1, x_2\}$ and $\{x_6, x_7\}$ can be regarded as two sets A and B. All the arrows on the path form A to B meet head-to-tail, therefore all the paths are blocked given x_4.

  (2) The statement is False. The arrow on the path from x_3 to x_4 meets head-to-head. Since the node x_6 is observed, the path from x_3 to x_4 is not blocked. The path from x_6 to x_2 is also unblocked, therefore x_3 and x_2 are not conditionally independent.

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

$$
\begin{aligned}
\log p(\mathcal{D} \mid \theta) &= \log \prod_{k=1}^{K} p(x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k}, x_{7k} \mid \theta) \\
&= \log \prod_{k=1}^{K} p(x_{1k} \mid \theta) p(x_{2k} \mid \theta) p(x_{3k} \mid \theta) p(x_{5k} \mid \theta) p(x_{4k} \mid x_{1k}, x_{2k}, \theta) p(x_{6k} \mid x_{3k}, x_{4k}, \theta) p(x_{7k} \mid x_{4k}, x_{5k}, \theta) \\
&= \sum_{k=1}^{K} \log p(x_{1k} \mid \theta) + \log p(x_{2k} \mid \theta) + \log p(x_{3k} \mid \theta) + \log p(x_{5k} \mid \theta) + \log p(x_{4k} \mid x_{1k}, x_{2k}, \theta) \\
&\quad + \log p(x_{6k} \mid x_{3k}, x_{4k}, \theta) + \log p(x_{7k} \mid x_{4k}, x_{5k}, \theta).
\end{aligned}
$$

Then we derive the gradient of $\log p(\mathcal{D} \mid \theta)$ with respect to $\theta_{x_6|i,j}$

$$
\frac{\partial \log p(\mathcal{D} \mid \theta)}{\partial \theta_{x_6|i,j}} = \sum_{k=1}^{K} \frac{\partial p(x_{6k} \mid x_{3k}, x_{4k}, \theta)}{\partial \theta_{x_6|i,j}}
$$

Set the derivative to 0 and then obtain the parameter $\theta_{x_6|i,j}$

$$
\theta_{x_6|i,j} = \frac{\sum_{k=1}^{K} \mathbb{I}(x_{6k} = 1, x_{3k} = i, x_{4k} = j)}{\sum_{k=1}^{K} \mathbb{I}(x_{3k} = i, x_{4k} = j)},
$$

where $\mathbb{I}(\cdot)$ is the indicator function.