CS 188 Midterm

JUNHONG WANG

TOTAL POINTS

93.25 / 100

QUESTION 1

1 Problem 1 16 / 16

✓ - **0 pts** Correct

- **5 pts** Did not specify article

- **5 pts** Not clear what the article got wrong

- **8 pts** Invalid argument on article inaccuracies

- **8 pts** Did not propose alternate solution

- **16 pts** no attempt

QUESTION 2

kMeans Classification 28 pts

2.1 Centroids 5 / 5

✓ - **0 pts** Correct

- **3 pts** Incorrect Centroid value calculation

- **2 pts** Incorrect plotting of Centroid

- **5 pts** Did not Attempt

- **1 pts** Plotted correctly but incorrect values

- **2 pts** Click here to replace this description.

2.2 Vertex + Voronoi 5 / 5

✓ - **0 pts** Correct

- **3 pts** Incorrect Vertex Calculation

- **2 pts** Incorrect plotting of Vertex

- **2 pts** Incorrect drawing of Voronoi diagram

- **5 pts** Did not attempt solution

- **2 pts** Failed to calculate vertex value (even though plot appears correct)

2.3 Classification and Metrics Calculation 10 / 10

✓ - **0 pts** Correct

- **4 pts** Improperly calculated the predicted labels

- **4 pts** Improperly calculated metric

- **2 pts** Improperly calculated metric (error carried over from mislabeling

- **10 pts** Did not attempt

- **2 pts** missing or incorrect F1 score

2.4 Metrics Evaluation 8 / 8

✓ - **0 pts** Correct

- **2 pts** Failed to discuss each of the metrics

- **4 pts** Failed to differentiate the Recall performance against the rest of the metrics

- **8 pts** Did not attempt

QUESTION 3

3 Perceptron 28 / 28

✓ - **0 pts** Correct

- **18 pts** Concluded that it won't converge

- **5 pts** calculation mistake on step

- **10 pts** Did not iterate through dataset examples

- **28 pts** no attempt

QUESTION 4

Project 28 pts

4.1 Visualizations 7 / 7

- **0 pts** Correct: B + C

✓ - **0 pts** Did not select B, but project 2 misled you about the answer

- **1.75 pts** Did not select C

- **3.5 pts** Selected A

- **3.5 pts** Selected D

4.2 Filtering 6.5 / 7

- **0 pts** Correct

- **2 pts** Used cereal['shelf'...] instead of cereal[cereal['shelf'] ...]

- **1.5 pts** Returned boolean mask rather than the actual rows

- **0 pts** Used "and" instead of &

- **0 pts** Used "&&" instead of &

- **0 pts** Used "," instead of &

- **1.5 pts** Used "|" instead of &

- **0 pts** Used iloc instead of loc

- **1 pts** Used where instead of loc

- **3.5 pts** Used group instead of loc

- **0 pts** No parentheses to group operations, would result in logic error

- **0 pts** Passed boolean series into select instead of a function

- **1 pts** Unnecessary call to index (isn't callable)

✓ - **0.5 pts** **Used "df" or another name instead of "cereal"**

- **0 pts** No quotations around column name

- **0 pts** Used groupby, filter, where, or select instead of query

- **5 pts** Used unique instead of query

- **0 pts** Unnecessary quotations around column name

- **0 pts** Incorrect quotation in query

- **3.5 pts** Wrote two solutions, one of which was incorrect

✓ - **0 pts** **Used "=" instead of "=="**

- **1 pts** Used <= instead of <

- **0 pts** Unnecessary inclusion of colon operator

- **5.5 pts** Used drop instead of cereals[]

