4780/6780 Fundamentals of Data Science
Final Exam

Instructor: Kiril Kuzmin

05/06/2025

Name: _____

HONOR CODE STATEMENT: I will not commit any act of academic dishonesty while completing this assignment. I am fully aware that any of my own personal actions while attempting this assignment that are interpreted as academic dishonesty, will be treated as such. I understand that if I am held accountable for an act of academic dishonesty that I will receive a grade of "0" (zero) for this assignment and the incident will be reported to the Dean of Students Office.

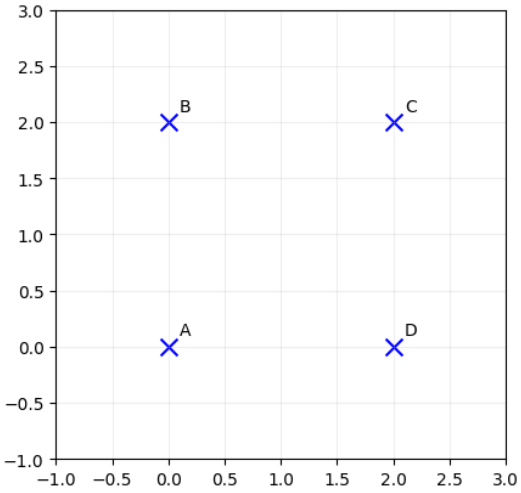
This exam contains 22 pages (including this cover page) and 10 problems. Total of points is 100.

Good luck and productive work!

Distribution of grades

Question	Points	Score
1	12	
2	18	
3	6	
4	6	
5	16	
6	4	
7	12	
8	10	
9	10	
10	6	
Total:	100	

1. Consider running the hierarchical agglomerative clustering algorithm on the following set of four points in $\mathbb{R}^2$, breaking ties *arbitrarily*:



We stop when only two clusters remain. Which of the following linkage methods **ensures** that the resulting clusters are balanced (i.e., each cluster has exactly two points)? For each method, **explain your reasoning**.

(a) (3 points) Complete linkage.

Solution: We may observe that the first merge is the same for all methods due to symmetry. Without loss of generality, we can assume that A merges with B first.

Next step: After merging A and B, the distances are:

- Distance between $\{A, B\}$ and C is $2\sqrt{2}$.
- Distance between $\{A, B\}$ and D is $2\sqrt{2}$.

D and C are still available to merge (at distance 2), so they merge together.

Therefore, we end up with two balanced clusters: $\{A, B\}$ and $\{C, D\}$.

Answer: Complete linkage ensures balanced clusters.

(b) (3 points) Single linkage.

Solution: The first merge is the same for all methods. Without loss of generality, we can assume that A merges with B first.

Next step: After merging A and B, the distances are:

- Minimum distance between $\{A, B\}$ and C is 2 (from B to C).
- Minimum distance between $\{A, B\}$ and D is 2 (from A to D).
- Distance between C and D is 2.

Since all these distances are equal, depending on tie-breaking, either:

- C and D could merge, leading to balanced clusters ($\{A, B\}$ and $\{C, D\}$),

or

- D could merge with $\{A, B\}$, leading to an unbalanced cluster $\{A, B, D\}$ and singleton $\{C\}$.

Thus, single linkage does not guarantee balanced clusters — it depends on tie-breaking.

Answer: Single linkage may not ensure balanced clusters.

(c) (3 points) Centroid linkage.

Solution: We can assume that A merges with B first.

Next step: The centroid of $\{A, B\}$ is at $(0, 1)$.

- Distance from centroid $(0, 1)$ to $D(2, 0)$ is $\sqrt{5} \approx 2.2$.
- Distance from centroid $(0, 1)$ to $C(2, 2)$ is $\sqrt{5} \approx 2.2$.
- Distance between C and D is 2.

Since D and C are closer (distance 2) than either is to the centroid of $\{A, B\}$, they will merge first.

Answer: Centroid linkage ensures balanced clusters.

