CS 188 Midterm

STEWART DULANEY

TOTAL POINTS

73 / 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 6 / 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 23 / 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
- **5 pts** forgot bias
- **28 pts** no attempt

QUESTION 4

Project 28 pts

4.1 Visualizations 3.5 / 7

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

✓ - 3.5 pts **Did not select B**

- **3.5 pts** Did not select C
- **3.5 pts** Selected A
- **3.5 pts** Selected D

4.2 Filtering 3.5 / 7

- **0 pts** Correct

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

- **1.5 pts** Returned boolean mask rather than the actual rows
- **1 pts** Used "and" instead of &

- **1 pts** Used "&&" instead of &
- **1.5 pts** Used "," instead of &
- **1.5 pts** Used "|" instead of &
- **1 pts** Used iloc instead of loc
- **2 pts** Used where instead of loc
- **3.5 pts** Used group instead of loc
- **1 pts** No parentheses to group operations, would result in logic error
- **3.5 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"
- **0.5 pts** No quotations around column name
- **0.5 pts** Unnecessary quotations around column name
- **0.5 pts** Incorrect quotation in query
- **3.5 pts** Wrote two solutions, one of which was incorrect
- **0.5 pts** Used "=" instead of "=="
- **0.5 pts** Used <= instead of <
- **1 pts** Unnecessary inclusion of colon operator
- **7 pts** Incorrect
- **7 pts** Blank

4.3 Plotting 3 / 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** Forgot quotes around column name

✓ **- 0.5 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 0 / 7

- **7 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
- **7 pts** B: A negative correlation indicates that cereals with higher protein tend to have lower ratings

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

✓ **- 7 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

Name(last, first): Dulaney, Stewart

UCLA Computer Science Department

CS 188 **Data Science Fundamentals** **UID:** 904-064-791

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) Coffee, drink it to perform better & live longer
"Study shows 6 cups/day lowers diabetes risk by 50%" (bulletproof.com)

b) Read paper, question answered was not same as assertion by article. Also, sample only 11 young healthy males. Not data neutral. No validation on if subjects got diabetes later. Did not even calc diabetes risk, only measured insulin/glucose levels. Author twisting results to sell his product. Bulletproof coffee.

c) Question: Does drinking coffee cause reduced risk for diabetes?

Features: age, gender, smoke/no smoke, bmi, activity level, glucose/insulin levels

Labels: high / low / medium risk for diabetes

Data collection: smart phone app where users log features

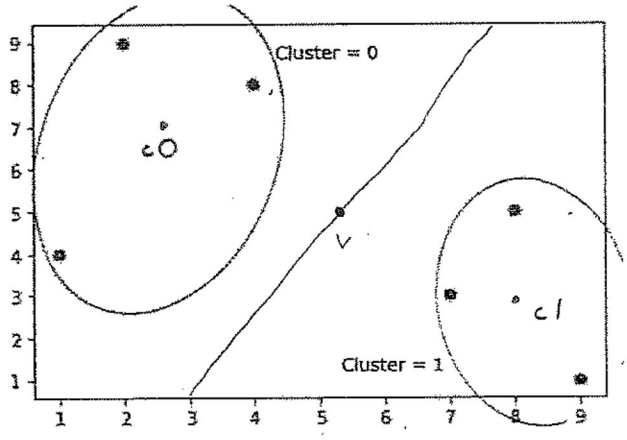
Model: KNN b/c the data will likely be multi-clustered and non-linear, will use small dataset w/ small amt of features at first.

Validation: use labeled dataset, based on medical evaluation

Name(last, first): _____

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.



label = 1

Cluster = 0
(1,4)
(2,9)
(4,8)

Cluster = 1
(7,3)
(8,5)
(9,1)

label = 0

Part a.

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

$c0 = \left(\frac{1+2+4}{3}, \frac{4+9+8}{3}\right) = \left(\frac{7}{3}, 7\right)$ $c1 = \left(\frac{7+8+9}{3}, \frac{3+5+1}{3}\right) = (8, 3)$

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).