- **7 pts** Incorrect

- **5 pts** Half Correct

- **7 pts** Blank

4.3 Plotting 2.5 / 7

- **0 pts** Correct

- **4 pts** Called cereal.plot.barh without any aggregation

✓ - **2.5 pts** **Incorrect attempt at aggregation**

- **1.5 pts** Slight error in aggregation

Axes

✓ - **2 pts** **Didn't set axis labels**

- **1 pts** Incorrect attempt to set axis labels

- **0.5 pts** Labels flipped

- **0.5 pts** Axis labels set with incorrect names

- **1 pts** Passed extra argument to barh

- **0.5 pts** Called groupby or barh on pd or on its own instead of your dataframe, or on wrong dataframe

- **0.5 pts** Called cereal.plot(kind=hbar) instead of cereal.plot.barh()

- **0 pts** Forgot quotes around column name

✓ - **0 pts** **Called cereal.plot.barh() with pandas series/dataframe (or other python object) instead of strings**

- **0.5 pts** Missing argument to x.plot.barh()

- **0.5 pts** Order of arguments to barh flipped

- **0.5 pts** plt.plot() instead of plt.show()

- **2 pts** Called .plot.barh() on something other than a dataframe

- **1 pts** Called hist instead of barh()

- **7 pts** Incorrect

- **7 pts** Blank

4.4 Correlations 5.25 / 7

- **3.5 pts** A: It is possible that the rating is of the form $z = ax + by$, and that adding a second feature could explain all variation

- **3.5 pts** B: A negative correlation indicates that cereals with higher protein tend to have lower ratings

✓ - **1.75 pts** **C: A correlation coefficient of 0 means there is no linear association, but not that there is no association at all**

- **3.5 pts** D: It's possible that sugar and fat are inversely correlated with one another. However, while both tend to be higher when scores are higher, there is no guarantee both are high together.
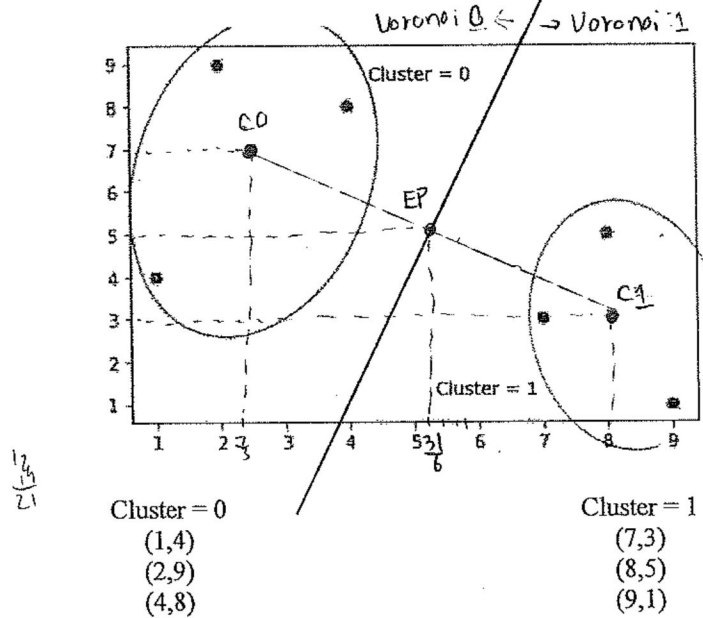
- **0 pts** E: Correct

- **7 pts** Blank

U C L A Computer Science Department

CS 188 **Data Science Fundamentals** UID: ___504941113___

Midterm **Total Time: 1 hour** **February 26, 2020**

Problem 1 has **16 points**. All other problems have **28 points**.
You **cannot** use the back of the pages for your answers (you are welcome to use them as a scratch paper)

Problem 1: a. Describe a news headline that you have read recently that seems odd/wrong. Include a citation.
b. How were you able to prove/observe the result was odd/wrong.
c. Describe, step-by-step, how would you go about finding the right information using data science.

a. This is an article I found when I was doing CS 35L project. I think the source was CNN or something similar. The headline was this : AI is racist. They were saying AI could recognize faces of white people better than faces of black people.

b. Ok, first of all, AI is just an algorithm, so it doesn't make sense for it to be racist. My guess is that the training data set they used was biased. In other words, most of the training samples are faces of white people, thus the algorithm had hard time detecting faces of black people in the real world application (i.e. test set). Therefore, the AI "seems" like it's racist, but it is not. It's just that the model was designed in such a way that it works well only on white people.

c.   1. Make sure you have similar distributions of faces of different races in the training set.