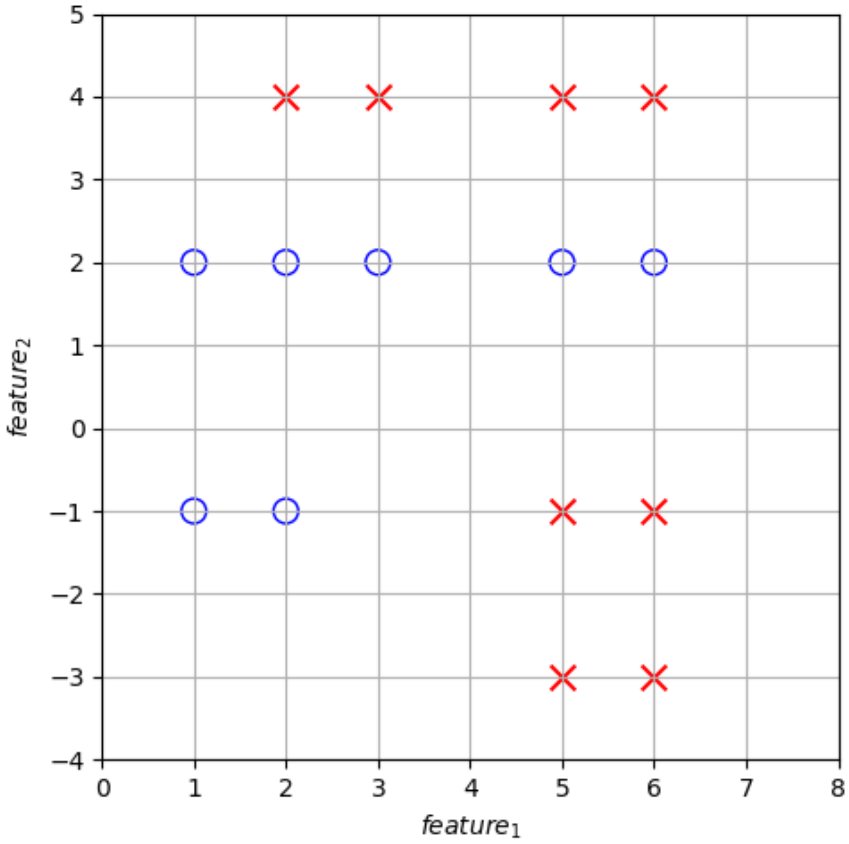
(d) (3 points) Average linkage.

Solution: We assume that A merges with B first.
Next step: After merging A and B:

- Distance between A and D is 2, and between B and D is $2\sqrt{2}$.
- Average distance between $\{A, B\}$ and D is $\frac{2+2\sqrt{2}}{2} > 2$.
- Similarly, average distance between $\{A, B\}$ and C is also > 2.
- Distance between C and D is 2.

Since C and D are closer to each other (distance 2) than to cluster $\{A, B\}$, they will merge first.
Answer: Average linkage ensures balanced clusters.

2. We seek to learn a classifier on the dataset shown below, consisting of 15 data points labeled as red crosses (class -1) and blue circles (class $+1$):

