CSE 575: Homework #1

Due: September 27, 2022

Problem 1

a) Suppose that X and Y are **independent** events, and $P(Y) > 0$. What is the value of $P(X|Y)$? (5pt)

Solution: Since the events X and Y are independent, $P(X \cap Y) = P(X) \times P(Y)$. Hence

$$\begin{aligned} P(X|Y) &= \frac{P(X \cap Y)}{P(Y)} \\ &= \frac{P(X) \times P(Y)}{P(Y)} \\ &= P(X) \end{aligned}$$

b) Suppose that X and Y are **disjoint** events, and $P(Y) > 0$. What is the value of $P(X|Y)$? (5pt)

Solution: Since the events X and Y are disjoint, $P(X \cap Y) = 0$. Hence

$$P(X|Y) = \frac{P(X \cap Y)P(X)}{P(Y)} = 0$$

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? (5pt)

$$HT, HT, TT, TT$$

Solution: For coin C_1, $P(C_1 = H) = 0.6$, $P(C_1 = T) = 0.4$. For coin C_2, $P(C_2 = H) = 0.4$, $P(C_2 = T) = 0.6$. Hence

$$\begin{aligned} P(HT) &= P(C_1 = H) \times P(C_2 = T) = 0.6 \times 0.6 = 0.36. \\ P(TT) &= P(C_1 = T) \times P(C_2 = T) = 0.4 \times 0.6 = 0.24. \end{aligned}$$

Therefore, probability of the given result is

$$P(HT) \times P(HT) \times P(TT) \times P(TT) = 0.36 \times 0.36 \times 0.24 \times 0.24 = 0.00746496.$$

d) Suppose you are given a coin and are asked to toss as many times as you wish to decide the probability of getting heads from the coin toss. 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? (5pt)

Solution: The likelihood is

$$P(X|\theta) = \theta^{15} \times (1-\theta)^5.$$

The MLE estimation of the coin toss is

$$\theta_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{15}{15+5} = 0.75.$$

Problem 2

You are given a training data set $\{x_n, t_n\}$ of size $N = 4$. Each input vector x_n is a point in the 2-dimensional Euclidean space R^2. We have $x_1 = (1, 0)$, $x_2 = (2, 1)$, $x_3 = (2, 3)$, $x_4 = (3, 3)$.

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 equations regarding least-squares linear discriminant and Fisher's linear discriminant, and have $t_n = -1$ in the question on the perception algorithm.

Hint: you may use software like python, Matlab or R for matrix computation, especially inverses.

a) Compute the least-square linear classifier based on the training data. You need to write out (a) the error function, (b) the computed parameters (w_0, w_1, w_2). (6pt)

Solution: (a) We have:

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 3 \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

(b)

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} = \begin{bmatrix} 0.944 \\ 0.167 \\ -0.444 \end{bmatrix}$$

b) Compute the linear classifier based on the training data using Fisher's linear discriminant. You need to write out (a) the error function, (b) the computed parameters (w_1, w_2) given $||w|| = 1$, and (c) choose a threshold for $w^T x$ such that all the training examples are correctly classified. (6pt)

Problem 2 continued on next page...

Solution:
(a)

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

We wish to maximize Fisher criterion (equivalently, minimize its negation):

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

(b)

$$\mathbf{m}_1 = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}, \quad \mathbf{m}_2 = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

$$\mathbf{S}_w = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1) = \begin{bmatrix} -3 \\ 8 \end{bmatrix}$$

Note that $\|(-3, 8)^T\| = \sqrt{73}$. Normalizing such that $\|\mathbf{w}\| = 1$ we get:

$$\mathbf{w} = \begin{bmatrix} -0.351 \\ 0.936 \end{bmatrix}$$

(c) $\mathbf{w}^T\mathbf{x}_1 = -0.35$, $\mathbf{w}^T\mathbf{x}_2 = 0.23$, $\mathbf{w}^T\mathbf{x}_3 = 2.11$, $\mathbf{w}^T\mathbf{x}_4 = 1.76$, therefore any threshold between 0.23 and 1.76 is acceptable.

