CSE575 HW01, Monday, 09/26/2022, Due: Friday, 10/07/2022

Please note that you have to typeset your assignment using either LATEX or Microsoft Word, and produce a PDF file for submission. Hand-written assignment (or photo of it) will not be graded. You need to submit an electronic version (in PDF form) on the canvas. You should name your file using the format CSE575-HW01-LastName-FirstName.

1. **Probability, MLE, and PAC [20 pts: 4+4+4+4+4]**

(A) Suppose that X and Y are **independent** events, and $p(Y) > 0, p(X) = 0.5$. What is the value of $p(X|Y)$?

Ans: Since the events X and Y are independent, $p(X,Y) = p(X) \times p(Y)$. Hence $p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X) \times p(Y)}{p(Y)} = p(X) = 0.5$. [4pts]

(B) Suppose that X and Y are **disjoint** events (i.e. $p(X,Y) = 0$) and $p(Y) > 0$. What is the value of $p(X|Y)$?

Ans: Since the events X and Y are disjoint, $p(X,Y) = 0$. Hence $p(X|Y) = \frac{p(X,Y)}{p(Y)} = 0$. [4pts]

(C) Suppose that we have two coins C_1 and C_2. The probability of C_1 having head is 0.6, and the probability of C_2 having head is 0.4. In each test, we toss both coins, and read the faces of C_1 and C_2 (note that we read C_2 **after** reading C_1). For example, if the toss resulted in C_1 head up and C_2 tail up, we will record the result as HT. Suppose we perform the test 4 times. What is the probability for us to observe the following result?

$$HT,\ HT,\ TT,\ TT?$$

Ans: For Coin C1, $p\,(C_1 = H) = 0.6,\ p(C_1 = T) = 0.4$.

For Coin C_2, $p(C_2 = H) = 0.4,\ p(C_2 = T) = 0.6$. Hence

$p(HT) = p(C_1 = H) \times p(C_2 = T) = 0.6 \times 0.6 = 0.36$. [1pt]

$p(TT) = p(C_1 = T) \times p(C_2 = T) = 0.4 \times 0.6 = 0.24$. [1pt]

Therefore, probability of the given result is,
$p(HT) \times p(HT) \times p\,(TT) \times p\,(TT) = 0.36 \times 0.36 \times 0.24 \times 0.24 = 0.00746496$. [2pts]

(D) You are given a coin and are asked to toss as many times as you wish to decide the probability of having heads-up for a toss of the coin. You tossed the coin 20 times, and observed 15 heads and 5 tails. What is your best estimate of the probability θ of having heads-up?

Ans: The likelihood is

$p(X|\theta) = \theta^{15} \times (1 - \theta)^5$ [2pts]

The MLE estimation of coin toss is

$\theta_{\mathrm{ML}} = \frac{15}{15+5} = 0.75.$ [2pts]

(E) If you want to be at least 99% sure that the difference between your estimated value of θ and the true probability of the coin having heads-up is no more than 0.1, how many tosses can guarantee this (hint: use the Hoeffding's inequality on slide 11 of Lecture05)? Please give the minimum number of tosses.

Ans: We will make use of Hoeffding's inequality (slide 11 of Lecture05), for any $\varepsilon > 0$,

$$p\left(\left|\theta_{\mathrm{ML}} - \theta^*\right| \geq \varepsilon\right) \leq 2\exp\left\{-2N\varepsilon^2\right\},$$

where θ^* is the true parameter, N is the total number of coin tosses.

Note that we have

$p(|\ \theta_{\mathrm{ML}} - \theta^*| \leq 0.1) = 1 - p(|\ \theta_{\mathrm{ML}} - \theta^*| \geq 0.1) \geq 1 - 2\exp\{-2N \times 0.01\}$ (1)

Therefore, in order to guarantee that $p(|\ \theta_{\mathrm{ML}} - \theta^*| \leq 0.1) \geq 0.99$, we must have

$1 - 2\exp\{-2N \times 0.01\} \geq 0.99$, which is equivalent to $\exp\left\{-\frac{N}{50}\right\} \leq \frac{1}{200}$. We then have

$N \geq 50 \ln 200 \approx 264.9$. Hence tossing 265 times [4pts] is sufficient.