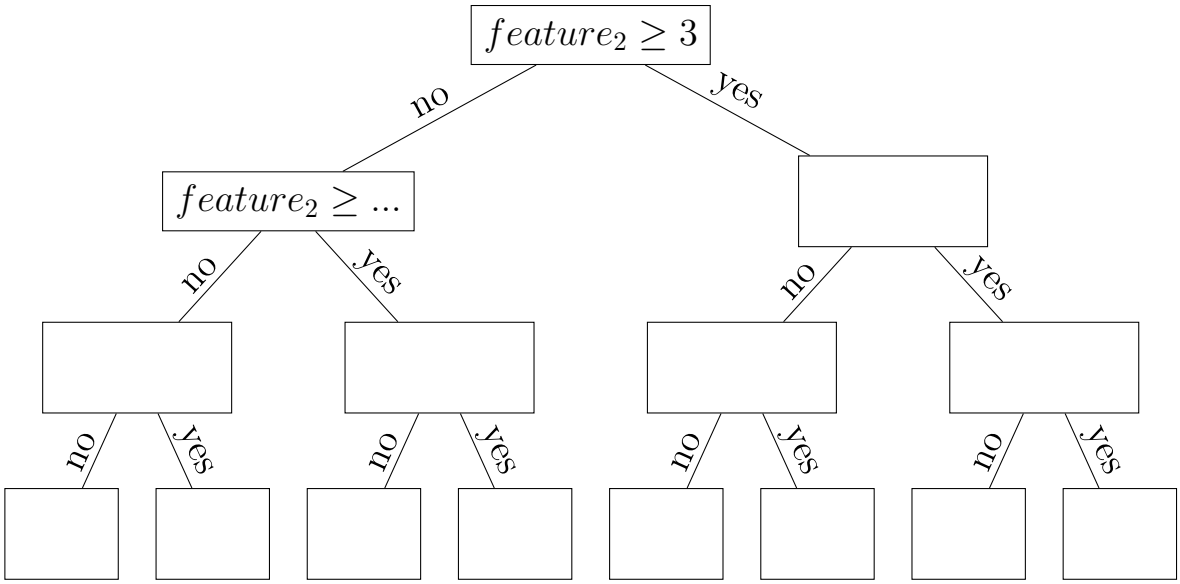


(a) (2 points) We first train a `depth-1` decision tree on this dataset. Will it achieve 100% training accuracy? Briefly explain.

(b) (6 points) We now train a `depth-3` decision tree on this dataset using the **entropy** criterion[1], with `min_samples_split = 4`. The partially completed tree is shown below, where the root node splits on `feature2 >= 3`, and one second-level split is also on `feature2`.

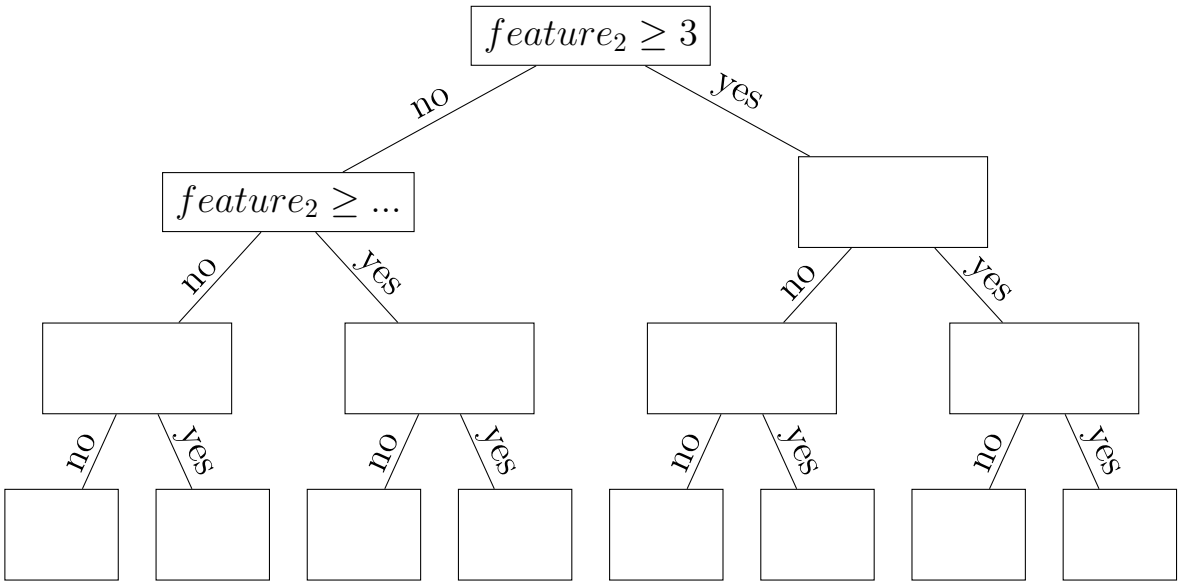
Complete the tree by filling in the missing split conditions and class labels at the leaves. Leave any unused boxes empty. Explain why you choose such splits (calculating entropy).

[1]The table of values for $-\frac{x}{y}\log\frac{x}{y}$ can be found on the second-to-last page, if needed.



(c) (2 points) What is the accuracy of the decision tree model from the task (b)?

(d) (6 points) Now suppose we set `min_samples_split = 7`. Again, complete the tree shown below (the partially completed tree has the root node which splits on `feature2 >= 3`, and the second-level splits are also on `feature2`). Fill in the leaf nodes where applicable, and leave other boxes empty if a split does not occur.

