CS534 Written Homework

Yongsung CHO

1. (a) $f(z) = \log(1+z)$, $z = \mathbf{x}^T\mathbf{x}$, $\mathbf{x} \in R^D$

$$z = g(x) = x^T x$$

$$f(g(x)) = \log(1+g(x))$$

$$f'(g(x))\, g'(x) = \frac{1}{1+g(x)} \times g'(x)$$

$$\therefore f'(z) = \frac{1}{1+g(x)} = \frac{1}{1+x^T x}$$

(b)
$$f(z) = \exp^{-\frac{1}{2}z}$$
$$z = g(\mathbf{y}) = \mathbf{y}^T S^{-1}\mathbf{y}$$
$$\mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \mu$$
where $\mathbf{x}, \mu \in R^D, S \in R^{D\times D}$

$$f(g(h(x))) = \exp^{-\frac{1}{2}g(h(x))}$$

$$f'(g(h(x)))\cdot \cancel{g'(h(x))}\cdot \cancel{h'(x)} = -\frac{1}{2}\cancel{g'(h(x))}\,\cancel{h'(x)} \cdot \exp^{-\frac{1}{2}g(h(x))}$$

$$f'(g(h(x))) = -\frac{1}{2}\exp^{-\frac{1}{2}(x-\mu)^T S^{-1}(x-\mu)}$$

2. (a) $\dfrac{n_{fair}}{n_{all}} = \dfrac{1}{2}$

(b) $\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{10} = \frac{1}{4} + \frac{1}{20} = \frac{6}{20} = \frac{3}{10}$

(c) $P(\text{fair}) = \frac{1}{2}$ $\quad P(\text{head2} \mid \text{fair}) = \frac{1}{4}$

$P(\text{false}) = \frac{1}{2}$ $\quad P(\text{head2} \mid \text{false}) = \frac{1}{100}$

$P(\text{head2}) = \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{10}\times\frac{1}{10} = \frac{13}{100}$

The question that I need to know is $p(\text{fair} \mid \text{head2})$

According to Bayes rule, $p(\text{fair} \mid \text{head2}) = p(\text{head2} \mid \text{fair})\,\dfrac{p(\text{fair})}{p(\text{head2})}$

$$= \frac{1}{4} \times \frac{\frac{1}{2}}{\frac{13}{100}} = \frac{25}{26}$$

3. (a) (3pts) Write down the likelihood function of θ.

$x_1, x_2, x_3, \cdots x_n$ are independent and distributed uniform from 0 to θ

let $\begin{cases} \frac{1}{\theta} & x_i = 1 \\ 1-\frac{1}{\theta} & x_i = 0 \end{cases}$

so, $f_{x_i}(x_i \mid \theta) = \frac{1}{\theta}^{x_i} (1-\frac{1}{\theta})^{1-x_i}$

likelihood function $L(\theta) = \prod_{i=1}^{n} f_{x_i}(x_i \mid \theta)$

$$= \prod_{i=1}^{n} \frac{1}{\theta}^{x_i} (1-\frac{1}{\theta})^{1-x_i}$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log (\frac{1}{\theta})^{x_i} (1-\frac{1}{\theta})^{1-x_i}$$

(b) (4pts) Find the maximum likelihood estimator for θ.

$$l(\theta) = \sum_{i=1}^{n} x_i \log(\frac{1}{\theta}) + \sum_{i=1}^{n} (1-x_i) \log(1-\frac{1}{\theta})$$

$$= n_1 \log(\frac{1}{\theta}) + n_2 \log(1-\frac{1}{\theta})$$

$$\frac{dl(\theta)}{d\theta} = n_1 \theta \times (-\frac{1}{\theta^2}) + \frac{n_2 \theta}{\theta - 1} \times (\frac{1}{\theta^2})$$

$$0 = -\frac{n_1}{\theta} + \frac{n_2}{\theta(\theta-1)}$$

$$\frac{n_1}{\cancel{\theta}} = \frac{n_2}{\cancel{\theta}(\theta-1)} \quad \because \theta > 0$$

$$n_1 \theta - n_1 = n_2$$

$$\theta = \frac{n_1 + n_2}{n_1} = \frac{N}{n_1}$$

4. (12pts) (Maximum likelihood estimation of categorical distribution.) A DNA sequence is formed using four bases Adenine(A), Cytosine(C), Guanine(G), and Thymine(T). We are interested in estimating the probability of each base appearing in a DNA sequence. Here we consider each base as a random variable x following a categorical distribution of 4 values (a, c, g and t) and assume a sequence is generated by repeatedly sampling from this distribution. This distribution has 4 parameters, which we denote as p_a, p_c, p_g, and p_t. Given a collection of DNA sequences with accumulated length of N, we counted the number of times that we observe of the four values, denoted by n_a, n_c, n_g and n_t respectively. Please show that the maximum likelihood estimation for p_x is $\frac{n_x}{N}$, where $x \in \{a, c, g, t\}$. Note that the four parameters are constrained to sum up to 1. This can be captured as a constrained optimization problem, solved using the method of Lagrange multiplier.

