CSE 142: Machine Learning, Fall 2021

Assignment #1
Due Monday, October 18 by 23:59pm PT

Notes:

- *This assignment is to be done individually. You may discuss the problems at a general level with others in the class (e.g., about the concepts underlying the question, or what lecture or reading material may be relevant), but the work you turn in must be solely your own.*
- Be sure to re-read the "Policy on Academic Integrity" on the course website.
- Be aware of the late policy in the course syllabus so turn in what you have by the due time.
- Justify every answer you give – **show the work** that achieves the answer or **explain** your response.
- Any updates or corrections will be posted on the Assignments page (of the course website), so check there occasionally.
- To turn in your assignment:
  - Submission through Gradescope.
  - Clearly indicates which part of your submission belongs to which question when submitting. If you don't do it right, grader might see "no content is available".

Problem #1 [5 points]
Consider the problem of an adult learning to speak and understand a foreign language. Explain how this process fits into the general learning model (Fig. 3 in the textbook) – i.e., describe the domain objects, training data, model, learning algorithm, and output for this scenario. Discuss what kind(s) of learning takes place.

Problem #2 [6 points]
Choose one of the active data science competitions at Kaggle (www.kaggle.com) and describe the core machine learning problem by answering these questions:

(a) What is the task? (Please include the Kaggle link to the task)
(b) What is the experience or data used to learn a model?
(c) What is (are) the primary performance measure(s)?
(d) What prior assumptions are reasonable to make?
(e) What kind(s) of learning will be done – supervised, semi-supervised, unsupervised, reinforcement?
(f) Briefly describe the data that is made available.

Problem #3 [6 points]
You are asked to build a machine learning system to estimate someone's blood pressure (two numbers: systolic and diastolic; consider them to be real-valued) based on the following inputs: the patient's sex, age, weight, average grams of fat consumed per day, number of servings of red meat per week, servings of fruits and vegetables per day, smoker or non-smoker. You are given a training data set of values for all of these variables and the blood pressure numbers for 10,000 patients.

Answer (and explain) the following questions:

(a) What kind of machine learning problem is this?
(b) Is it a predictive task or a descriptive task?
(c) Are you likely to use a geometric model, a probabilistic model, or a logical model?
(d) Will your model be a grouping model or a grading model?
(e) What is the label space for this problem?
(f) What is the output space for this problem?

Problem #4 [8 points]
We (simplistically) describe a basketball player's value in terms of the following statistics:

	LeBron James ($x1$)	Kevin Durant ($x2$)
Minutes played per game	38.2	36.7
Points scored per game	27.0	27.0
Rebounds per game	7.4	7.1
Assists per game	7.4	4.2
Steals per game	1.6	1.1
Blocks per game	0.7	1.1
Turnovers per game	3.5	3.1

Treating the statistics for each player ($x1$ and $x2$) as a feature vector, what is the distance between them, measured in terms of (a) L1 distance, (b) L2 distance, (c) L_{10} distance, (d) L_{100} distance?
(e) If a constant vector $v = [5\ 5\ 2\ 2\ 0.5\ 0.1\ 1]^T$ is added to both $x1$ and $x2$, which (if any) of L1, L2, L_{10}, or L_{100} will change?
(f) If $x1$ and $x2$ are multiplied by a constant k, which (if any) of L1, L2, or L_{10} will change?

Problem #5 [12 points]
The joint probability distribution of three variables, *class*, *grade* and *effort* can be computed from the following table that shows numbers of students in each bin:

	class = machine learning			**class** = surfing		
grade	**effort**=Small	Medium	Large	**effort**=Small	Medium	Large
A	0	25	75	50	125	125
B	5	50	50	50	50	25
C	25	50	5	50	25	0
D	50	25	0	0	0	0
F	50	0	0	0	0	0

(a) What is the conditional probability distribution P(*grade* | *class*, *effort*)?
(b) What is the marginal probability distribution P(*grade*, *effort*)?
(c) What is the marginal probability distribution P(*effort*)?
(d) What is P(*grade*=A | *class*)?