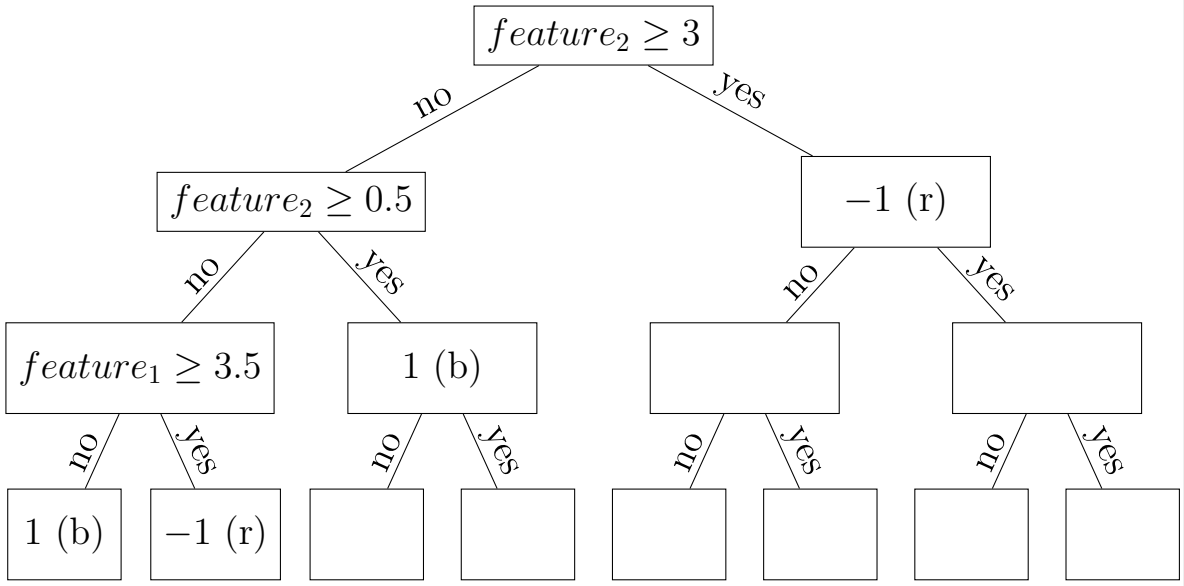


(e) (2 points) What is the accuracy of the decision tree model from the task (d)?

Solution:

(a) No, a depth-1 decision tree will not achieve perfect training accuracy, because no single split (on $feature1$ or $feature2$) can fully separate the red and blue classes in the dataset.

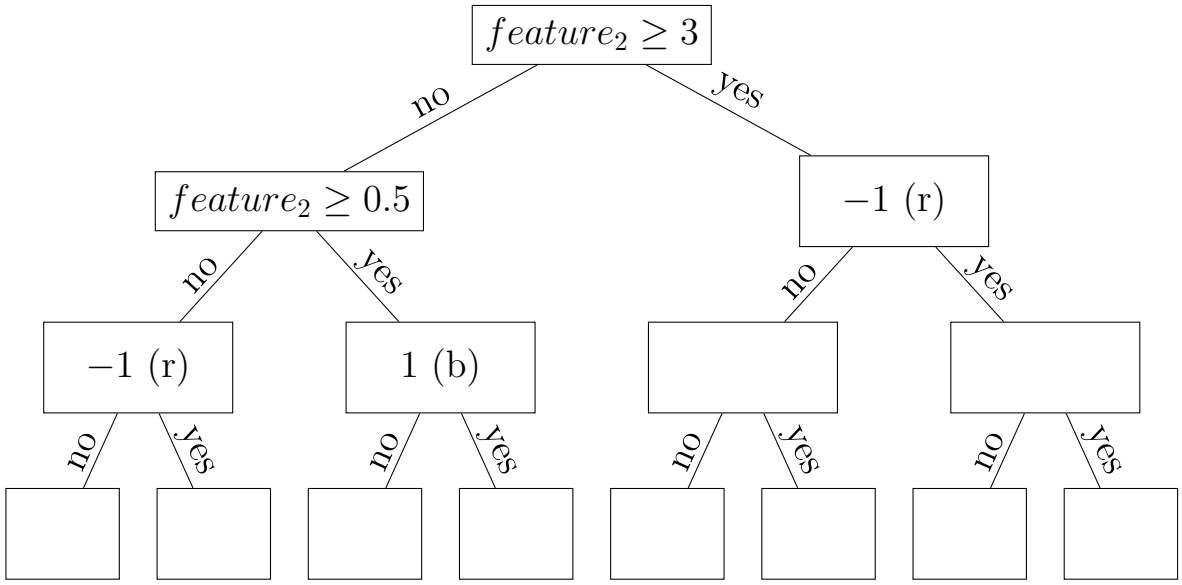
(b) The completed decision tree is shown below:



