COM S 474/574 Spring 2020
Introduction to Machine Learning

Final Exam

1 Directions:

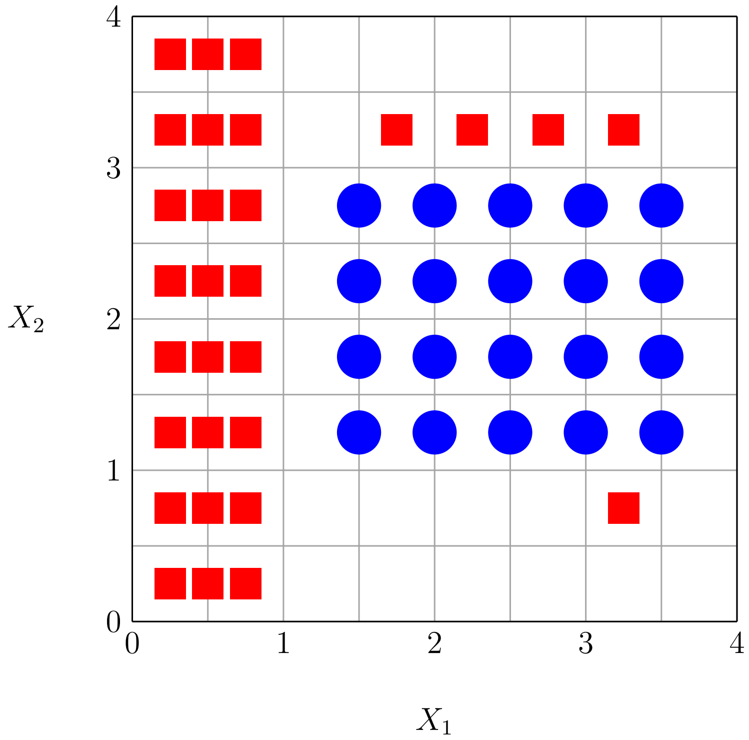
- **Due: Friday May 8, 2020 at noon.** Late submissions will be penalized 0.5 points per minute it is late. You are strongly encouraged to complete the exam and upload your answers on Thursday May 7.
- Upload your solutions to Canvas as a pdf file. Your solutions may be hand-written or typed. Plots can be hand-drawn.

  The exam can be done entirely by hand. But if you do any calculations on the computer or write any scripts to make plots as part of doing this exam, also upload a file or zipped folder with that information (such as a .ipynb if everything is in one notebook, .zip if there were multiple files; contact us if you have any questions regarding this)
- The exam is designed to fit within a two-hour block. However, you are not required to finish it within a two-hour window.
- Unless specified otherwise, show all your work.
- You are allowed to use
  - a computer and/or calculator.
  - the textbook and Chris Bishop's machine learning textbook
  - our lecture notes (video recordings and pdfs)
  - our homeworks and solutions
  - Piazza/email to ask clarifying questions about the exam

  You are **not** allowed to
  - communicate with other students
  - access websites other than canvas, piazza, or email (for communicating with the staff)
  - do web searches
  - do anything else not explicitly permitted above or approved through piazza/email by the staff.

Problem 1. [16 points]

Consider the following scatter plot of training samples. There are two classes, Y = red and Y = blue. We will use a single decision tree to achieve perfect classification.



A. In 2-5 sentences, briefly explain how a decision tree is constructed.

B. Draw the decision tree diagram, labeling edges and the predictions. You do ***not*** need to explicitly calculate Gini values for this problem. You should be able to use your knowledge of what Gini values measure to determine the splits.

C. Sketch the corresponding partition of the X_1, X_2 plane. Label each region with the corresponding prediction. You do *not* need to redraw the training samples themselves. Just the decision boundaries corresponding to a decision tree.

Problem 2. [20 points]

Suppose we have a training data set of 100 samples with a feature Y that we want to predict on new data. We have five features X_1, X_2, ..., X_5 that we can use to predict Y. We will use (multiple) linear regression.

The mean value of Y across training samples is

$$\frac{1}{100}\sum_{i=1}^{100} Y(i) = 3.8.$$

We are concerned that this model may under-fit.

Using all of the features, we find that

$$\underset{a_0,a_1,\ldots,a_5}{\arg\min} \quad \frac{1}{100}\sum_{i=1}^{100}\Big(Y(i) - (a_0 + a_1X_1(i) + \cdots + a_5X_5(i))\Big)^2$$

is

$$\widehat{Y} = 2.1 - 0.4X_1 + 1.3X_2 + 3.8X_3 - 0.9X_4 + 0.5X_5.$$

We are concerned that this model may over-fit.

