ESE 402-542 : Statistics for Data Science
Instructor: Hamed Hassani
Fall 2020

Final Examination

NAME	

	Grade (y/n)	Score	Max. Score
Problem 1			30
Problem 2			40
Problem 3			30
TOTAL			100

Problem 1 (30 points)
Assume that $X_1, X_2, \cdots, X_n$ are generated i.i.d. according to the following distribution:

$$\Pr(X = i) = \theta(1-\theta)^i, \text{ for } i = 0, 1, 2, 3, \cdots$$

In other words, the pdf (pmf) of the distribution that generates the data is of the form $f(X = i|\theta) = \theta(1-\theta)^i$, where θ is an unknown parameter. Consider the following hypothesis testing problem:

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_\text{a} : &\quad \theta = \theta_a, \end{aligned}$$

where we assume that $\theta_a > \theta_0$.

Derive the most powerful test for this hypothesis testing problem and specify what the acceptance/rejection regions are for a given significance level α_0 (you can assume that n is large).

Problem 2 (40 points)

In this question we assume that data is generated according to a distribution $P(X = x, Y = y)$ given as follows: $x \in \mathbb{R}$ and $y \in \{-1, 1\}$, i.e. the data is one-dimensional and the label is binary. Write $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$. We let $P(y = +1) = \frac{3}{4}$, and $P(Y = -1) = \frac{1}{4}$, and

$$P(X = x|Y = +1) = \frac{1}{2}\exp(-|x - 2|), \text{ and,}$$

$$P(X = x|Y = -1) = \frac{1}{2}\exp(-|x + 2|).$$

I.e. $P(X = x|Y = +1)$ and $P(X = x|Y = -1)$ follow the Laplace distribution. (The mean of a Laplace distribution with pdf $\frac{1}{2b}e^{(-\frac{|x-\mu|}{b})}$ is μ and the variance is $2b^2$.)

1. Derive the expression for $P(X = x, Y = y)$.

2. Plot the conditional distributions $P(X = x, Y = +1)$ and $P(X = x, Y = -1)$ in one figure.

3. Write the Bayes optimal classification rule given the above distribution P and simplify it (hint: in the end you should reach to a very simple classification rule that classifies an input x based on whether or not its value is greater than a threshold).

4. Compute the probability of classification error for the Bayes optimal classifier.

5. Let us now consider QDA (this part and the next can be answered independently from the previous parts). Given training data $(x_1, y_1), \cdots, (x_n, y_n)$, explain briefly the main steps of training the QDA model. I.e. what quantities/probabilities are being estimated by QDA? What is the parametric model used? How are the parameters of the model found? (Recall that QDA is similar to LDA with the exception that the variance per class can be different.)

6. Assume that the number of training data points is very large (i.e. $n \to \infty$); What will be the exact value of the parameters of the trained QDA model in this case? (Hint: You don't really need much calculation to derive the parameters.)

7. Simplify the QDA classifier that you obtained in the previous part (hint: in the end you should reach to a very simple classification rule that classifies an input x based on whether or not its value is greater than a threshold).

8. Given your answers in Parts 2 and 5, what do you think about the performance of QDA compared to what can be done optimally? Does QDA perform optimally when we have many training data points?

Problem 3 (30 points)

1. Find the VC-dimension of the function class $\mathcal{H}_1 = \{h_a; a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0, 1\}$ is defined as

$$h_a(x) = \begin{cases} 1, & \text{if } x \in [a, a+1] \\ 1, & \text{if } x \in [a+5, a+6] \\ 0 & \text{otherwise.} \end{cases}$$

2. Find the VC-dimension of the function class $\mathcal{H}_2 = \{h_{a,b}; a, b \in [0, +\infty)\}$, where $h_{a,b} : \mathbb{R} \to \{0, 1\}$ is defined as

$$h_{a,b}(x) = \begin{cases} 1, & \text{if } |x| \leq a \\ 1 & \text{if } |x| \geq b \\ 0 & \text{otherwise.} \end{cases}$$

3. Consider the class $\mathcal{H}_1$. Given a training data set S, let h_S^* be the outcome of the empirical risk minimization procedure restricted to class $\mathcal{H}_1$. How many training samples do we need (ignoring constants) to ensure that for any distribution of the data, $\mathcal{D}$, the following event occurs with probability $1 - \delta$?

$$L_{\mathcal{D}}(h_S^*) - \min_{h \in \mathcal{H}_1} L_{\mathcal{D}}(h) \leq \epsilon.$$

(Justify your answer in one sentence.)

4. Consider again the same class $\mathcal{H}_1$. Recall that we defined h_S^* to be the outcome of the empirical risk minimization procedure restricted to class $\mathcal{H}_1$. How many training samples do we need (ignoring constants) to ensure that for any distribution of the data, $\mathcal{D}$, the following event occurs with probability $1 - \delta$?

$$L_{\mathcal{D}}(h_S^*) - \min_{h \in \{\text{all the possible functions}\}} L_{\mathcal{D}}(h) \leq \epsilon.$$

(Justify your answer in one sentence.)