(c) Since all points are correctly identified,

$$\text{Accuracy} = 1 \quad \text{or} \quad 100\%.$$

(d) The completed decision tree using `min_samples_split = 7` is shown below. Some branches remain unexpanded due to the minimum split requirement:



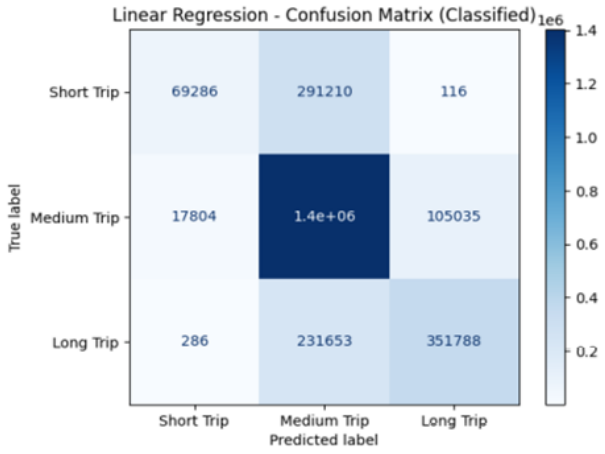
(e) Since two blue points are misclassified,

$$\text{Accuracy} = \frac{13}{15} \approx 0.87 \quad \text{or} \quad 87\%.$$

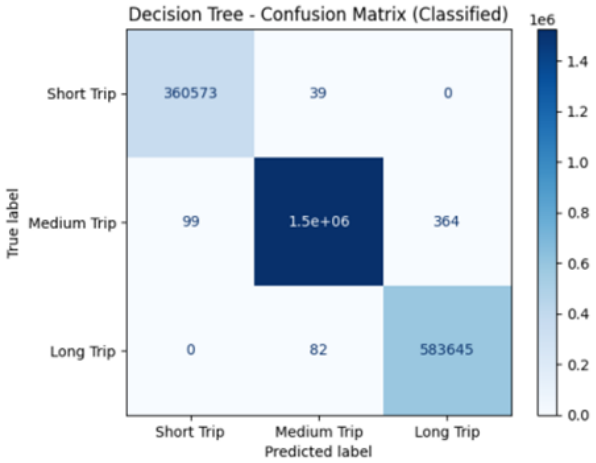
3. A group of beginner data scientists is working on a project to classify trips into *Short*, *Medium*, or *Long*. They report the following results for two models:

   - Model A: Linear Regression
   - Model B: Decision Tree

   The group presents the confusion matrices and classification reports for each model as shown below:



	precision	recall	f1-score	support
Short Trip	0.77	0.60	0.68	583727
Medium Trip	0.73	0.92	0.81	1524841
Long Trip	0.79	0.19	0.31	360612
accuracy			0.74	2469180
macro avg	0.76	0.57	0.60	2469180
weighted avg	0.75	0.74	0.71	2469180



	precision	recall	f1-score	support
Short Trip	1.00	1.00	1.00	583727
Medium Trip	1.00	1.00	1.00	1524841
Long Trip	1.00	1.00	1.00	360612
accuracy			1.00	2469180
macro avg	1.00	1.00	1.00	2469180
weighted avg	1.00	1.00	1.00	2469180

(a) (2 points) What is problematic about how Model A? Explain briefly.

Solution: Model A is a Linear Regression model, which is not suitable for direct classification.

(b) (4 points) What is suspicious about Model B? What could be the cause?

Solution: Model B shows nearly perfect classification performance across all classes, which is highly unusual. This could indicate data leakage (e.g., label-related features or incorrect splitting of train/test data), or extreme overfitting of the decision tree. Either way, the reported performance is likely unrealistic and should be carefully investigated.

