CISC 3225: Final review

The final includes all material from the midterm and the following topics:

_k_-Nearest-Neighbors and _k_-Means Clustering

- Classification and regression as machine learning tasks
- Types of machine learning: supervised vs. unsupervised
- The *k*-Nearest-Neighbors algorithm
- Evaluating *k*-Nearest-Neighbors classification with test/validation sets and accuracy
- *k*-Means Clustering algorithm, finding a good value of *k*
- Analyzing and interpreting clusters
- Data normalization
- Encoding categorical variables

Visualization

- Basic differences in functionality between Matplotlib and Seaborn
- The Matplotlib global state machine
- Subplots

Principal component analysis

- What it is for and basic mechanics of how it works
- Interpretability of principal components

Statistics and Hypothesis Testing

- What statistical tests are for
- Components of a statistical test: Your hypothesis, null/alternative hypothesis
- Errors: Type 1 and Type 2
- Coin flipping example
- *p*-value: What it is and how to interpret it
- Types of statistical tests:
  - Pearson r: What it tests, r value and p value, how to interpret
  - Linear regression: What it tests, slope and intercept, how to interpret
  - T-test: What it tests, how to interpret
  - Chi-squared test: What it tests, how to interpret, contingency tables, expected vs. observed frequency

1. True/False

____ If you perform *k*-means clustering on a dataset containing different types of tea, the clustering can help you classify new teas you have never seen before.

____ You discover that A is positively correlated with variable B, and the effect is statistically significant. You can conclude that A caused the change in B.

____ Grouping and aggregating is a good way to smooth noisy data for visualization.

____ In statistical tests, a large *p*-value is justification for rejecting the null hypothesis.

Multiple Choice

2. How does the *k*-means clustering algorithm know when to stop iterating?

a) When all instances within each cluster are of the same type.
b) When the centroids stop changing position.
c) When there are exactly *k* instances within each cluster.
d) The *k*-means clustering algorithm is not iterative.

3. You want to find out how much a change in A contributes to a change in B. You should perform:

a) Pearson's correlation test
b) A linear regression
c) A T-test
d) A chi-squared test

4. How does principal component analysis find the top principal component?

a) Look for the direction with the highest variance.
b) Look for the direction with the lowest variance.
c) Look for the axis with the highest variance.
d) Look for the axis with the least variance.

5. What is a Type 2 error?

a) You reject the null hypothesis, but it is actually true.
b) You fail to reject the null hypothesis, but it is actually false.
c) The test shows that the magnitude of the effect is too small.
d) None of these are true.

Short Answer and Analysis Questions

6. Imagine you are performing *k*-Nearest-Neighbors classification in Pandas on a dataset of different types of coffee. You are trying to predict a quality rating ("High", "Medium", or "Low"). If you call .info() on your training data, you see the following output:

```
RangeIndex: 893 entries, 0 to 892
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  -----------  --------------  -----
 0   bitterness   893 non-null    float64
 1   sweetness    893 non-null    float64
 2   acidity      891 non-null    float64
 3   moisture     893 non-null    float64
 4   roast_level  880 non-null    float64
 5   color        893 non-null    float64
 6   oil          893 non-null    float64
 7   price        152 non-null    float64
 8   sources      893 non-null    int64
 9   quality      893 non-null    object
```

Discuss what you should do to preprocess the dataset. Address the following points:

- Is it a good idea to perform *k*-NN classification with the dataset as-is? Why or why not?
- Which row(s) should you drop and why?
- Which column(s) should you drop and why?

7. Below are two clusterings of the nutritional content of Starbucks beverages, one with k=2 and one with k=3. The charts show the mean value for each variable within each cluster.

	0	1
Calories	282.5000	197.938095
Total Fat (g)	2.1250	2.952381
Trans Fat (g)	0.8625	1.420000
Saturated Fat (g)	0.0125	0.040476
Sodium (mg)	6.2500	6.690476
Total Carbohydrates (g)	130.6250	137.095238
Cholesterol (mg)	54.7500	36.490476
Dietary Fibre (g)	6.7500	0.619048
Sugars (g)	37.1250	33.819048
Protein (g)	16.7500	6.945238
Vitamin A (% DV)	21.7500	9.876190
Vitamin C (% DV)	71.2500	1.166667
Calcium (% DV)	12.5000	22.033333
Iron (% DV)	13.2500	7.842857
Caffeine (mg)	5.6250	93.142857

	0	1	2
Calories	101.320988	323.333333	232.000000
Total Fat (g)	1.172840	7.380952	2.442105
Trans Fat (g)	0.413580	3.995238	1.092632
Saturated Fat (g)	0.013580	0.138095	0.017895
Sodium (mg)	3.024691	18.809524	4.421053
Total Carbohydrates (g)	67.777778	205.833333	165.263158
Cholesterol (mg)	17.765432	54.547619	46.010526
Dietary Fibre (g)	0.271605	1.214286	1.168421
Sugars (g)	16.222222	49.857143	42.010526
Protein (g)	4.574074	12.357143	7.400000
Vitamin A (% DV)	6.802469	16.333333	10.642105
Vitamin C (% DV)	0.098765	3.476190	6.957895
Calcium (% DV)	14.814815	37.023810	20.757895
Iron (% DV)	3.209877	13.714286	9.652632
Caffeine (mg)	96.604938	111.071429	74.894737

Which clustering do you think is more useful? Pick the best one and broadly characterize the different clusters based on the mean values shown above. As part of your answer, discuss specific aspects of the better cluster that make it superior.