Suppose we decide to include a penalty on complexity of the form

$$\underset{a_0,a_1,\ldots,a_5}{\arg\min} \quad \frac{1}{100}\sum_{i=1}^{100}\Big(Y(i) - (a_0 + a_1X_1(i) + \cdots + a_5X_5(i))\Big)^2 + \lambda \sum_{j=1}^{5}|a_j|^p. \tag{1}$$

A. For any exponent $p \geq 0$, what solution do we get when

   1. $\lambda = 0$
   2. $\lambda \to \infty$

B. Consider the case with $p = 0$. Unfortunately, we cannot solve (1) with a single optimization call.

   1. In a few words, what is the penalty term measuring?
   2. If we wanted to solve (1), how many optimization calls would be needed? In a few words, explain the procedure.
   3. In 3-6 sentences, describe two heuristic procedures we discussed in class that run much faster than the previous method, though may not find the best solution to (1).

C. If we use $p = 1$ or $p = 2$ instead, for a fixed λ, how many optimization calls are necessary to solve (1)?

D. In 1-3 sentences, describe what the major difference between the types of solutions (eg sets of coefficients) we get with $p = 1$ and $p = 2$ are as we vary λ from 0 to ∞.

Problem 3. [20 points]

In this problem, we consider k nearest neighbor regression. Suppose we have the following (training) data set.

X	Y
2	10
4	2
6	0
8	4

For each $k \in \{1, 2, 3, 4\}$,

- Make a scatter plot with X values on horizontal axis, Y values on the vertical axis.
- Plot the (unweighted) k-nearest neighbor prediction function over the range $X \in [-2, 12]$.
- Show the calculations for determining the prediction function over that range and explain the process.

Problem 4. [12 points]

In 4-6 sentences, describe

- what the terms “under-fitting” and “over-fitting” refer to,
- why both phenomena are undesirable, and
- how we can identify when they happen.

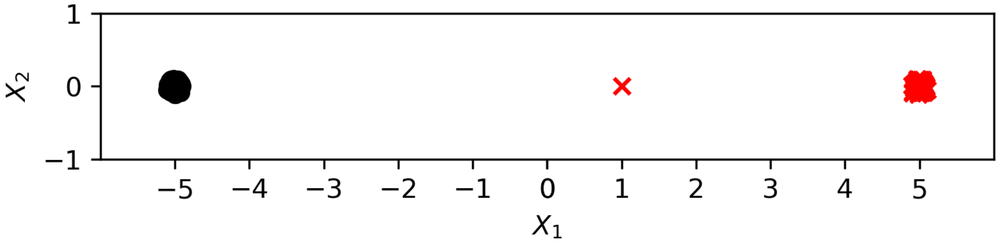
You may use drawings to help illustrate your point.

Problem 5. [12 points]

In 3-5 sentences, describe the goal of PCA, including how it differs from linear regression.

Problem 6. [20 points]

Below is a scatter plot of a training data set.



- Class $Y = -1$: there are 1000 samples, plotted as black circles. They follow a multivariate normal distribution
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.001 & 0 \\ 0 & 0.001 \end{bmatrix} \right)$$
For the purposes of this problem you can treat all of the 1000 samples in class -1 as having $X_1 = -5$ and $X_2 = 0$.

- Class $Y = +1$: there are 1000 samples, plotted as red 'x's. They follow a multivariate normal distribution
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.001 & 0 \\ 0 & 0.001 \end{bmatrix} \right)$$
For the purposes of this problem you can treat 999 samples in class $+1$ as having $X_1 = 5$ and $X_2 = 0$ and there is one (outlier) sample at $X_1 = 1$ and $X_2 = 0$.

Solve the following questions by inspection (e.g. no need to do calculations or solve algebraic formulas; you can round values to the nearest integer).

A. If we fit an LDA classifier on this data, where would the boundary be? Briefly explain why.

B. If we fit a QDA classifier on this data, where would the boundary be? Briefly explain why.

C. If we fit a max margin classifier (linear SVM with no mistakes allowed), where would the boundary be? Where would the margins be? Briefly explain why.

D. Consider fitting a linear SVM classifier where we have a budget C for mistakes,
$$Y(i)\,[b + w_1 X_1(i) + w_2 X_2(i)] \geq 1 - \xi_i \quad \text{for all } i \in \{1, 2, \ldots, 2000\}$$
$$\xi_i \geq 0 \qquad \text{for all } i \in \{1, 2, \ldots, 2000\}$$
$$C = \sum_{i=1}^{2000} \xi_i$$
Describe what happens to the boundary and the margins as we increase C beginning at 0.