CSE575: Statistical Machine Learning (Fall 2019)
Midterm Exam 1 Grading Keys

Instructions:

- Print your name and your ASU ID on the cover page of the paper.
- This is a **closed book** examination. One letter-sized cheat sheet (with your name printed on it) is allowed. You have to hand in your cheat sheet with your exam paper.
- Electronic devices (except a calculator) are not allowed.
- There are weights associated with the problems and sub-problems. Optimize your time/effort to achieve the highest possible score.
- Read the questions/instructions carefully.
- Use the space provided below each question to write your answer.

NAME	
ASUID	

Question	Score
P1	
P2	
P3	
P4	
P5	
Total	

P1. Maximum Likelihood Estimations (30 marks: 5+5+10+10)

(A) Assume that you have a data set $\mathcal{Y} = \{y_i | 1 \le i \le N\}$ where $y_i \in \{1, 2\}$, $i = 1, 2, \ldots, N$. The data are i.i.d., and follow the Bernoulli distribution: $y_i \sim$ Bernoulli(θ), *i.e.*,

$$P(y_i = 1|\theta) = \theta, \quad P(y_i = 2|\theta) = 1 - \theta.$$

Use MLE to estimate the parameter θ. In addition to the estimate of θ, you have to **write out the likelihood function** $P(\mathcal{Y}|\theta)$, or the log of the likelihood function. You may assume that N_1 of the y_i have value 1, and N_2 of the y_i have value 2.

Solutions: The likelihood function is $P(\mathcal{Y}|\theta) = \theta^{N_1} \times (1-\theta)^{N_2}$. The log of the likelihood is $\ln P(\mathcal{Y}|\theta) = N_1 \ln\theta + N_2 \ln(1-\theta)$. The MLE estimate is $\theta^{MLE} = \frac{N_1}{N_1+N_2} = \frac{N_1}{N}$.

Grading: 3 marks for θ^{MLE}; 2 marks for the likelihood or the log of the likelihood.

(B) Assume that you have a data set $\mathcal{X} = \{x_i | 1 \le i \le N\}$ where $x_i \in R$, $i = 1, 2, \ldots, N$. The data are i.i.d., and $x_i \sim \mathcal{N}(\mu, 1)$. In other words,

$$P(x_i|\mu) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{(x_i - \mu)^2}{2}\}.$$

Use MLE to estimate the parameter μ. In addition to the estimate of μ, you have to **write out the likelihood function** $P(\mathcal{X}|\mu)$, or the log of the likelihood function.

Solutions: The likelihood function is $P(\mathcal{X}|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} exp\{-\frac{(x_i-\mu)^2}{2}\}$. The log of the likelihood is $\ln P(\mathcal{X}|\mu) = -\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2} - \frac{N}{2}\ln(2\pi)$. The MLE estimate is $\mu^{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Grading: 3 marks for μ^{MLE}; 2 marks for the likelihood or the log of the likelihood.

(C) Assume that we have a dataset $\mathcal{D} = \{(x_i, y_i) | 1 \le i \le N\}$ where each $x_i \in R$ denotes an i.i.d sample and $y_i \in \{1, 2\}$ is the corresponding label. Assume $x_i | y_i = 1 \sim \mathcal{N}(\mu_1, 1)$, $x_i | y_i = 2 \sim \mathcal{N}(\mu_2, 1)$ and $y_i \sim$ Bernoulli(θ). That is,

$$\begin{aligned}
p(x_i|y_i = 1, \mu_1) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i - \mu_1)^2}{2}),\\
p(x_i|y_i = 2, \mu_2) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i - \mu_2)^2}{2}),\\
p(y_i = 1|\theta) &= \theta,\\
p(y_i = 2|\theta) &= 1 - \theta.
\end{aligned}$$

Derive the MLE estimates for parameters μ_1, μ_2, and θ, respectively. (You can assume that the first N_1 samples $\{x_1, \cdots, x_{N_1}\}$ are labeled as 1, and the last $N_2 = N - N_1$ samples $\{x_{N_1+1}, \cdots, x_N\}$ are labeled as 2.)

