

CS 188 Midterm

ANIRUDH MANI

TOTAL POINTS

93.75 / 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

- 8 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 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
 - **0.5 pts** Incorrect attempt to set axis labels
 - **0.25 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

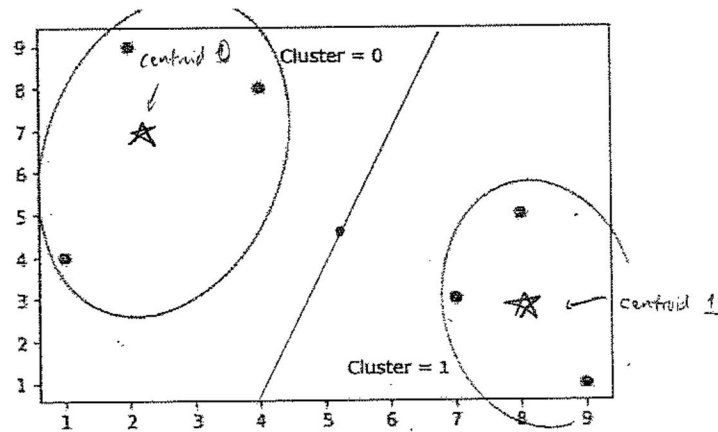
UCLA Computer Science Department**CS 188****Data Science Fundamentals****UID:** 005111192**Midterm****Total Time: 1 hour****February 26, 2020**Problem 1 has **16 points**. All other problems have **28 points**.You **cannot** use the back of the pages for your answers (you are welcome to use them as a 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. There was a headline that states "We were advised previously to eat less red meat. A new study says otherwise!" It was an article in the New York Times published September 2019.
- b. The new study cited in the article had a couple of problems. First, it was an aggregation of old studies (called a meta-analysis). However, they were not able to normalize across studies. For example, some studies would compare the health effects of a diet without red meat but still eating chicken, while other studies compared a red meat diet to a vegetarians. Another problem with the study is that it included information such as the opinions of whether or not people liked eating red meat and used that in consideration on whether eating red meat would be good. Obviously, people's opinions should not be factored into a health decision.
- c. I would organized a randomized controlled trial with 3 groups of people. One would be a group who continued to eat red meat, one who substituted red meat consumption with chicken and a third who switched to a vegetarian diet. I would control each group to have roughly the same variance in demographic (sex, age, race, etc). Then, I would continue this trial for a statistically significant amount of time (such as 2-3 years). The features I would note are their change in cholesterol, weight, likelihood of cancer, tracking vitals such as ¹vitamin levels and their ^{change in} blood pressure. All of the features would be changes, calculated by taking the difference between start and end levels.

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.



Cluster = 0

(1,4)
(2,9)
(4,8)

Cluster = 1

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

Part a.

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

Centroid for cluster 0: (2.33, 7) Centroid for cluster 1: (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).

vertex = (5.165, 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).

$$\text{Accuracy} = \frac{2}{4}$$

$$\text{Precision} = \frac{1}{3}$$

$$\text{Recall} = \frac{1}{1}$$

$$\text{F1} = \frac{2}{\frac{1}{\frac{2}{3}} + 1} = \frac{2}{\frac{3}{2} + 1} = \frac{2}{\frac{5}{2}} = \frac{2}{3} \cdot \frac{2}{5} = \frac{4}{15}$$

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

	Predict P	Predict N
Actual P	1	0
Actual N	2	1

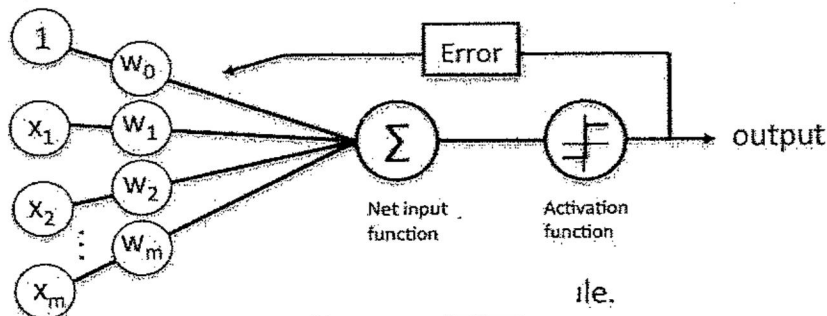
Cont. Problem 2:

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?