2. Make sure the test set / training set represents the samples from real world.
(i.e. don't train a model with human faces and tries to detect dog faces)

3. Make sure your model performs well on test set. In case you couldn't find balanced data in step 1, use F1 score to measure the performance (don't use accuracy!). In case you couldn't find much dataset in step 1, use cross validation to measure the test set performance

Problem 2: kMeans Classification

You are provided with the following training data that has been preassigned to a 2 Cluster (k = 2) kMeans clustering model. You will be leveraging kMeans as a Classifier for the following problem.

$$\frac{24}{3} + \frac{7}{3} \qquad \frac{31}{3}$$

$$\frac{12}{14}{21}$$

$$\frac{15}{24} \qquad 8+1$$

Cluster = 0	Cluster = 1
(1,4)	(7,3)
(2,9)	(8,5)
(4,8)	(9,1)

Part a.
Calculate the Centroid for each cluster, and then add it to the plot above.

$$C_0 = \frac{(1,4)+(2,9)+(4,8)}{3} = (\frac{7}{3}, 7) \qquad C1 = \frac{(7,3)+(8,5)+(9,1)}{3} = (8,3)$$

Part b.
Calculate the vertex (edge point) between the clusters and mark it on the plot above. Then use that point to draw the approximate Voronoi regions (you can eyeball the approximate separation, just ensure it cuts through the vertex).

$$EP = \frac{C0 + C1}{2} = \frac{(\frac{7}{3}, 7) + (8,3)}{2} = (\frac{31}{6}, 5)$$

Part c.
The following four datapoints are provided for test data, along with their actual labels. Use your trained models to predict the corresponding labels. Report on your model's **Accuracy, Precision, Recall,** and **F1 Score.** (Note: In this model a label of '1' corresponds to a positive label).

X	Y	Predicted Label	Actual Label
7	3	1	0
6	1	1	1
2	2	0	0
8	8	1	0

prediction

Actual		N	P
	N	1	2
	P	0	1

$TP = 1$
$TN = 1$
$FP = 2$
$FN = 0$

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} = \frac{2}{4} = \frac{1}{2}$$

$$precision = \frac{TP}{TP+FP} = \frac{1}{3}$$

$$recall = \frac{TP}{TP+FN} = \frac{1}{1} = 1$$

$$F1 = 2 \cdot \frac{precision \times recall}{\ \ } = 2 \cdot \frac{\frac{1}{3}}{\ } = 2 \cdot \frac{1}{4} = \frac{1}{2}$$

2

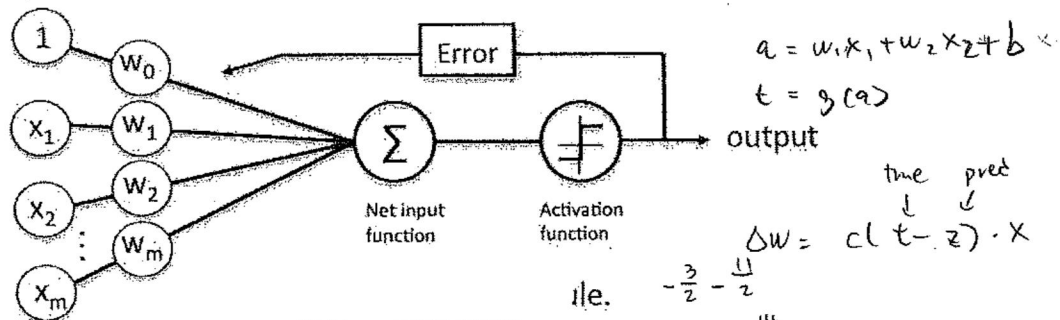**Cont. Problem 2:**

Part d.