Solutions: The MLE estimates are
$\theta^{MLE} = \frac{N_1}{N_1+N_2} = \frac{N_1}{N}$; $\mu_1^{MLE} = \frac{1}{N_1}\sum_{i=1}^{N_1} x_i$; and $\mu_2^{MLE} = \frac{1}{N_2}\sum_{i=N_1+1}^{N_1+N_2} x_i$.

We have

$$
\begin{aligned}
p(x, y|\mu_1, \mu_2, \theta) &= p(x|y, \mu_1, \mu_2, \theta) \times p(y|\mu_1, \mu_2, \theta) && (1)\\
&= p(x|y, \mu_y) \times p(y|\theta). && (2)
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
p(\mathcal{D}|\mu_1, \mu_2, \theta) &= \prod_{i=1}^{N} p(x_i|y_i, \mu_{y_i}) \times p(y_i|\theta)\\
&= (\prod_{i=1}^{N_1} p(x_i|y_i, \mu_{y_i}) \times p(y_i|\theta)) \times (\prod_{i=N_1+1}^{N} p(x_i|y_i, \mu_{y_i}) \times p(y_i|\theta))\\
&= (\prod_{i=1}^{N_1} \frac{1}{\sqrt{2\pi}} exp\{-\frac{(x_i-\mu_1)^2}{2}\} \cdot \theta) \cdot (\prod_{i=N_1+1}^{N} \frac{1}{\sqrt{2\pi}} exp\{-\frac{(x_i-\mu_2)^2}{2}\} \cdot (1-\theta)).
\end{aligned}
$$

Here from the second equality we have used the simplification assumption. Taking the derivative with respect to each of the parameters and set the result to zero, we obtain the MLE estimates.

Grading: 6 marks for the three estimations; 4 marks for the likelihood or the log of the likelihood.

(D) We are drawing i.i.d data $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ sampled from a uniform distribution $U(-w, w)$. That is

$$
p(x|w) = \begin{cases} \frac{1}{2w}, & \text{if } -w \le x \le w \\ 0, & \text{otherwise.} \end{cases}
$$

What is the likelihood function $P(\mathcal{X}|w)$? Suppose that we have observed $\mathcal{X} = \{1, 2, 3, 4\}$. What is the maximum likelihood estimation of the parameter w? Write down your derivations.

Solutions: The likelihood function is

$$
P(\mathcal{X}|w) = \begin{cases} (\frac{1}{2w})^4, & \text{if } |x_i| \le w, i = 1, 2, 3, 4 \\ 0, & \text{otherwise.} \end{cases}
$$

Suppose we have observed $\mathcal{X} = \{1, 2, 3, 4\}$. The likelihood is 0 if $w < 4$, and $(\frac{1}{2w})^4$ otherwise. Therefore the MLE estimation is $w^{MLE} = 4$.

Grading: 5 marks for the first part; 5 marks for the second part. Take 2 marks off if 0 is not considered in the likelihood.

P2. Continuous Bayes Classifier (20 marks: 3+3+3+4+4+3)

We want to build a Bayes classifier for a binary classification task ($y = 1$ or $y = 2$) with a 1-dimensional input feature (x). We know the following quantities: (1) $P(y = 1) = 0.8$; (2) $P(x|y = 1) = \frac{1}{3}$ for $2 \le x \le 5$ and $P(x|y = 1) = 0$ otherwise; and (3) $P(x|y = 2) = \frac{1}{3}$ for $3 \le x \le 6$ and $P(x|y = 2) = 0$ otherwise.

(A) What is the prior for class label $y = 2$?

Solutions: $P(y = 2) = 1 - P(y = 1) = 1 - 0.8 = 0.2$.

Grading: 3 marks for correct answer, 1 mark for reasonable derivation but incorrect answer.

(B) What is $P(y = 1|x)$?