Grading: students get 2pts for the correct inequality (1), and 3pts if their final answer is 264.

2. Discriminant Linear Classifiers [20 pts: 10+10]

You are given a training data set $\{x_n, t_n\}$ of size N = 21. Each input vector x_n is a point in the 2-dimensional Euclidean space R^2. We have $x_1 = (0,0)^\mathrm{T}$, $x_2 = (1,0)^\mathrm{T}$, $x_3 = (2,0)^\mathrm{T}$, $x_4 = (0,1)^\mathrm{T}$, $x_5 = (1,1)^\mathrm{T}$, $x_6 = (2,1)^\mathrm{T}$, $x_7 = (3,1)^\mathrm{T}$, $x_8 = (4,1)^\mathrm{T}$, $x_9 = (5,1)^\mathrm{T}$, $x_{10} = (100,1)^\mathrm{T}$, $x_{11} = (0,2)^\mathrm{T}$, $x_{12} = (1,2)^\mathrm{T}$, $x_{13} = (2,2)^\mathrm{T}$, $x_{14} = (3,2)^\mathrm{T}$, $x_{15} = (4,2)^\mathrm{T}$, $x_{16} = (5,2)^\mathrm{T}$, $x_{17} = (100,2)^\mathrm{T}$, $x_{18} = (3,3)^\mathrm{T}$, $x_{19} = (4,3)^\mathrm{T}$, $x_{20} = (5,3)^\mathrm{T}$, and $x_{21} = (100,3)^\mathrm{T}$. Each point is represented as a column vector.

There are two target classes C$_1$ and C$_2$. For each point x_n in the training set, x_n belongs to C$_1$ if its second coordinate is less than or equal to 2, and belongs to C$_2$ otherwise. If $x_n \in$ C$_1$, we have , t_n = 1. If $x_n \in$ C$_2$, we have t_n = 0 in the questions regarding least-squares linear discriminant and Fisher's linear discriminant.

(A) Compute the least-square linear classifier based on the training data (using $K = 2$ in slides of Lecture08 or textbook chapter 4.1.3). You need to write out (a) the error function [5pts], (b) the computed parameter matrix $\widetilde{\mathrm{W}}$ (a 3 by 2 matrix) [5pts].

Ans:

(a) Parameter matrix $\widetilde{\mathrm{W}}$ which is a 3 by 2 matrix.

$\tilde{x}_n = (1 \quad x_n^\mathrm{T})^\mathrm{T}$ is the augmented input vector for x_n

$y(x_n) = \widetilde{\mathrm{W}}^\mathrm{T}\tilde{x}_n$

$t_n = (1,0)^\mathrm{T}$ if $x_n \in$ C$_1$, and $t_n = (0,1)^\mathrm{T}$ if $x_n \in$ C$_2$.

The error function is

$\frac{1}{2}\sum_{n=1}^{N}\|y(x_n) - t_n\|^2$ [5pts]

(b) Let the data matrix be $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$, where the n-th row of $\mathbf{X}$ is $\tilde{x}_n^\mathrm{T}$. Let the

matrix $\mathbf{T}$ be a n by 2 matrix where the n-th row of $\mathbf{T}$ is t_n^T. Then

$$\widetilde{\mathrm{W}} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{T} = \begin{bmatrix} 1.2796 & -0.2796 \\ -0.0002 & 0.0002 \\ -0.2970 & 0.2970 \end{bmatrix}$$

Grading: if any numerical value of $\widetilde{\mathrm{W}}$ is wrong but the formula is correct, students will receive 3pts for (b). Students receive full points if they define $\tilde{x}_n = (x_n^\mathrm{T} \quad 1)^\mathrm{T}$ so that