The model has a very good recall of 1, meaning it does not have a lot of false negatives. This would be very good for when it is very costly to miss a positive case, such as an earthquake or tsunami. However the model is not good if we wanted to avoid false positives, as our precision is $\frac{1}{3}$. This might be in a case such as sentencing someone to death penalty, as we want to have a high probability that we don't wrongfully convict. The F1 score is the harmonic mean of the precision and recall, and a score of $\frac{1}{2}$ means it is not that great (in this case, really good at recall and bad at precision). The accuracy of the model of $\frac{1}{2}$ tells us it is about as good as a coin in determining the correct result, as it is just doing $\frac{TP+FP}{all}$

Problem 3



Assume that you are given observations $(x, y) \in \mathbb{R}^2 \times \{+1\}$ 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.

first epoch

	x_1	x_2	w_0	w_1	w_2	predicted	true
1	9	9	0	1	1	1	1
2	-1	-1	0	1	1	-1	-1
3	7	3	0	1	1	1	1
4	2	2	0	1	1	1	-1
5	3	7	-2	-3	-3	-1	1
6	-0.5	-0.5	0	3	11	-1	1
7	3	2	0	3	11	1	1
8	1	4	0	3	11	1	1

← weights don't change

$$\Delta w_0 = 1(-1-1) = -2$$

$$\Delta w_1 = 1(-1-1) \cdot x_1^2 = -4$$

$$\Delta w_2 = 1(-1-1) \cdot x_2^2 = -4$$

$$\Delta w_0 = 1(1-1) = 2$$

$$\Delta w_1 = 1(1-1) \cdot 3 = 6$$

$$\Delta w_2 = 1(2)(7) = 14$$

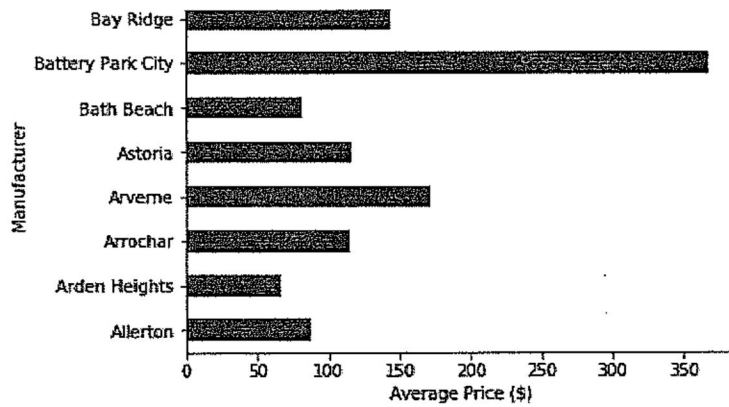
second epoch

1	9	9	0	3	11	1	1
2	-1	-1	0	3	11	-1	-1
3	7	3	0	3	11	1	1
4	2	2	0	3	11	1	-1
5	3	7	-2	-1	7	1	1
6	-0.5	-0.5	-2	-1	7	-1	1
7	3	2	-2	-1	7	1	1
8	1	4	-2	-1	7	1	1

third epoch

	w_0	w_1	w_2
9	9	-2	-1
10	9	-2	-1
11	9	-2	-1
12	9	-2	-1
13	9	-2	-1
14	9	-2	-1
15	9	-2	-1
16	9	-2	-1
17	9	-2	-1
18	9	-2	-1
19	9	-2	-1
20	9	-2	-1
21	9	-2	-1
22	9	-2	-1
23	9	-2	-1
24	9	-2	-1
25	9	-2	-1
26	9	-2	-1
27	9	-2	-1
28	9	-2	-1
29	9	-2	-1
30	9	-2	-1
31	9	-2	-1
32	9	-2	-1
33	9	-2	-1
34	9	-2	-1
35	9	-2	-1
36	9	-2	-1
37	9	-2	-1
38	9	-2	-1
39	9	-2	-1
40	9	-2	-1
41	9	-2	-1
42	9	-2	-1
43	9	-2	-1
44	9	-2	-1
45	9	-2	-1
46	9	-2	-1
47	9	-2	-1
48	9	-2	-1
49	9	-2	-1
50	9	-2	-1

it will eventually converge after many iterations of the algorithm



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.

- 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

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

selected_cereals = `data[data["shelf"] == 3 && data["sugars"] < 10]`

- 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_labels = data["mfr"].vals()
y_labels = data.groupby(["mfr"])["price"].avg()
plot.barh(x_labels, y_labels)

```

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