c) Compute the linear classifier based on the training data using the perceptron algorithm, starting with the initial parameter $(w_0, w_1, w_2) = (1.5, 0, 0)$. For each iteration, you need to specify (a) the iteration number, (b) the current parameters, and (c) the updating vector. (8pt)

Solution: The perceptron rule is:

$$\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} + \mathbf{x}_n t_n$$

An example run of the perceptron algorithm may be:

Iteration	**Current w**	**Misclassified x**	**Updating vector w′**
1	(1.5, 0, 0)	(2,3)	(0.5, -2, -3)
2	(0.5, -2, -3)	(1,0)	(1.5, -1, -3)
3	(1.5, -1, -3)	(2,1)	(2.5, 1, -2)

Note that this is not a unique solution: the algorithm may converge to a different set of parameters depending on which misclassified example was used in each iteration.

Problem 3

We want to build a Bayes classifier for a binary classification task ($y = 1$ or $y = 2$) with a 1-dimensional real-valued input feature (x). We know the following quantities: (1) $p(y = 1) = 0.4$; (2) $p(x|y = 1) = 0.5$

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

a) What is the prior for class label $y = 2$? (5pt)

Solution: $p(y=2) = 1 - p(y=1) = 1 - 0.4 = 0.6$

b) What is $p(y=1|x)$? (5pt)

Solution: For $0 \leq x \leq 2$, we have

$$\begin{aligned} P(y=1|x) &= \frac{p(x|y=1) \times p(y=1)}{P(x)} \\ &= \frac{p(x|y=1) \times p(y=1)}{p(x|y=1) \times p(y=1) + p(x|y=2) \times p(y=2)} \\ &= \frac{0.5 \times 0.4}{0.5 \times 0.4 + 0.125 \times 0.6} \\ &= \frac{0.2}{0.2 + 0.075} \\ &= \frac{8}{11} \end{aligned}$$

For $2 < x \leq 8$, we have

$$P(x|y=1) = 0.$$

Hence $p(y=1|x) = 0$ for $2 < x \leq 8$. Therefore, we have

$$P(y=1|x) = \begin{cases} \frac{8}{11} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{if } 2 < x \leq 8 \\ \text{undefined} & \text{otherwise} \end{cases}$$

c) For $x = 1$, what is the class label your classifier will assign? Why? What is the risk of this decision (the probability of misclassifying)? (5pt)

Solution:

$$p(y=1|x=1) = \frac{8}{11}$$
$$p(y=2|x=1) = \frac{3}{11}$$

Since

$$p(y=1|x=1) > p(y=2|x=1)$$

the decision is to give x the label $y = 1$. The corresponding risk is

$$p(y=2|x=1) = \frac{3}{11}$$

d) What is the decision function of your Bayes classifier? (describe the classification made for different values of x) (5pt)

Solution: The decision function for the classifier is:

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

Problem 4

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

a) What is the prior for class label $y = 2$? (5pt)

Solution: $P(y = 2) = 1 - P(y = 1) = 0.4$

b) What is $P(y = 1|x)$? (5pt)

Solution: We need to calculate $P(y = 1|x)$ for each possible combination of x.
We use Bayes' Theorem for all possible values of x_1, x_2.

$$P(y = 1 \mid x_1, x_2) = \frac{P(x_1, x_2 \mid y = 1)P(y = 1)}{P(x_1, x_2 \mid y = 1)P(y = 1) + P(x_1, x_2 \mid y = 2)P(y = 2)}$$