Solutions:

$$P(y=1|x) = \frac{P(x|y=1) \times P(y=1)}{P(x|y=1) \times P(y=1) + P(x|y=2) \times P(y=2)}.$$

We consider three cases: (1) $2 \le x < 3$, (2) $3 \le x \le 5$, and (3) otherwise.

When $x \in [2, 3)$, we have

$$P(y=1|x) = \frac{P(x|y=1) \times P(y=1)}{P(x|y=1) \times P(y=1) + P(x|y=2) \times P(y=2)} \tag{3}$$

$$= \frac{P(x|y=1) \times P(y=1)}{P(x|y=1) \times P(y=1)} = 1. \tag{4}$$

When $x \in [3, 5]$, we have

$$P(y=1|x) = \frac{P(x|y=1) \times P(y=1)}{P(x|y=1) \times P(y=1) + P(x|y=2) \times P(y=2)} \tag{5}$$

$$= \frac{\frac{1}{3} \times 0.8}{\frac{1}{3} \times 0.8 + \frac{1}{3} \times 0.2} \tag{6}$$

$$= \frac{4}{5}. \tag{7}$$

In all other cases, the numerator is zero, hence the probability is equal to zero. In summary, we have

$$P(y=1|x) = \begin{cases} 1, & 2 \le x < 3 \\ \frac{4}{5}, & 3 \le x \le 5 \\ 0, & \text{otherwise.} \end{cases}$$

Grading: 1 mark for each correct case.

(C) What is $P(y = 2|x)$?

Solutions:

$$P(y = 2|x) = \frac{P(x|y = 2) \times P(y = 2)}{P(x|y = 1) \times P(y = 1) + P(x|y = 2) \times P(y = 2)}.$$

We consider three cases: (1) $3 \le x \le 5$, (2) $5 < x \le 6$, and (3) otherwise.

When $x \in [3, 5]$, we have

$$\begin{aligned} P(y = 2|x) &= \frac{P(x|y = 2) \times P(y = 2)}{P(x|y = 1) \times P(y = 1) + P(x|y = 2) \times P(y = 2)} && (8) \\ &= \frac{\frac{1}{3} \times 0.2}{\frac{1}{3} \times 0.2 + \frac{1}{3} \times 0.8} && (9) \\ &= \frac{1}{5}. && (10) \end{aligned}$$

When $x \in (5, 6]$, we have

$$\begin{aligned} P(y = 2|x) &= \frac{P(x|y = 2) \times P(y = 2)}{P(x|y = 1) \times P(y = 1) + P(x|y = 2) \times P(y = 2)} && (11) \\ &= \frac{P(x|y = 2) \times P(y = 2)}{P(x|y = 2) \times P(y = 2)} = 1. && (12) \end{aligned}$$

In all other cases, the numerator is zero, hence the probability is equal to zero. In summary, we have

$$P(y = 2|x) = \begin{cases} \frac{1}{5}, & 3 \le x \le 5 \\ 1, & 5 < x \le 6 \\ 0, & \text{otherwise.} \end{cases}$$

Grading: 1 mark for each correct case.

(D) For $x = 2$, what is the class label your classifier will assign? What is the risk of this decision?

Solutions: $P(y = 1|x = 2) = 1$ and $P(y = 2|x = 2) = 0$. Hence we assign the class label 1. The risk of this assignment if 0.

Grading: 2 marks for each correct answer.

(E) For $x = 4$, what is the class label your classifier will assign? What is the risk of this decision?

Solutions: $P(y = 1|x = 4) = \frac{4}{5}$ and $P(y = 2|x = 4) = \frac{1}{5}$. Hence we assign the class label 1. The risk of this assignment if $\frac{1}{5}$.

Grading: 2 marks for each correct answer.

(F) What is the decision boundary of your Bayes classifier?

Solutions:

$$\text{label} = \begin{cases} 1, & 2 \le x \le 5 \\ 2, & 5 < x \le 6 \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

Grading: 1 mark for each correct answer. Take 1 mark off (overall) if the boundary is not clearly specified.