Assuming the testing sample is representative of the overall performance of the model, interpret the overall performance of your model. In what circumstances would you consider using it, and when would you consider it insufficient?

It has good recall but bad in accuracy, precison, and F1.

This is a good model to use if you don't care about false positive, but care about false negative.

Otherwise, this is not a good model to use.

Problem 3

$a = w_1 x_1 + w_2 x_2 + b$

$t = g(a)$

output

$$\Delta w = c(\underset{true}{t} - \underset{pred}{z}) \cdot x$$

ıle. $-\frac{3}{2} \quad -\frac{11}{2}$

$-\frac{14}{2}$

$-9 \quad -21 \quad -2 \qquad -30$

$9 + 22$

Assume that you are given observations $(x, y) \in \mathfrak{R}^2 \times \{\pm 1\}$ in the following order

Instance	1	2	3	4	5	6	7	8
Label y	+1	-1	+1	-1	+1	-1	+1	+1
Data (x_1, x_2)	(9, 9)	(-1, -1)	(7, 3)	(2, 2)	(3, 7)	(-0.5, -0.5)	(3, 2)	(1, 4)

$\frac{22}{\frac{5}{3}}$

$3 + 44$

Show how the perceptron algorithm would apply updates for the above sequence of observations. We start with the initial set of weights $w = (1, 1)$ and $b = 0$. Use a learning weight of one $\leftarrow c = 1$

---

1 : $x_1 = 9, x_2 = 9, y = 1$

$y_{pred} = 1 \cdot 9 + 1 \cdot 9 + 0 = 18 > 0$

$z = 1$

$(y - z) = 0$

Thus $\Delta b = \Delta w_1 = \Delta w_2 = 0$

---

2 : $x_1 = -1, x = -1, y = -1$

$y_{pred} = 1 \cdot (-1) + 1 \cdot (-1) + 0 = -2 < 0$

$z = -1$

$y - z = 0$

Thus $\Delta b = \Delta w_1 = \Delta w_2 = 0$

---

3 : $x_1 = 7, x_2 = 3, y = 1$

$y_{pred} = 1 \cdot 7 + 1 \cdot 3 + 0 = 10 > 0$

$z = 1$

$y - z = 0$

Thus $\Delta b = \Delta w_1 = \Delta w_2 = 0$

---

4 : $x_1 = 2, x_2 = 2, y = -1$

$y_{pred} = 1 \cdot 2 + 1 \cdot 2 + 0 = 4 > 0$

$z = 1$

$c(y - z) = 1 \cdot (-2) = -2$

$\Delta b = (-2) \cdot 1 = -2 \quad \Delta w_1 = 1 \cdot 2 = -4 \quad \Delta w_2 = (-2) \cdot 2 = -4$

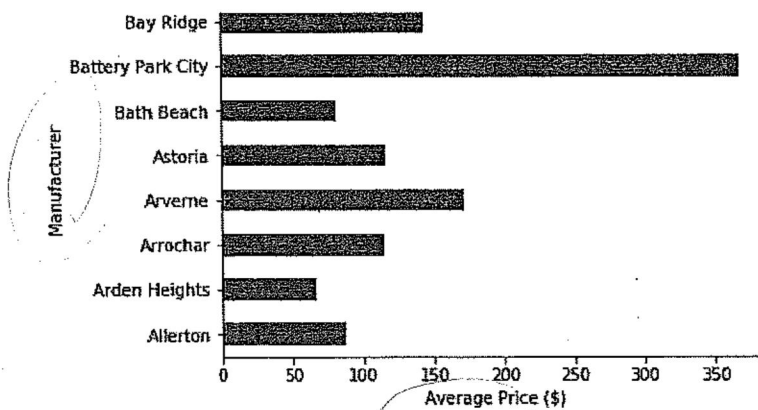
5 $x_1 = 3 \quad x_2 = 7 \quad y = 1$

$y_{pred} = (-3) \cdot 3 + (-3) \cdot 7 + (-2) = -32 < 0$

$z = -1$

$c(y - z) = 1(1 - (-1)) = 2$

$\Delta b = 2 \cdot 1 = 2 \quad \Delta w_1 = 2 \cdot 3 = 6 \quad \Delta w_2 = 2 \cdot 7 = 14$