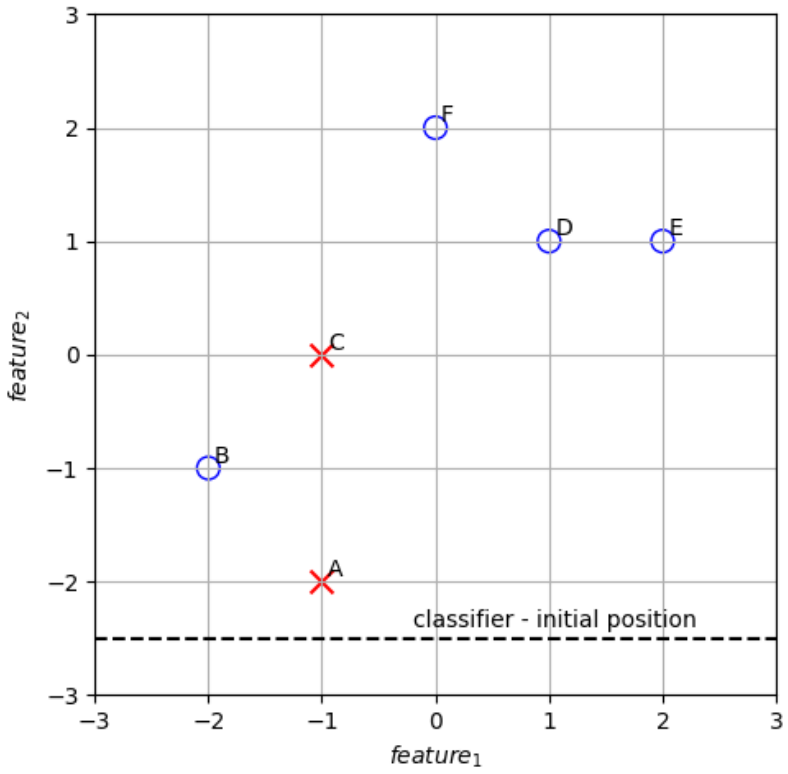
4. (a) (4 points) Why is *L1 regularization* called "L1"? And what is being "regularized"?

Solution: It is called L1 because it adds a penalty term equal to the L1 norm (the sum of absolute values) of the model coefficients to the loss function. The loss function (you can also say objective) is being regularized [in order to discourage large weights and help prevent overfitting].

(b) (2 points) What does the term *filter method* refer to in feature selection? What is being "filtered"?

Solution: A filter method is a feature selection technique that ranks features based on their individual relationship with the target variable, independently of the model. Features that are weakly related to the target are filtered out.

5. Suppose we are evaluating a binary classifier that predicts all points **above** a horizontal line as belonging to the **positive class** (blue circles), and all points **below** the line as the **negative class** (red crosses). The classifier is a horizontal line that **can only move up or down** – it cannot rotate or tilt. The figure below shows six labeled points in $\mathbb{R}^2$, with the initial position of the classifier indicated as a dashed line.



(a) (12 points) As the horizontal line moves up, it changes the classification of points. Draw the Receiver Operating Characteristic (ROC) curve. Provide a completed table of ROC coordinates used to draw the curve.

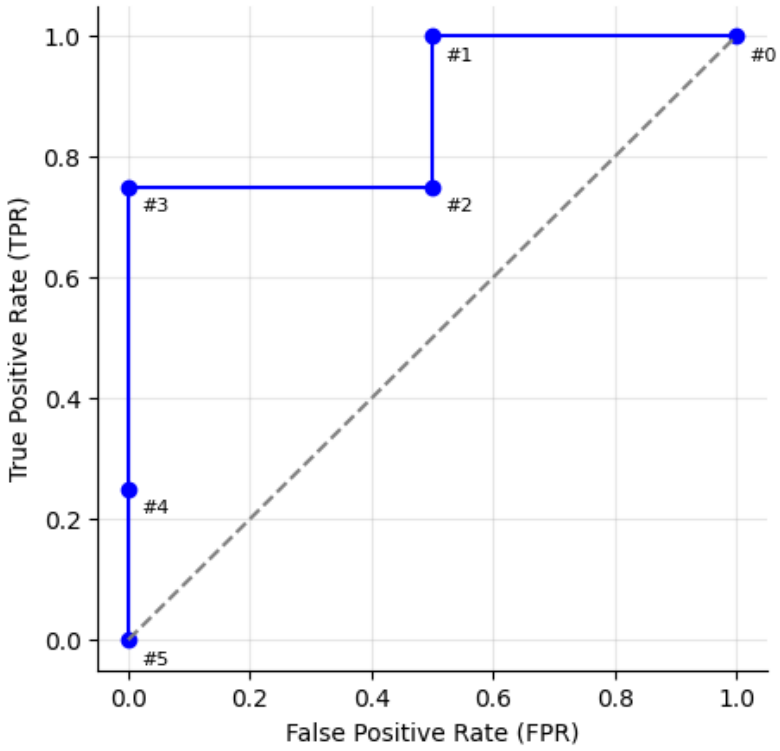
(b) (2 points) What is the highest precision that the classifier achieves on this dataset?