Thus,

$$P(y = 1|x_1 = 0, x_2 = 0) = \frac{0.4 \times 0.6}{0.4 \times 0.6 + 0.3 \times 0.4} = \frac{2}{3} = 0.67$$
$$P(y = 1|x_1 = 0, x_2 = 1) = \frac{0.3 \times 0.6}{0.3 \times 0.6 + 0.1 \times 0.4} = \frac{9}{11} = 0.82$$
$$P(y = 1|x_1 = 1, x_2 = 0) = \frac{0.2 \times 0.6}{0.2 \times 0.6 + 0.4 \times 0.4} = \frac{3}{7} = 0.43$$
$$P(y = 1|x_1 = 1, x_2 = 1) = \frac{0.1 \times 0.6}{0.1 \times 0.6 + 0.2 \times 0.4} = \frac{3}{7} = 0.43$$

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? (5pt)

Solution:

$$P(y = 1|x_1 = 0, x_2 = 1) = \frac{9}{11} = 0.82, \quad P(y = 2|x_1 = 0, x_2 = 1) = \frac{2}{11} = 0.18$$

Since $P(y = 1|x_1 = 0, x_2 = 1) > P(y = 2|x_1 = 0, x_2 = 1)$, the classifier will label $x_1 = 0, x_2 = 1$ as 1.

$$\text{Risk} = P(y = 2|x_1 = 0, x_2 = 1) = \frac{2}{11} = 0.18$$

Problem 4 continued on next page...

d) What is the decision function of your Bayes classifier? (5pt)

Solution: The decision function for the classifier is:

$$y = \begin{cases} 1 & \text{if } x_1 = 0, x_2 = 0 \text{ or } x_1 = 0, x_2 = 1 \\ 2 & \text{if } x_1 = 1, x_2 = 0 \text{ or } x_1 = 1, x_2 = 1 \end{cases}$$

Problem 5

Given the training data below, we want to train a binary classifier using Naive Bayes. The last column is the class label y, and each column of **x** denotes a categorical feature. Features **Sky**, **Humid**, and **Wind** have two possible values; feature **Temp** can take three values (cold, mild, hot).

input features $x = (x_1, x_2, x_3, x_4)$				Class label y
Sky	**Temp**	**Humid**	**Wind**	**Play Sport**
sunny	mild	normal	strong	yes
rainy	cold	high	mild	no
rainy	hot	normal	strong	no
sunny	hot	high	mild	no
sunny	cold	normal	mild	yes
sunny	mild	normal	strong	no
rainy	mild	high	strong	no
rainy	mild	normal	mild	yes
sunny	hot	normal	strong	yes
sunny	cold	high	strong	no

a) How many independent parameters are there in your Naive Bayes classifier? Justify your answer. (5pt)

Solution:
For the class prior, number of independent parameters = 1:

$$P(y = \text{yes})$$

For the class-conditional distribution, number of independent parameters = 10:

$P(x_1 = \text{sunny} \mid y = \text{yes}),\ P(x_1 = \text{sunny} \mid y = \text{no})$
$P(x_2 = \text{cold} \mid y = \text{yes}),\ P(x_2 = \text{mild} \mid y = \text{yes}),\ P(x_2 = \text{cold} \mid y = \text{no}),\ P(x_2 = \text{mild} \mid y = \text{no})$
$P(x_3 = \text{normal} \mid y = \text{yes}),\ P(x_3 = \text{normal} \mid y = \text{no})$
$P(x_4 = \text{mild} \mid y = \text{yes}),\ P(x_4 = \text{mild} \mid y = \text{no})$

Therefore, the total number of parameters is: 1 + 10=11

b) What are the maximum likelihood estimates for these parameters? (10pt)

Problem 5 continued on next page...