New params
$b = -2$
$w_1 = -3$
$w_2 = -3$

---

6 $x_1 = -0.5 \quad x_2 = -0.5 \quad y = -1$

$y_{pred} = 3 \cdot (-0.5) + 11(-0.5) + 0 = -7 < 0$

$z = -1$

$y - z = 0$

$\Delta b = \Delta w_1 = \Delta w_2 = 0$

new params
$b = 0$
$w_1 = 3$
$w_2 = 11$

---

7 $x_1 = 3 \quad x_2 = 2 \quad y = 1$

$y_{pred} = 3 \cdot 3 + 11 \cdot 2 + 0 = 31 > 0$

$z = 1$

$y - z = 0$

$\Delta b = \Delta w_1 = \Delta w_2 = 0$

---

8 $x_1 = 1 \quad x_2 = 4 \quad y = 1$

$y_{pred} = 3 \cdot 1 + 11 \cdot 4 + 0 = 47 > 0$

$z = 1$

$y - z = 0$

$\Delta b = \Delta w_1 = \Delta w_2 = 0$

4

Average Price ($)

Name(last, first): ___Wang Junhong___

Problem 4 A table cereal contains one row with nutritional information for each of a set of cereal brands.

The table contains ten columns:

- **name**: a string, the name of the cereal
- **mfr**: a string, the initial of the manufacturer
- **calories**: an int, calories per serving
- **protein**: an int, grams of protein per serving
- **fat**: an int, grams of fat per serving
- **fiber**: an int, grams of fiber per serving
- **sugars**: an int, grams of sugar per serving
- **shelf**: an int, what shelf the cereal is displayed on at a certain supermarket (1, 2, or 3)
- **organic**: a boolean, whether the cereal is organic
- **rating**: a float, giving a rating of the cereal on a scale of 1–100 from Consumer Reports (CR)
- price , mfr

For each of the following questions, shade in one or more circles corresponding to your answer.

1. Given the data types of this dataset, which of the following visualizations are not appropriate?

   ○ A histogram of calories

   ○ A histogram of organic

   ◉ A scatter plot with mfr on the x axis and rating on the y-axis

   ○ A scatter plot with sugars on the x axis and calories on the y axis

2. Write a statement to select only cereals on aisle 3 whose sugar content is less than 10g per serving

   selected_cereals = ___selected_cereals [(selected_cereals ['shelf'] = 3) &___
   (selected_cereals ['sugars'] < 10)]

3. Write code that would generate a plot identical to the one below. Hint: the pd.DataFrame class has a .plot.barh() method that will make a horizontal bar plot with the specified x (labels) and y (values) axes. One approach takes only three lines. **Answer on next page**

3 (answer here)

x = selected_cereals . groupby (" mfr ") . [" mfr "]

Y = 'selected_ cereals . groupby (" mfr ")["price"] / len(y)

pd . plot . barh (x= x , Y = y)

4. What conclusion or conclusions are justified by the table of correlations below? (select all that apply)

Correlation of Features with Rating

Feature	Pearson's Correlation Coefficient
Sugar	.5
Fat	.3
Protein	-.4
Calories	0

○ A model that uses both sugar and fat to predict rating can do no better than a model that only uses sugar as the correlation between sugar and rating is stronger than that between fat and rating

○ Protein is not useful for predicting rating as its correlation coefficient is less than zero

◉ There is no association between calories and rating

○ Increased sugar is positively associated with increased fat, as both are positively associated with higher ratings

○ None of the above

6