(c) (2 points) Which point(s) *on the ROC curve* should we select if minimizing false positives is a primary concern?

Solution:

(a) We compute TPR and FPR as the horizontal threshold moves upward and more points are predicted negative.

# of position	FPR	TPR
0 (initial position)	1	1
1	$\frac{1}{2}$	1
2	$\frac{1}{2}$	$\frac{3}{4}$
3	0	$\frac{3}{4}$
4	0	$\frac{1}{4}$
5	0	0



(b) The highest precision is for positions #3, #4, #5, and it is 100%.

(c) Since we care a lot about false positives, we want to choose a point with FPR = 0. Among the points with FPR = 0, we select the one with the highest TPR. Thus, we choose the point #3 $(0, 3/4)$.

6. (4 points) You are building a Naive Bayes model to classify news articles as related to "sports" or "technology." In your training data, the word "touchdown" appears multiple times in sports articles, but never in technology articles.

   Now you receive a new article that contains the word "touchdown." You use your trained model *without Laplace smoothing* to predict the category of this article.

   What probability will the model assign to the "technology" class? Justify your answer.

   Solution: The model will assign a probability of 0 to the "technology" class.

   Without Laplace smoothing, any word that has zero frequency in a given class will result in a zero conditional probability. Since Naive Bayes multiplies conditional probabilities, this zero makes the entire probability for that class zero.

7. Consider the dataset shown below, consisting of 10 examples with a single real-valued feature x and binary class label $y \in \{0, 1\}$:

x	y
1	1
2	1
3	0
4	1
5	1
6	0
7	0
8	0
9	0
10	0

We apply a Majority Classifier, which always predicts the class label that appears most frequently in the training set. In the case of a tie, the classifier defaults to predicting class 0.

(a) (2 points) What is the training set accuracy of the Majority Classifier?

(b) (5 points) What is the average Leave-One-Out Cross-Validation accuracy?

(c) (5 points) What is the average 2-Fold Cross-Validation accuracy? Assume the dataset is split into two equal halves according to increasing order of x.

Solution:

(a) Since class 0 is the majority, the Majority Classifier always predicts 0.

The training set accuracy is:

$$\text{Accuracy} = \frac{6}{10} = 0.6 \quad \text{or} \quad 60\%.$$

(b) For Leave-One-Out Cross-Validation (LOOCV), we train on 9 examples and test on the 1 left-out example, repeating this for each example. The Majority Classifier is recomputed each time.

Proceed case by case:

- For each point with $y = 0$: 4 ones, 5 zeros $\rightarrow$ predict 0 $\rightarrow$ correct
- For each point with $y = 1$: 3 ones, 6 zeros $\rightarrow$ predict 0 $\rightarrow$ wrong

Summary:

- Correct predictions: 6
- Incorrect predictions: 4

Thus, the LOOCV accuracy is:

$$\text{Accuracy} = \frac{6}{10} = 0.6 \quad \text{or} \quad 60\%.$$

(c) For 2-Fold Cross-Validation:

Split the data into two folds based on x values:

- Fold 1: $x = 1, 2, 3, 4, 5$
- Fold 2: $x = 6, 7, 8, 9, 10$

First, train on Fold 2, test on Fold 1:

- Majority class = 0 $\rightarrow$ predict 0 for all Fold 1 examples.
- 1 correct out of 5

Second, train on Fold 1, test on Fold 2:

- Fold 1 has 4 ones and 1 zero.
- Majority class = 1 $\rightarrow$ predict 1 for all Fold 2 examples.
- 0 correct out of 5

Thus:

$$\text{Total correct} = 1 \quad \text{out of} \quad 10.$$
$$\text{Accuracy} = \frac{1}{10} = 0.1 \quad \text{or} \quad 10\%.$$

8. For each of the following True/False questions, please **answer and provide a brief explanation to justify your choice**. Answers without a correct explanation will not receive credit.

