ESE 402/542: Statistics for Data Science
Instructor: Hamed Hassani
Fall 2021

Final Examination

NAME	

Additional Information:

- You may use any class material (homeworks, class notes, listed textbooks) on the exam, but nothing else.
- There should be very little tedious computation if you approach the problem correctly.
- If you are stuck in any part of a problem do not dwell on it, try to move on and attempt it later.
- There are 10 bonus points – i.e. the total number of points is 70 but the exam will be graded out of 60. If your final grade is above 60, the extra points will be added to the midterm after proper normalization.

	Score	Max. Score
Problem 1		20
Problem 2		20
Problem 3		10
Problem 4		20
TOTAL		60

Problem 1. Classification [20 pts]

Consider the following two-class data distribution:

$$\begin{aligned}
&\mathbb{P}[Y=0]=p, \quad \mathbb{P}[Y=1]=1-p \\
&\mathbb{P}[X=x|Y=0]=1/5, \quad \text{for } -4 \leq x \leq 1 \\
&\mathbb{P}[X=x|Y=1]=1/4, \quad \text{for } 0 \leq x \leq 4,
\end{aligned}$$

where $0 < p < 1$ is a fixed constant.

a) (2 pts) Plot the conditional distributions $\mathbb{P}[X=x|Y=0]$ and $\mathbb{P}[X=x|Y=1]$ in a single plot.

b) (3 pts) *An exercise in conditional probability.* Write $\mathbb{P}[Y=y|X=x]$ in terms of $\mathbb{P}[X=x|Y=y]$, $\mathbb{P}[Y=y]$, and $\mathbb{P}[X=x]$.

c) (10 pts) Recall the Bayes optimal classifier is defined $h^*(x) := \text{argmax}_{y\in\{0,1\}} \mathbb{P}[Y=y|X=x]$. Provide the Bayes optimal classifier for the given data distribution. Explain why the classifier you derived is optimal, intuitively.

Hint 1: part b) should come in handy. Hint 2: your answer should depend on p.

d) (5 pts) Derive the Bayes error rate of this data distribution, that is, the error rate of the Bayes optimal classifier $\mathbb{P}_{X,Y}[h^*(X) \neq Y]$?

Problem 2. k**-means Clustering** [20 pts]

a) (5 pts) Plot the following 16 points on the 2-D plane:

$$\{(-5, \pm 1), (-3, \pm 1), (\pm 1, 5), (\pm 1, 3), (5, \pm 1), (3, \pm 1), (\pm 1, -5), (\pm 1, -3)\}.$$

Make the plot large, as you will be using it for the next few parts. We want to use k-means to cluster these points into 4 clusters.

Plot the following initial cluster centers $C_0 = \{(-5, 5), (-5, -5), (5, -5), (5, 5)\}$.

b) (10 pts) Run k-means for one full iteration: First, report which points are closest to which cluster center in C_0. Then, update the clusters centers by finding the centroids of each cluster–you do not need to compute this exactly. Mark the approximate location of the new cluster centers on your plot.

c) (3 pts) What is the *final* output of the k-means algorithm after it terminates in this instance? Justify.

d) (2 pts) Determine the *optimal* cluster centers and mark them on your plot. Did k-means successfully find the optimal clusters in this instance?

Problem 3. Variations of Linear Regression [10 pts]
Suppose we have data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ are fixed, and the random variables $y_i \in \mathbb{R}$ are generated under the following model:

$$y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \epsilon_i$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$. Different variants of linear regression can be recovered depending on different assumptions made about ϵ_i and $\boldsymbol{\beta}$.

a) (5 pts) Suppose $\boldsymbol{\beta}$ is some fixed d-dimensional vector, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, 1)$, i.e. the PDF of ϵ_i is given by $f_{\epsilon_i}(z) = \frac{1}{2}e^{-|z|}$. Show that the likelihood $f(y_1, \ldots, y_n|\boldsymbol{\beta})$ is given by

$$f(y_1, \ldots, y_n|\boldsymbol{\beta}) = \left(\frac{1}{2}\right)^n \prod_{i=1}^n e^{-|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|}$$

Hint: We have done several similar exercises in class–For example, you should use the fact that ϵ_i's are i.i.d. to obtain the product.

b) (5 pts) Under the same setup as part a), show that the maximum likelihood estimator ($\hat{\boldsymbol{\beta}}^{\text{MLE}} \triangleq \text{argmax}_{\boldsymbol{\beta}} f(y_1, \ldots, y_n|\boldsymbol{\beta})$) of $\boldsymbol{\beta}$ is equivalent to the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{\text{MLE}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^n |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|$$

Hint: Recall that in maximum likelihood estimation we always took the 'log' of the likelihood function

Problem 4. VC Dimension and Shattering Number [20 pts]
Let $\mathcal{H}$ be a function class containing unions of 3 disjoint intervals of $\mathbb{R}$, i.e.

$$\mathcal{H} \triangleq \{h_{a_1,a_2,a_3,b_1,b_2,b_3} : a_i < b_i < a_{i+1} \forall i\}$$

where

$$h_{a_1,a_2,a_3,b_1,b_2,b_3}(x) = \begin{cases} 1 & \text{if} \quad x \in [a_1, b_1] \cup [a_2, b_2] \cup [a_3, b_3] \\ 0 & \text{if} \quad x \notin [a_1, b_1] \cup [a_2, b_2] \cup [a_3, b_3] \end{cases}$$

You will show why the VC dimension of $\mathcal{H}$ is 6.

a) (10 pts) Show that the set $\{1, 2, 3, 4, 5, 6\}$ is shattered by $\mathcal{H}$. *Hint: You don't need to explicitly list all the possible binary* 6*-tuples. Instead, take a binary* 6*-tuple of the form* $(c_1, c_2, \cdots, c_6)$*, where* $c_i \in \{0, 1\}$*, and provide a function* $h \in \mathcal{H}$ *such that* $h(i) = c_i$ *for* $i = 1, \cdots, 6$.

b) (10 pts) Show that there in no set of 7 points in $\mathbb{R}$ that can be shattered by $\mathcal{H}$. *Hint: We solved a similar exercise in class.*