Solution:

$$P(y = \text{yes}) = \frac{4}{10} = \frac{2}{5}$$
$$P(x_1 = \text{sunny} \mid y = \text{yes}) = \frac{3}{4}, \quad P(x_1 = \text{sunny} \mid y = \text{no}) = \frac{3}{6} = \frac{1}{2}$$
$$P(x_2 = \text{cold} \mid y = \text{yes}) = \frac{1}{4}, \quad P(x_2 = \text{mild} \mid y = \text{yes}) = \frac{2}{4} = \frac{1}{2},$$
$$P(x_2 = \text{cold} \mid y = \text{no}) = \frac{2}{6} = \frac{1}{3}, \quad P(x_2 = \text{mild} \mid y = \text{no}) = \frac{2}{6} = \frac{1}{3}$$
$$P(x_3 = \text{normal} \mid y = \text{yes}) = \frac{4}{4} = 1, \quad P(x_3 = \text{normal} \mid y = \text{no}) = \frac{2}{6} = \frac{1}{3}$$
$$P(x_4 = \text{mild} \mid y = \text{yes}) = \frac{2}{4} = \frac{1}{2}, \quad P(x_4 = \text{mild} \mid y = \text{no}) = \frac{2}{6} = \frac{1}{3}$$

c) Suppose we have a new input vector $\mathbf{x} = (\text{sunny}, \text{cold}, \text{normal}, \text{strong})$. What is $P(y = 1|\mathbf{x})$? Which class label will the Naive Bayes classifier (with the parameters from part (b)) assign to this example? Justify your answer. (5pt)

Solution:

$$\begin{aligned} P(y = 1 \mid \mathbf{x}) &= \frac{P(\mathbf{x} \mid y = 1)P(y = 1)}{P(\mathbf{x} \mid y = 1)P(y = 1) + P(\mathbf{x} \mid y = 0)P(y = 0)} \\ &= \frac{\frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{1}{2} \times \frac{2}{5}}{\frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{3}{5}} \\ &= \frac{27}{43} = 0.63 \end{aligned}$$

Because $P(y = 1 \mid \mathbf{x}) \geq 0.5$, the naive Bayes classifier will assign $y = 1$ to this example.

d) (Bonus question) Suppose we have a new input vector $\mathbf{x} = (\text{sunny}, \text{cold}, \text{normal}, \textbf{missing})$. That is, you are given values for the first three features only. What is $P(y = 1|\mathbf{x})$ according to the Naive Bayes model from part (b)? Justify your answer. (5pt)

Problem 5 continued on next page...

Solution: Bayes' Theorem holds regardless of missing values in the input:

$$P(y=1 \mid x_1, x_2, x_3) = \frac{P(x_1, x_2, x_3 \mid y=1)P(y=1)}{P(x_1, x_2, x_3 \mid y=1)P(y=1) + P(x_1, x_2, x_3 \mid y=0)P(y=0)}$$

Using sum rule and the naive Bayes assumption:

$$\begin{aligned}
&P(x_1, x_2, x_3 \mid y=1) \\
&= P(x_1, x_2, x_3, x_4 = \text{mild} \mid y=1) + P(x_1, x_2, x_3, x_4 = \text{strong} \mid y=1) \\
&= P(x_1 \mid y=1) \times P(x_2 \mid y=1) \times P(x_3 \mid y=1) \times P(x_4 = \text{mild} \mid y=1) \\
&\quad + P(x_1 \mid y=1) \times P(x_2 \mid y=1) \times P(x_3 \mid y=1) \times P(x_4 = \text{strong} \mid y=1) \\
&= P(x_1 \mid y=1) \times P(x_2 \mid y=1) \times P(x_3 \mid y=1) \times [P(x_4 = \text{mild} \mid y=1) + P(x_4 = \text{strong} \mid y=1)] \\
&= P(x_1 \mid y=1) \times P(x_2 \mid y=1) \times P(x_3 \mid y=1)
\end{aligned}$$

Note that this is a direct implication of naive Bayes assumption: variables are independent given the class label.

Thus,

$$P(y=1 \mid x_1 = \text{sunny}, x_2 = \text{cold}, x_3 = \text{normal}) = \frac{\frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{2}{5}}{\frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{2}{5} + \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{3}{5}} = \frac{27}{39} = 0.69$$