Machine Learning **(Due: 16th May)**

Assignment #3 (Probability theory)

Instructor: Beilun Wang *Name:* Qipeng Zhu, *ID:* 58119304

Problem Description:

Problem 1: Probability and statistics

(Note: the tables of related statistics used in problems are attached at the end of the document)

(1) Let C and D be two events. Suppose $P(C) = 0.5$, $P(C \cap D) = 0.2$ and $P(\overline{C \cup D}) = 0.4$. What is $P(D)$?

(2) Suppose X is a random variable with CDF(Cumulative Distribution Function)

$$F(x) = \begin{cases} 0 & for\ x < 0 \\ x(2-x) & for\ 0 \leqslant x \leqslant 1 \\ 1 & for\ x > 1 \end{cases}$$

(a) Find $E(X)$.
(b) Find $P(X < 0.4)$.

(3) Let X have range $[0, 3]$ and density $f_X(x) = kx^2$ among it. Let $Y = X^3$.
(a) Find k and the cumulative distribution function of X.
(b) Find the probability density function $f_Y(y)$ for Y.

(4) Data was taken on height and weight from the entire population of 700 mountain gorillas living in the Democratic Republic of Congo:

height \ weight	light	average	heavy
short	170	70	30
tall	85	190	155

Let X encode the weight, taking the values of a randomly chosen gorilla: 0, 1, 2 for light, average, and heavy respectively.
Likewise, let Y encode the height, taking values 0 and 1 for short and tall respectively.

(a) Determine the joint PMF(Probability Mass Function) of X and Y and the marginal PMFs of X and of Y.
(b) Are X and Y independent?
(c) Find the covariance of X and Y.
(d) Find the correlation of X and Y.
For part (c) and (d), you need to give a numerical (no variables inside) expression, but you can leave it unevaluated.

(5) Suppose a researcher collects $x_1, ..., x_n$ i.i.d. measurements of the background radiation in Boston. Suppose alse that these observations follow a Rayleigh distribution with parameter τ, with PDF(Probability Density Function) given by

$$f(x) = x\tau e^{-\frac{1}{2}\tau x^2}.$$

Find the maximum likelihood estimate for τ.

(6) You independently draw 100 data points from a normal distribution.

Suppose you know the distribution is $\mathcal{N}(\mu, 4)(\sigma^2 = 4)$ and you want to test the null hypothesis $H_0 : \mu = 3$ against the alternative hypothesis $H_A : \mu \neq 3$. If you want a significance level of $\alpha = 0.05$. What is your rejection region?

You must clearly state what test statistic you are using.

(**Hint:** for $Z \sim \mathcal{N}(0, 1)$, we have $\Phi(-1.96) = P(Z \leqslant -1.96) = 0.025$).

(7) Data is collected on the time between arrivals of consecutive taxis at a downtown hotel. We collect a data set of size 45 with sample mean $\bar{x} = 5.0$ and sample standard deviation $s = 4.0$.

Assume the data follows a normal random variable.

(a) Find an 80% confidence interval for the mean μ of X.

(b) Find an 80% χ^2-confidence interval for the variance.

Problem 2: Classification and Logistic Regression

Let $(X, C) \in \mathbb{R}^p \times \{0, 1\}$ be a random vector pair subject to $P(C = c) = \pi_c\,(\pi_0 + \pi_1 = 1)$. Here we treat C as the "class" of X, and the class for sample X_i is C_i.

(1) Assume that the conditional distribution of X given C is $X|C \sim \mathcal{N}(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C)$, where $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1 \in \mathbb{S}_{++}^{p \times p}$ are the mean vectors and covariance matrices(both are PD) for each class respectively. Write down the PDF(Probability Density Function) for X without given C.

(2) Under the assumption above, write down the condition that the given observation X_i will be classified into either class through Bayes Classifier. Recall that Bayes Classifier selects the class that maximizes the conditional probability of C given X.