If necessary, use conventional nutritional wisdom in your analysis:
Bad: Excessive calories, fat, sodium, carbohydrates, cholesterol, sugar.
Good: Fiber, protein, vitamins, and minerals.

You may make the following assumptions:

- All drinks contain coffee
- Coffee by itself has very few calories, fat, vitamins, and minerals
- Drinks consist of coffee mixed with any number of additions, including dairy or imitation milks (nut, oat, soy), flavoring syrups, spices, and sugar.

8. You want to perform clustering on the following dataset of penguins. The dataset does not contain missing values.

Identify two problems with the data, and discuss how you will fix them. If necessary, give the formula for any mathematical operations on the data..

	culmen_depth_mm	flipper_length_mm	species
0	18.7	181.0	Gentoo
1	17.4	186.0	Adelie
2	18.0	195.0	Gentoo
4	19.3	193.0	Chinstrap
5	20.6	190.0	Chinstrap
...	...	...	...
338	13.7	214.0	Adelie
340	14.3	215.0	Gentoo
341	15.7	222.0	Gentoo
342	14.8	212.0	Gentoo
343	16.1	213.0	Adelie

9. Below is the result of a chi-squared test performed on several people surveyed about their dominant hand. Discuss the input and output of the chi-squared test. What does the contingency table and the results mean?

Input
Contingency table:

	Right-Handed	**Left-Handed**
American	236	19
Canadian	157	16

Output
Statistic: 0.24

p: 0.63

Expected frequency:

234.14 20.85
158.85 14.15

1. True/False

F___ If you perform *k*-means clustering on a dataset containing different types of tea, the clustering can help you classify new teas you have never seen before.

F___ You discover that A is positively correlated with variable B, and the effect is statistically significant. Because of this, you can conclude that A caused the change in B.

T___ Grouping and aggregating is a good way to smooth noisy data for visualization.

F___ In statistical tests, a large *p*-value is justification for rejecting the null hypothesis.

Multiple Choice

2. How does the *k*-means clustering algorithm know when to stop iterating?

a) When all instances within each cluster are of the same type.
b) When the centroids stop changing position.
c) When there are exactly *k* instances within each cluster.
d) The *k*-means clustering algorithm is not iterative.

3. You want to find out how much a change in A contributes to a change in B. You should perform:

a) Pearson's correlation test
b) A linear regression
c) A T-test
d) A chi-squared test

4. How does principal component analysis find the top principal component?

a) Look for the direction with the highest variance.
b) Look for the direction with the lowest variance.
c) Look for the axis with the highest variance.
d) Look for the axis with the least variance.

5. What is a Type 2 error?

a) You reject the null hypothesis, but it is actually true.
b) You fail to reject the null hypothesis, but it is actually false.
c) The test shows that the magnitude of the effect is too small.
d) None of these are true.

Short Answer

6.

It is not a good idea to perform *k*-nearest-neighbor classification with the dataset as-is.

The dataset contains missing values, which would impact the algorithm's distance calculations. The distance between a new coffee instance and an existing coffee instance with a missing value is NaN.

I would drop rows that are missing values for acidity and roast_level. Very few rows would be affected, so most of the data would remain intact.

I would drop the entire price column. It has too many missing values to be useful for k-nearest-neighbor classification. If we dropped rows with missing price values, only 152 rows would remain.

7.

k=3 is the better clustering. It has a pretty clear progression of "less" (cluster 0), "medium" (cluster 2), and "most" (cluster 1) nutritional information. That is, cluster 0 contains beverages with generally low nutrition values, cluster 2 contains beverages with generally moderate nutrition values, and cluster 1 contains beverages with generally high nutrition values. The k=2 example did not have a high enough value of k to identify extreme examples of beverages.

This makes sense given the assumptions: you can assume that cluster 0 contains drinks with few additions, cluster 2 contains drinks with some (possibly low-fat) additions, and cluster 1 contains drinks with lots of additions. There are some exceptions (notably caffeine), but the less-medium-most pattern applies to most of the variables.

8.

Problem 1: The “species” column is a categorical variable, and cannot be used to find clusters. It should be one-hot encoded. This would result in the dataset looking like this (first 4 rows shown):

	culmen_depth_mm	flipper_length_mm	species_Gentoo	species_Adelie	species_Chinstrap
0	18.7	181.0	1	0	0
1	17.4	186.0	0	1	0
2	18.0	195.0	1	0	0
4	19.3	193.0	0	0	1

Problem 2: culmen_depth_mm and flipper_length_mm are on different scales. Culmen depths are tens of millimeters, and flipper lengths are hundreds of millimeters. We should normalize them to bring them to approximately the same scale. We can do this with z-score normalization, which makes each column have a mean of 0 and variance of 1. We can compute it as $z = (x - \mu) / \sigma$, where x is a value in a column, μ is the mean of a column, and σ is the standard deviation of a column.

9.

The contingency table shows a count of the number of people who are left- and right-handed who are American and Canadian. That is, there are 236 American right-handed people, 19 American left-handed people, 157 Canadian right-handed people, and 16 Canadian left-handed people.

The p-value in the chi-squared test output is very large, above any reasonable cutoff ($p>0.05$ and $p>0.1$). Thus, we fail to reject the null hypothesis. That is, there no relationship between nationality and dominant hand.

This lack of relationship is reflected in both the statistic and the expected frequency. The statistic is fairly small, and the expected frequency table is very similar to the input contingency table. This means that our observed right- and left-handedness is very similar to the frequency we would expect to see if there is no relationship between nationality and dominant hand.