$v = \left(\frac{\frac{7}{3}+8}{2}, \frac{7+3}{2}\right) = \left(\frac{\frac{7}{3}+\frac{24}{3}}{2}, 5\right) = \left(\frac{31}{6}, 5\right)$

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	0	0
6	1	0	1
2	2	1	0
8	8	0	0

$Acc = \frac{2}{4}$

$Prec = \frac{0}{0+1} = 0$

$Recall = \frac{0}{0+1} = 0$

$F1 = \frac{2 \cdot 0 \cdot 0}{0+0} = 0$

$\frac{\#\text{ correct}}{\text{total}}$ $\frac{TP}{TP+FN}$ $\frac{2 \cdot Prec \cdot Rec}{Prec + Rec}$

Name(last, first): _____

Cont. Problem 2:

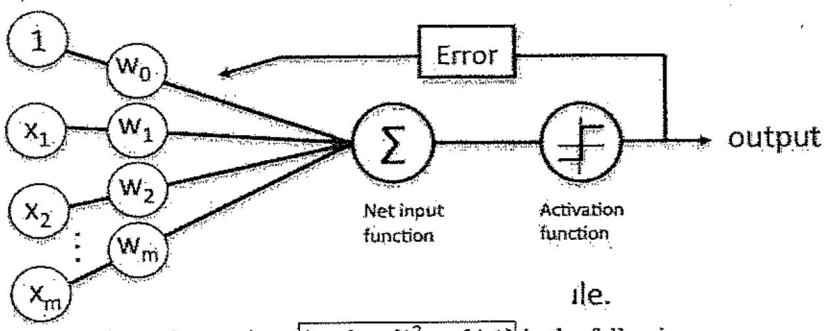
Part d.

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?

Overall very poor b/c 50% accuracy is not even better than random so not useful. Insufficient in almost all cases b/c of this. Would consider using if FP's and FN's are not critical, b/c precision and recall indicate these would happen at high rate.

Name(last, first): _____

Problem 3



Assume that you are given observations $(x, y) \in \Re^2 \times \{\pm 1\}$ 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)

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.

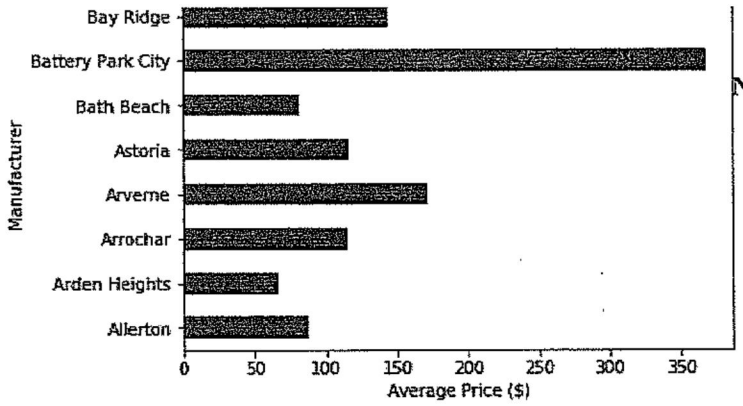
$b + \sum x_i w_i > 0$ 1
else -1

$c = 1$

$\Delta w = c(t - z)x_i$

$b = 0 = x_0 w_0$

x_0	x_1	x_2	t	w_0	w_1	w_2	net	z	Δw_0	Δw_1	Δw_2
0	9	9	1	1	1	1	18	1	0	0	0
0	-1	-1	-1	1	1	1	-2	-1	0	0	0
0	7	3	1	1	1	1	10	1	0	0	0
0	2	2	-1	1	1	1	4	1	0	-4	-4
0	3	9	1	1	-3	-3	-30	-1	0	6	14
0	-0.5	-0.5	-1	1	3	11	-7	-1	0	0	0
0	3	2	1	1	3	11	31	1	0	0	0
0	1	4	1	1	3	11	47	1	0	0	0



Name(last, first): _____

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)

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 = `cereal["shelf"] == 3 ]["sugar" < 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**

Name(last, first): _____

3 (answer here)

```
x1 = cereal["mfr"]
y1 = cereal.groupby("mfr","price").mean()
cereal.plot.bar(x=x1,y=y1)
plt.show()
```

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