(3) Under the assumption above, further assume that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. Write down the decision boundary of Bayes Classifier, and show that it forms a hyperplane. What about the boundary under the general condition $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$? You can illustrate it with specific $\boldsymbol{\Sigma}_c$ you choose.

(4) In Logistic Regression settings, we estimate that

$$\hat{P}(C = 1|X = \boldsymbol{x}; \boldsymbol{\theta}) = 1/(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x}))$$

and

$$\hat{P}(C = 0|X = \boldsymbol{x}; \boldsymbol{\theta}) = 1 - \hat{P}(C = 1|X = \boldsymbol{x}; \boldsymbol{\theta}).$$

Recall that the Kullback–Leibler divergence D_{KL} from distribution Q to P is defined by

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Show that minimizing the summation of the Kullback–Leibler divergence from $\hat{P}(C = c|X = X_i; \boldsymbol{\theta})$ to $P(C = c|X = X_i)$ for each sample $X_i (i = 1, 2, \cdots, n)$ is equivalent to the maximum likelihood estimate for $\boldsymbol{\theta}$. Here $P(C = c|X = X_i)$ represents the real probability and $\hat{P}(C = 1|X = \boldsymbol{x}; \boldsymbol{\theta})$ denotes the estimate.

Answer:

Problem 1: Probability and statistics

(1) $P(C \cup D) = 1 - P(\overline{C \cup D}) = 0.6$ And $\because$ $P(C \cup D) = P(C) + P(D) - P(C \cap D)$ $\therefore$ $P(D) = 0.3$

(2)

(a)

$$f(x) = F'(x) = \begin{cases} 2 - 2x & \text{if } 0 \leqslant x \leqslant 1 \\ 0 & \text{others} \end{cases}$$

$$E(x) = \int_{-\infty}^{+\infty} xf(x)\ dx = \int_0^1 2x - 2x^2\ dx = \frac{1}{3}$$

(b)

$$P(x < 0.4) = F(0.4) = 0.64$$

(3)

(a)

$$\int_{-\infty}^{+\infty} f_x(x)\ dx = \int_0^3 kx^2\ dx = 9k = 1$$

$$\therefore k = \frac{1}{9}$$

$\because X \in [0, 3]$ $\therefore Y = X^3 \in [0, 27]$ $\therefore$ *when* $y < 0, F_Y(Y \leqslant y) = 0;$ *when* $y > 27, F_Y(Y \leqslant y) = 1$

$$when\ 0 \leqslant y \leqslant 27:\ F_Y(Y \leqslant y) = F_Y(X^3 \leqslant y) = F_Y(X \leqslant \sqrt[3]{y}) = \int_0^{\int_0^3 kx^2\ dx} \frac{1}{9}x^2\ dx = \frac{y}{3}$$

So

$$F_Y(Y \leqslant y) = \begin{cases} \frac{y}{3} & \text{if } 0 \leqslant x \leqslant 27 \\ 0 & \text{others} \end{cases}$$

(b)

$$f_Y(y) = F'_Y = \begin{cases} \frac{1}{3} & \text{if } 0 \leqslant y \leqslant 27 \\ 0 & \text{others} \end{cases}$$

(4)

(a)

$$f(x, y) = P(X = x, Y = y) = \begin{cases} \frac{17}{70} & \text{if } x = 0, y = 0 \\ \frac{1}{10} & \text{if } x = 1, y = 0 \\ \frac{3}{70} & \text{if } x = 2, y = 0 \\ \frac{17}{140} & \text{if } x = 0, y = 1 \\ \frac{19}{70} & \text{if } x = 1, y = 1 \\ \frac{31}{140} & \text{if } x = 2, y = 1 \end{cases}$$

$$f_X(x) = P(X = x) = \begin{cases} \frac{51}{140} & \text{if } x = 0 \\ \frac{13}{35} & \text{if } x = 1 \\ \frac{37}{140} & \text{if } x = 2 \end{cases}$$

$$f_Y(y) = P(Y = y) = \begin{cases} \frac{27}{70} & \text{if } y = 0 \\ \\ \frac{43}{70} & \text{if } y = 1 \end{cases}$$