P3. Discrete Bayes Classifier (20 marks: 3+3+3+4+4+3)

We want to build a Bayes classifier for a binary classification task ($y = 1$ or $y = 2$) with one discrete feature x, where $x \in \{0, 1, 2, 3\}$. We know the following quantities: (1) $P(y = 1) = 0.6$; (2) $P(0|y = 1) = 0.3$, $P(1|y = 1) = 0.1$, $P(2|y = 1) = 0.4$, $P(3|y = 1) = 0.2$, and (3) $P(0|y = 2) = 0.4$, $P(1|y = 2) = 0.3$, $P(2|y = 2) = 0.2$, $P(3|y = 2) = 0.1$,

(A) What is the prior for class label $y = 2$?

Solutions: $P(y = 2) = 1 - P(y = 1) = 1 - 0.6 = 0.4$.

Grading: 3 marks for the correct answer; earn 1 mark for reasonable derivation, but incorrect answer.

(B) What is $P(y = 1|x = 3)$?

Solutions:

$$P(y = 1|x = 3) = \frac{P(x = 3|y = 1) \times P(y = 1)}{P(x = 3|y = 1) \times P(y = 1) + P(x = 3|y = 2) \times P(y = 2)} \quad (13)$$

$$= \frac{0.2 \times 0.6}{0.2 \times 0.6 + 0.1 \times 0.4} = \frac{3}{4}. \quad (14)$$

Grading: 3 marks for the correct answer; earn 1 mark for the correct formula, but incorrect answer.

(C) What is $P(y = 2|x = 3)$?

Solutions:

$$P(y = 2|x = 3) = \frac{P(x = 3|y = 2) \times P(y = 2)}{P(x = 3|y = 1) \times P(y = 1) + P(x = 3|y = 2) \times P(y = 2)} \quad (15)$$

$$= \frac{0.1 \times 0.4}{0.2 \times 0.6 + 0.1 \times 0.4} = \frac{1}{4}. \quad (16)$$

Grading: 3 marks for the correct answer; earn 2 mark for the correct formula, but incorrect answer.

(D) For $x = 1$, what is the class label your classifier will assign? What is the risk of this decision?

Solutions:

$$P(y=1|x=1) = \frac{P(x=1|y=1) \times P(y=1)}{P(x=1|y=1) \times P(y=1) + P(x=1|y=2) \times P(y=2)} \tag{17}$$

$$= \frac{0.1 \times 0.6}{0.1 \times 0.6 + 0.3 \times 0.4} = \frac{1}{3}. \tag{18}$$

Hence the class label is 2, with a risk of $\frac{1}{3}$.

Grading: 2 marks for each correct answer.

(E) For $x = 3$, what is the class label your classifier will assign? What is the risk of this decision?

Solutions: We have already computed $P(y = 1|x = 3 = \frac{3}{4}$. Hence the class label is 1, with a risk of $\frac{1}{4}$.

Grading: 2 marks for each correct answer.

(F) What is the decision boundary of your Bayes classifier?

Solutions:

$$P(y=1|x=0) = \frac{9}{17} > 0.5,$$
$$P(y=1|x=1) = \frac{1}{3} < 0.5,$$
$$P(y=1|x=2) = \frac{3}{4} > 0.5,$$
$$P(y=1|x=3) = \frac{3}{4} > 0.5.$$

Hence the decision boundary is as follows:

x	label
0	1
1	2
2	1
3	1

Grading: 1 mark each for $x = 0$ and $x = 2$; 0.5 mark each for $x = 1$ and $x = 3$.

P4. Naive Bayes Classifier (20 marks: 2+2+2+2+2+2+2+6)

Given the training data in Table 1, we want to train a binary classifier using Naive Bayes, with (1) the last column being the class label y, and (2) each column of X being a binary feature.