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & 1 \\ x_{21} & x_{22} & 1 \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & 1 \end{bmatrix} \text{ and obtain } \widetilde{\mathrm{W}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{T} = \begin{bmatrix} \text{-0.0002} & 0.0002 \\ \text{-0.2970} & 0.2970 \\ 1.2796 & \text{-0.2796} \end{bmatrix}.$$

(B) Compute the linear classifier based on the training data using Fisher's linear discriminant by $\mathbf{w}=\mathbf{S}_{\mathrm{W}}^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1)$ where $\mathbf{S}_{\mathrm{W}}$ is the within-class covariance matrix. You need to write out (a) the Fisher criterion [5pts], (b) the computed parameter $\mathbf{w}=(w_0,w_1)^{\mathrm{T}}$ [5 pts].

Ans:

(a) Projection vector is $\mathbf{w}$ which is a 2 by 1 column vector. $\mathbf{S}_{\mathrm{B}}$ is the between-class covariance matrix and $\mathbf{S}_{\mathrm{W}}$ is the within-class covariance matrix. The Fisher criterion is

$$J(\mathbf{w})=\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \quad \text{[5pts]}$$

(b) $\mathbf{w}=(w_0,w_1)^{\mathrm{T}}=(0.0001, 0.1940)^{\mathrm{T}}$ [5pts]

Grading: any scaled version of $\mathbf{w}$ is also considered correct, i.e. $\mathbf{w}=(\text{-0.0001}, \text{-0.1940})^{\mathrm{T}}$.

3. Continuous Bayes Classifier [20 pts: 5+5+5+5]

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) $p\,(y = 1) = 0.6$; (2) $p\,(x|y = 1) = 0.5$ for $0 \le x \le 2$ and $p\,(x|y = 1) = 0$ otherwise; and (3) $p\,(x|y = 2) = 0.125$ for $0 \le x \le 8$ and $p\,(x|y = 2) = 0$ otherwise.

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

Ans: $\boldsymbol{p(y = 2)} = 1 - \boldsymbol{p(y = 1)} = 1 - 0.6 = 0.4$ [5pts]

(B) What is $p\,(y = 1|x)$ for $0 \le x \le 8$?

Ans: for $0 \le x \le 2$, we have

$$\begin{aligned} p(y = 1|x) &= \frac{p(x|y = 1) \times p(y = 1)}{p(x)} \\ &= \frac{p(x|y = 1) \times p(y = 1)}{p(x|y = 1) \times p(y = 1) + p(x|y = 2) \times p(y = 2)} \\ &= \frac{0.5 \times 0.6}{0.5 \times 0.6 + 0.125 \times 0.4} = \frac{6}{7} \end{aligned}$$

For $2 < x \leq 8$, we have $p(x|y = 1) = 0$, so $p(y = 1|x) = 0$.

Therefore, we have

$$p(y = 1|x) = \begin{cases} \frac{6}{7} & 0 \leq x \leq 2 \\ 0 & 2 < x \leq 8 \end{cases}$$

Grading: 3pts for correct $p(y = 1|x)$ for $0 \leq x \leq 2$, 2pts for correct $p(y = 1|x)$ for $2 < x \leq 8$.

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

Ans:

$p(y = 1|x = 1) = \frac{6}{7}$, $p(y = 2|x = 1) = \frac{1}{7}$.

Because $p(y = 1|x = 1) > p(y = 2|x = 1)$, the Bayes classifier will assign the label $y = 1$ for $x = 1$ [3pts]. The risk of this decision is $p(y = 2|x = 1) = \frac{1}{7}$. [2pts]

(D) What are the decision regions of your Bayes classifier?

Ans: The decision regions of the Bayes classifier is

$$\begin{cases} y = 1 & 0 \leq x \leq 2 \\ y = 2 & 2 < x \leq 8 \\ y=1 \text{ or } 2 & \text{otherwise} \end{cases}$$

Grading: 2pts for $y = 1$ $(0 \leq x \leq 2)$, 2pts for $y = 2$ $(2 < x \leq 8)$, 1pt for y=1 or 2 (otherwise). Equivalent results receive corresponding points, e.g. 3pts for $y = 2$ $(x > 2)$.