(b)

$$\because P(X = 0)P(Y = 0) = \frac{1377}{9800} \neq P(X = 0, Y = 0), \ \therefore X\ and\ Y isnotindependent$$

(c)

$$cov(X, Y) = EXY - EX \cdot EY$$

$$EXY = \sum_{i=0}^{2}\sum_{j=0}^{1} x_i y_j P_{ij} = \frac{5}{7}$$

$$EX = \sum_{i=0}^{2}\sum_{j=0}^{1} x_i P_{ij} = \frac{9}{10}$$

$$EY = \sum_{i=0}^{2}\sum_{j=0}^{1} 1 y_j P_{ij} = \frac{43}{70}$$

$$So\ cov(X, Y) = \frac{113}{200}$$

(d)

$$DX = EX^2 - E^2X = \frac{10}{7}$$

$$DY = EY^2 - E^2Y = \frac{1161}{4900}$$

$$\rho = \frac{cov(X, Y)}{\sqrt{DX \cdot DY}} \approx 0.9711$$

(5)

$$Let\ lnL(\tau) = ln\prod_{i=1}^{n}(x_i \tau e^{-\frac{1}{2}\tau x_i^2}) = nln\tau\ + \sum_{i=1}^{n} lnx_i - \frac{\tau}{2}\sum_{i=1}^{n} x_i^2$$

$$Then\ let\ \nabla_\tau lnL(\tau) = \frac{n}{r} - \frac{1}{2}\sum_{i=1}^{n} x_i^2 = 0,\ we\ can\ get\ \tau = \frac{2n}{\sum_{i=1}^{n} x_i^2}$$

(6)

$$Let\ U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n},\ then\ we\ can\ get\ U \sim N(0, 1)$$

$$\because\ significance\ level\ of\ \alpha = 0.05,\ \Phi(-1.96) = 0.025$$

$$\therefore\ \mu_{\frac{\alpha}{2}} = 1.96\ \therefore\ the\ rejection\ region\ is\ \{|U| \leqslant 1.96.\}$$

(7)

(a)

$$let\, T = \frac{\overline{x} - \mu}{s/\sqrt{n}},\ \because\ T \sim t(n-1)\ \therefore\ the\ 80\%\ confidence\ interval\ for\ the\ mean\ \mu\ of\ X\ is\ [\overline{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}, \overline{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}]$$

$$\because\ n = 45,\ \overline{x} = 5,\ s = 4,\ \alpha = 0.2,\ \therefore\ the\ 80\%\ confidence\ interval\ for\ the\ mean\ \mu\ of\ X\ is\ [4.2242, 5.7758].$$

(b)

$$Let\ \chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\therefore\ 80\%\ \chi^2 - confidence\ interval\ for\ the\ variance\ is\ [\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}]$$

$$\because\ n = 45,\ s = 4\ ,\therefore\ 80\%\ \chi^2 - confidence\ interval\ for\ the\ variance\ is\ [0.2296, 0.6670].$$

Problem 2: Classification and Logisitic Regression

(1)
It's clear that:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{i=0}^{1} P(\boldsymbol{X}|C = c_i)P(C = c_i)$$

$$= \frac{\Pi_0}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}} exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_0)) + \frac{\Pi_1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1))$$

(2)
From Bayes Classification Rule:

$$P(C = c_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C = c_i)P(C = c_i)}{P(\boldsymbol{X})} = \alpha P(\boldsymbol{X}|C = c_i)P(C = c_i)$$

So we can easily know:

$$lnP(C = c_i|\boldsymbol{X}) = K + ln\Pi_c - \frac{1}{2}ln|\boldsymbol{\Sigma}_c| - \frac{1}{2}ln(\boldsymbol{x} - \boldsymbol{\mu}_c)^t \boldsymbol{\Sigma}_c^{-t}(\boldsymbol{x} - \boldsymbol{\mu}_c)(K\ is\ a\ constant)$$