Helpful starting point: the probability mass function for the discrete random variable can be written compactly as

$$p(x) = \prod_{s=a,c,g,t} p_s^{I(x=s)}$$

Here $I(x = s)$ is an indicator function, and takes value 1 if x is equal to s, and 0 otherwise.

$$I(x=s) \begin{cases} 1 & \text{true} \\ 0 & \text{false} \end{cases}$$

$$L(p) = \prod_{i=1}^{N} p_a^{x_a} p_c^{x_c} p_g^{x_g} p_t^{x_t}$$

$$L(p) = \prod_{i=1}^{N} p_a^{x_a} (p_c + p_g + p_t)^{x_c + x_g + x_t}$$

$$= \prod_{i=1}^{N} p_a^{x_a} (1-p_a)^{1-x_a}$$

$$l(p) = \log L(p) = \sum_{i=1}^{N} \log p_a^{x_a} (1-p_a)^{1-x_a}$$

$$= \sum_{i=1}^{N} x_a \log p_a + \sum_{i=1}^{N} (1-x_a) \log(1-p_a)$$

$$l(p) = n_a \log p_a + n_{others} \log(1-p_a)$$

$$\frac{dl(p)}{dp} = \frac{n_a}{p_a} - \frac{n_{others}}{1-p_a} \qquad \frac{n_a}{p_a} = \frac{n_{others}}{1-p_a}$$

$$\because n_a + n_{others} = N \qquad p_a = \frac{n_a}{N}$$

in the same way, $P_c = \frac{n_c}{N}$

$P_g = \frac{n_g}{N}$

$P_t = \frac{n_t}{N}$

$\therefore P_x = \frac{n_x}{N}$

5. (Expected loss). Sometimes the cost of classification is not symmetric, one type of mistake is much more costly than the other. For example, the cost of misclassifying a normal email as spam can be substantially higher than letting a spam slip through. This can be captured by using a mis-classification loss matrix like the following:

predicted label $\hat{y}$	true label y 0	1
0	0	10
1	5	0

where misclassifying a positive example ($y = 1$, $\hat{y} = 0$) has a cost of 10, and misclassifying a negative example ($y = 0$, $\hat{y} = 1$ has a smaller cost of 5.

Suppose we have a probabilistic model that estimates $P(y = 1|\mathbf{x})$ for given $\mathbf{x}$. Here we will go through some questions to figure out how to prediction for $\mathbf{x}$ so what the expected loss is minimized.

(a) (2pts) Say $P(y = 1|\mathbf{x}) = 0.4$, what is the expected loss of predicting $\hat{y} = 1$?

$P(y=1|x) = 0.4$

$P(y=0|x) = 0.6$

$\therefore$ expected loss of predicting $\hat{y}=1$ $= 0.6 \times 5 = 3.$

(b) (3pts) What is the best prediction that minimizes the expected loss?

expected loss of predicting $y=1$ $= 0.4 \times 10 = 4$

$E(L(y,\hat{y})|X=x) = \Sigma p(y|x) = 1 - p(\hat{y}|x)$

$= -2$

(c) (4pts) Show that to minimize the expected loss for our decision for this loss matrix, we should set a probability threshold θ and predict $\hat{y} = 1$ if $P(y = 1|x) > \theta$ and $\hat{y} = 0$ otherwise.

$$P(\hat{y}|x) = \theta \times L(\hat{y}, 0) + (1-\theta)\, L(\hat{y}, 1)$$

$$P(0|x) = 10 - \theta \times 10$$
$$P(1|x) = \theta \times 5 \qquad \theta = \frac{1}{20} = 0.05$$

(d) (3pts) Show a loss matrix where the threshold is 0.1.

$$\theta = \frac{\alpha}{\alpha+\beta} = 0.1 \qquad \alpha = 1 \quad \beta = 9$$

$$1-\theta = \frac{\beta}{\alpha+\beta} = 0.9 \qquad \therefore \begin{pmatrix} 0 & 9 \\ 1 & 0 \end{pmatrix}$$