Feature $X = (x_1, x_2, x_3)$			Class Label y
Sky	**Humid**	**Wind**	**Enjoy Sport**
sunny	warm	strong	1
rainy	cold	mild	2
sunny	warm	mild	1
rainy	cold	strong	2
sunny	warm	strong	1
rainy	cold	mild	2

Table 1: Training Data Set for Naive Bayes Classifier

(A) What is $P(y = 1)$?

Solutions: $P(y = 1) = \frac{3}{6} = \frac{1}{2}$.

Grading: 2 marks for the correct answer.

(B) What is $P(x_1 = \text{rainy}|y = 1)$?

Solutions: $P(x_1 = \text{rainy}|y = 1) = \frac{0}{3} = 0$.

Grading: 2 marks for the correct answer.

(C) What is $P(x_2 = \text{cold}|y = 1)$?

Solutions: $P(x_2 = \text{cold}|y = 1) = \frac{0}{3} = 0$.

Grading: 2 marks for the correct answer.

(D) What is $P(x_3 = \text{strong}|y = 1)$?

Solutions: $P(x_3 = \text{strong}|y = 1) = \frac{2}{3}$.

Grading: 2 marks for the correct answer.

(E) What is $P(x_1 = \text{rainy}|y = 2)$?

Solutions: $P(x_1 = \text{rainy}|y = 2) = \frac{3}{3} = 1$.

Grading: 2 marks for the correct answer.

(F) What is $P(x_2 = \text{cold}|y = 2)$?

Solutions: $P(x_2 = \text{cold}|y = 2) = \frac{3}{3} = 1$.

Grading: 2 marks for the correct answer.

(G) What is $P(x_3 = \text{strong}|y = 2)$?

Solutions: $P(x_3 = \text{strong}|y = 2) = \frac{1}{3}$.

Grading: 2 marks for the correct answer.

(H) Suppose we have a new input vector $x = (\text{sunny}, \text{cold}, \text{strong})$. What is $P(y = 1|x)$? What is $P(y = 2|x)$? Which class label will the Naive Bayes classifier assign to this input?

Solutions:

$$P(y = 1|x) = \frac{P(x|y = 1) \times P(y = 1)}{P(x|y = 1) \times P(y = 1) + P(x|y = 2) \times P(y = 2)} = 0.$$

$$P(y = 2|x) = \frac{P(x|y = 2) \times P(y = 2)}{P(x|y = 1) \times P(y = 1) + P(x|y = 2) \times P(y = 2)} = 0.$$

NB will assign no label to this input.

Grading: 2 marks for the correct answer.

P5. Optimization (10 marks: 3+3+4)

(A) Let $f(x) = x^2 - 3x + 18$. What is the value of x that solves the following unconstrained optimization problem? $x^{opt} =$

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & -\infty < x < \infty \end{aligned}$$

Solutions: $f(x)$ is a convex quadratic function. $f'(x) = 2x - 3$. It takes zero when $x = 1.5$. Hence $x^{opt} = 1.5$.

Grading: 3 marks for the correct answer; 1 mark for knowing to set the derivative equal to zero, but with incorrect answer.

(B) Let $f(x) = x^2 - 3x + 18$. What is the value of x that solves the following constrained optimization problem? $x^{opt} =$

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & 4 \leq x \leq 8 \end{aligned}$$

Solutions: $f'(x) = 2x - 3 > 0$ for $x \in [4, 8]$. Hence $f(x)$ is monotonically increasing in the interval $[4, 8]$. Hence $x^{opt} = 4$.

Grading: 3 marks for the correct answer; 1 mark for reasonable argument, but with incorrect answer.

(C) Let $f(x)$ be a twice continuously differentiable function, and $\bar{x}$ minimizes $f(x)$ in $-\infty < x < \infty$. What can you say about $f'(\bar{x})$? What can you say about $f''(\bar{x})$?

Solutions: $f'(\bar{x}) = 0$. $f''(\bar{x}) \geq 0$.

Grading: 2 marks for each correct answer.