Suppose:

$$g_i(\boldsymbol{x}) = lnP(C = c_i|\boldsymbol{X} = \boldsymbol{x}_i) = ln\Pi_c - \frac{1}{2}ln|\boldsymbol{\Sigma}_c| - \frac{1}{2}ln(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^t \boldsymbol{\Sigma}_c^{-t}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)(i = 0, 1)$$

So the discriminant function is:

$$g(\boldsymbol{x}) = g_0(\boldsymbol{x}) - g_1(\boldsymbol{x})$$

When $g(\boldsymbol{x}_i) > 0$, $\boldsymbol{x}_i$ should be classfied into class0; when $g(\boldsymbol{x}_i) < 0$, $\boldsymbol{x}_i$ should be classfied into class1.

(3)
From problem(2), let g($\boldsymbol{x}$)=0, then we can get the classification boundary, if $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$:

$$g_0(\boldsymbol{x}) - g_1(\boldsymbol{x}) = ln\frac{\Pi_0}{\Pi_1} - \frac{1}{2}ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}ln(\boldsymbol{x}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_c^{-t}(\boldsymbol{x}_i - \boldsymbol{\mu}_0) + \frac{1}{2}ln(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}_c^{-t}(\boldsymbol{x}_i - \boldsymbol{\mu}_1) = 0$$

if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}[-2\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\mu_i] + ln\Pi_i\ (i = 0, 1)$$

Let:

$$g_0(\boldsymbol{x}) - g_1(\boldsymbol{x}) = 0$$

Then we can get the classification boundary:

$$[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)]^t[\boldsymbol{x} - (\frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} - \frac{1}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)} ln\frac{\Pi_0}{\Pi_1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1))] = 0$$

(4)
From the title, we can know:

$$let\ D_{\mathrm{KL}}(P||\hat{P}) = \sum_{i=1}^{n} P(C = c|X = X_i) \log \frac{P(C = c|X = X_i)}{\hat{P}(C = c|X = X_i; \boldsymbol{\theta})}.$$

Simplify it, we can get:

$$-\sum_{i=1}^{n} \log(\hat{P}(C = c|X = X_i; \boldsymbol{\theta})$$

. So minimizing it is equal to maximizing:

$$\sum_{i=1}^{n} \log(\hat{P}(C = c|X = X_i; \boldsymbol{\theta}).$$

On the other hand, Suppose:

$$\log L(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} (\hat{P}(C = c|X = X_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log(\hat{P}(C = c|X = X_i; \boldsymbol{\theta}).$$

So when we use the maximize likelihood method, we should maximize it to estimate $\boldsymbol{\theta}$. So minimizing the summation of the Kullback–Leibler divergence from $\hat{P}(C = c|X = X_i; \boldsymbol{\theta})$ to $P(C = c|X = X_i)$ for each sample $X_i (i = 1, 2, \cdots, n)$ is equivalent to the maximum likelihood estimate for $\boldsymbol{\theta}$.

t-table of left tail probabilities (The table shows $P(T < t)$ for $T \sim t(n)$.)

n \ t	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
44	0.5000	0.5788	0.6545	0.7242	0.7860	0.8386	0.8817	0.9157	0.9416	0.9606	0.9742
45	0.5000	0.5788	0.6545	0.7242	0.7860	0.8387	0.8818	0.9158	0.9417	0.9607	0.9742

Table of χ^2 critical values (right-tail) (The table shows $c_{n,p}$ = the $1 - p$ quantile of $\chi^2(n)$.)

n \ p	0.010	0.025	0.050	0.100	0.200	0.300	0.500	0.700	0.800	0.900	0.950	0.975	0.990
44	68.71	64.20	60.48	56.37	51.64	48.40	43.34	38.64	35.97	32.49	29.79	27.57	25.15
45	69.96	65.41	61.66	57.51	52.73	49.45	44.34	39.58	36.88	33.35	30.61	28.37	25.90