4. **Discrete Bayes Classifier [20 pts: 5+5+5+5]**

We want to build a Bayes classifier for a binary classification task ($y = 1$ or $y = 2$) with input feature x of two binary features (x_1 and x_2). We know the following quantities: (1) $p(y = 1) = 0.6$; (2) $p(x_1 = 0, x_2 = 0|y = 1) = 0.3$, $p(x_1 = 0, x_2 = 1|y = 1) = 0.1$, $p(x_1 = 1, x_2 = 0|y = 1) = 0.4$, $p(x_1 = 1, x_2 = 1|y = 1) = 0.2$, and (3) $p(x_1 = 0, x_2 = 0|y = 2) = 0.4$, $p(x_1 = 0, x_2 = 1|y = 2) = 0.3$, $p(x_1 = 1, x_2 = 0|y = 2) = 0.2$, $p(x_1 = 1, x_2 = 1|y = 2) = 0.1$.

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

Ans: $\boldsymbol{p(y = 2) = 1 - p(y = 1) = 0.4}$ [5pts]

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

Ans: We need to calculate $p\ (y = 1|x)$ for each possible combination of features . For $x_1 = 0, x_2 = 0$:

$$p(y = 1|x_1 = 0, x_2 = 0) = \frac{p(0,0|y=1)p(y=1)}{p(0,0|y=1)p(y=1)+p(0,0|y=2)p(y=2)}$$
$$= \frac{0.3 \times 0.6}{0.3 \times 0.6 + 0.4 \times 0.4} = \frac{9}{17} \approx 0.53$$

For $x_1 = 0, x_2 = 1$:

$$p(y = 1|x_1 = 0, x_2 = 1) = \frac{p(0,1|y=1)p(y=1)}{p(0,1|y=1)p(y=1)+p(0,1|y=2)p(y=2)}$$
$$= \frac{0.1 \times 0.6}{0.1 \times 0.6 + 0.3 \times 0.4} = \frac{1}{3} \approx 0.33$$

For $x_1 = 1, x_2 = 0$:

$$p(y = 1|x_1 = 1, x_2 = 0) = \frac{p(1,0|y=1)p(y=1)}{p(1,0|y=1)p(y=1)+p(1,0|y=2)p(y=2)}$$
$$= \frac{0.4 \times 0.6}{0.4 \times 0.6 + 0.2 \times 0.4} = \frac{3}{4} = 0.75$$

For $x_1 = 1, x_2 = 1$:

$$p(y = 1|x_1 = 1, x_2 = 1) = \frac{p(1,1|y=1)p(y=1)}{p(1,1|y=1)p(y=1)+p(1,1|y=2)p(y=2)}$$
$$= \frac{0.2 \times 0.6}{0.2 \times 0.6 + 0.1 \times 0.4} = \frac{3}{4} = 0.75$$

Grading: 1pt for correct result for each combination of x_1 and x_2, 5pts if results of all the combinations are correct.

(C) For an example with $x_1 = 0$ and $x_2 = 1$, what is the class label your classifier will assign? Why? What is the risk of this decision?

Ans: $p(y = 1|x_1 = 0, x_2 = 1) = \frac{1}{3}$, $p(y = 2|x1 = 0, x2 = 1) = \frac{2}{3}$

Because $p(y = 2|x1 = 0, x2 = 1) > p(y = 1|x_1 = 0, x_2 = 1)$, the Bayes classifier will assign the label $y = 2$. [3pts] The risk of this decision is

$$p(y = 1|x_1 = 0, x_2 = 1) = \frac{1}{3}.\ \text{[2pts]}$$

(D) What are the decision regions of your Bayes classifier?

Ans: The decision regions of the Bayes classifier is

$$\begin{cases} y = 1 & \begin{array}{l} x_1 = 0, x_2 = 0 \\ x_1 = 1, x_2 = 0 \\ x_1 = 1, x_2 = 1 \end{array} \\ \\ y = 2 & x_1 = 0, x_2 = 1 \end{cases}$$

Grading: 1pt for correct y value for each combination of x_1 and x_2, 5pts if results of all the combinations are correct.