Problem #6 [12 points]
Here are some data with features {F1, F2, F3} used in a machine learning problem (six instances in a three-dimensional feature space):

#	F1	F2	F3	Label
1	2	5	6	A
2	5	8	9	B
3	1.4	4.4	5.4	A
4	4.7	7.7	8.7	B
5	1.7	4.7	5.7	B
6	4.4	7.4	8.4	B

(a) What is the intrinsic dimensionality of the data?
(b) After dimensionality reduction is applied to the data to produce a transformed feature space, and assuming that these examples are linearly separable into two classes, what is the geometric form of the discriminating function that separates the data? (E.g., a point, a line, a plane, a sphere, a non-linear contour, ….)
(c) Are the examples in this training set linearly separable in the intrinsic feature space?
(d) Are the examples in this training set linearly separable in the original feature space?

As always, don't just give the answer – explain how you arrived at the answer.

Problem #7 [9 points]
A ranking classifier ranks 25 training examples $\{x_i\}$, from highest to lowest rank, in the following order:

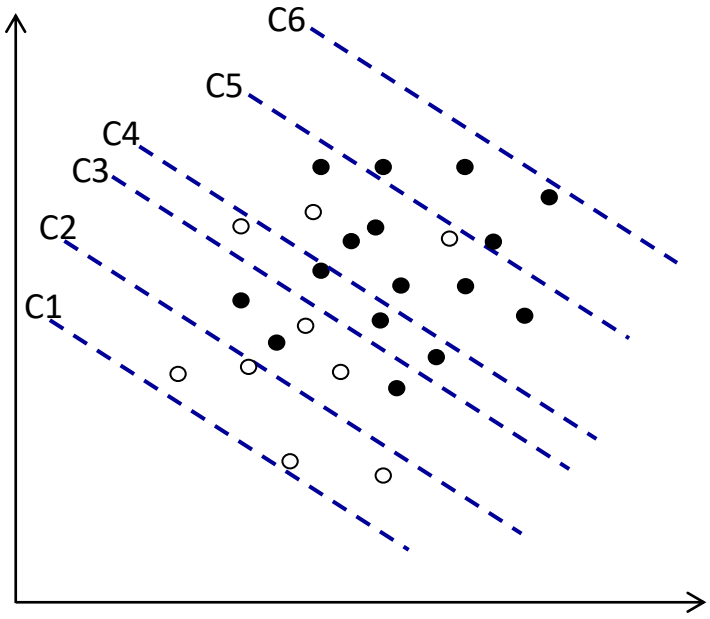
Highest Lowest

$x_9, x_{10}, x_{12}, x_{11}, x_7, x_{15}, x_4, x_2, x_3, x_{22}, x_{16}, x_{19}, x_{25}, x_{23}, x_{24}, x_1, x_5, x_{13}, x_6, x_8, x_{14}, x_{21}, x_{17}, x_{20}, x_{18}$

Examples x_1 through x_{12} are in the positive class (which should be ranked higher); examples x_{13} through x_{25} are in the negative class (which should be ranked lower).

(a) How many ranking errors are there?
(b) What is the ranking error rate?
(c) What is the ranking accuracy?

Problem #8 [14 points]



The figure below shows training data with two features, with each example labeled as being in the positive (filled-in points) or negative (open points) class. Proposed linear discriminant functions (C1 through C6) are shown as dotted lines, each one indicating a different classifier for this data. Each classifier classifies points to the upper-right of its dotted line as positive and points to the lower-left of its dotted line as negative.

(a) Draw the coverage plot for this data and plot the different classifiers (and label them as C1, C2, etc.).
(b) Draw the ROC plot and label the classifiers on the plot.

(c) Which classifiers have the highest and lowest accuracy?
(d) Which classifiers have the highest and lowest precision?
(e) Which classifiers have the highest and lowest recall?
(f) Which classifiers (if any) are complete?
(g) Which classifiers (if any) are consistent?