5. Naive Bayes Classifier [20 pts:5+10+5]

Given the training data set in Table 2, 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.

Input Feature $X = (x_1, x_2, x_3, x_4, x_5)$					Class Label y
Sky	**Temp**	**Humid**	**Wind**	**Water**	**Enjoy Sport**
sunny	warm	normal	strong	warm	Yes
rainy	cold	high	mild	warm	No
sunny	warm	high	mild	warm	Yes
rainy	cold	high	strong	warm	No
sunny	warm	high	strong	cool	Yes
sunny	cold	normal	mild	warm	Yes
rainy	cold	normal	mild	cool	No

Table 2: Training Data Set for Naive Bayes Classifier

(A) How many independent parameters are there in your Naive Bayes classifier? What are they (only list the independent parameters)? Justify your answer.

Ans:

For Prior, the number of independent parameters is 1, and the parameter is $p(y = \text{Yes})$.

For Likelihood, the number of independent parameters is

5 (the number of features) × 2 (the number of classes) = 10. The parameters are $p(x_1 = \text{sunny}|y = \text{Yes}), p(x_1 = \text{sunny}|y = \text{No}), p(x_2 = \text{warm}|y = \text{Yes}), p(x_2 = \text{warm}|y = \text{No}), p(x_3 = \text{normal}|y = \text{Yes}), p(x_3 = \text{normal}|y = \text{No}), p(x_4 = \text{strong}|y = \text{Yes}), p(x_4 = \text{strong}|y = \text{No}), p(x_5 = \text{warm}|y = \text{Yes}), p(x_5 = \text{warm}|y = \text{No})$. [4pts for correctly listing all the 11 parameters, otherwise, 0.5pt for each correct parameter]

The total number of independent parameters is 1+10=11. [1pt]

(B) What are your estimations for these parameters?

Ans: we use $\boldsymbol{y} = \mathbf{1}$ to denote $\boldsymbol{y}$ = Yes, and $\boldsymbol{y} = \mathbf{0}$ to denote $\boldsymbol{y}$ = No.

(1) $p(y = 1) = \frac{4}{7}$.

(2) $p(x_1 = \text{sunny}|y = 1) = \frac{4}{4} = 1$; $p(x_1 = \text{sunny}|y = 0) = \frac{0}{3} = 0$.

(3) $p(x_2 = \text{warm}|y = 1) = \frac{3}{4}$; $p(x_2 = \text{warm}|y = 0) = \frac{0}{3} = 0$.

(4) $p(x_3 = \text{normal}|y = 1) = \frac{2}{4} = \frac{1}{2}$; $p(x_3 = \text{normal}|y = 0) = \frac{1}{3}$.

(5) $p(x_4 = \text{strong}|y = 1) = \frac{2}{4} = \frac{1}{2}$; $p(x_4 = \text{strong}|y = 0) = \frac{1}{3}$.

(6) $p(x_5 = \text{warm}|y = 1) = \frac{3}{4}$; $p(x_5 = \text{warm}|y = 0) = \frac{2}{3}$.

Grading: 1pt for each correct parameter, and 10pts for all correct parameters.

(C) Suppose we have a new input vector X = (sunny, cold, high, strong, cool). What is $p\,(y = 1|X)$? Which class label will the Naive Bayes classifier assign to this example? Justify your answer.

Ans:

$$p(y = 0|X) = \frac{p(X|y = 0) \times p(y = 0)}{p(X|y = 1) \times p(y = 1) + p(X|y = 0) \times p(y = 0)}$$

Since $p(x_1 = \text{sunny}|y = 0) = 0$, we have $p(X|y = 0) = 0$. Therefore, we have

$p(y = 0|X) = 0$ and $p(y = 1|X) = 1$. [3pts]

Since $p(y = 1|X) > p(y = 0|X)$, Naive Bayes classifier will assign the input vector X = (sunny, cold, high, strong, cool) to class 1 (Yes). [2pts]