   (a) (2 points) One-hot encoding is for numeric data types.

   A. True

   B. False

   Solution: False. One-hot encoding is for categorical (non-numeric) variables. It transforms each category into a separate binary feature rather than working directly with numeric values.

   (b) (2 points) A good baseline model for the classification task that involves many output classes is to always predict the output class that has the least training examples.

   A. True

   B. False

   Solution: False. A good baseline would be to predict the most frequent class (majority class baseline), not the rarest (least frequent). Predicting the least frequent class would result in very low accuracy and would not serve as a meaningful baseline.

   (c) (2 points) For a binary classification task, false negatives and false positives might have different degrees of practical significance depending on the circumstances.

   A. True

   B. False

   Solution: True. Depending on the application, false positives and false negatives may carry different costs. For example, in medical diagnosis, a false negative could delay treatment, while in fraud detection, a false positive might inconvenience a customer.

   (d) (2 points) Decreasing the regularization term prevents a model from overfitting.

   A. True

B. False

Solution: False. Decreasing the regularization term actually makes overfitting more likely, because the model becomes less penalized for complexity and can fit the training data too closely. Regularization is used to discourage overly complex models.

(e) (2 points) In `scikit-learn`, all features must be converted into numerical format before being used in a `NaiveBayesClassifier`, even if they are all categorical (including the target variable).

A. True

B. False

Solution: True. `scikit-learn` requires both input features and the target variable to be numeric. Categorical features must be encoded (e.g., using one-hot encoding or ordinal encoding) before being passed to a `NaiveBayesClassifier`.

9. Four different classification models (M1–M4) were trained and evaluated using the same training, validation, and test datasets, but each with a different set of features. The results are shown below:

Model	Train accuracy	Validation accuracy	Test accuracy
M1	0.95	0.40	0.40
M2	0.83	0.70	0.79
M3	0.75	0.73	0.62
M4	0.45	0.36	0.36

What are the correct assessments of the models? **Circle or underline the most appropriate option(s) in each blank.** Provide explanations where prompted (i.e., where the word "because" appears).

(a) (4 points) M1 *underfits / is a normal fit / overfits* because _____

Solution: M1 *overfits* because *it has high training accuracy but very low validation and test accuracy.*

(b) (2 points) If we select M2 during model development as the best model among the four, it is because it has high *train / validation / test* accuracy.

Solution: *validation* accuracy.

(c) (4 points) M4 *underfits / is a normal fit / overfits* because _____

Solution: M4 *underfits* because *it has very low accuracy on train, validation, and test sets.*

10. Each question may have **only one** correct answer. *No need for explanations here.*

    (a) (2 points) Which of statement about decision trees is correct?

    A. The more levels a tree has, the higher the chances of underfitting.

    B. A decision tree cannot predict labels for unseen data samples.

    C. A decision tree uses a greedy algorithm to build the tree.

    D. A decision tree always achieves perfect accuracy on the training data.

    (b) (2 points) What of the following describes a classification task?

    A. Estimating the age of a tree based on its height and diameter.

    B. Predicting the cost of living of cities around the world.

    C. Clustering news articles into different categories without any labelled tags.

    D. Diagnosing whether someone has a Covid-19 or not.

    (c) (2 points) Given a particular model, which of the following most likely **will not** help resolve an overfitting issue?

    A. Decreasing the number of training examples

    B. Decreasing the number of features

    C. Using a simpler model

    D. Applying regularization to the model

x	y	$-\frac{x}{y}\log\frac{x}{y}$	x	y	$-\frac{x}{y}\log\frac{x}{y}$
1	2	0.50	1	8	0.38
1	3	0.53	3	8	0.53
2	3	0.39	5	8	0.42
1	4	0.50	7	8	0.17
3	4	0.31	1	9	0.35
1	5	0.46	2	9	0.48
2	5	0.53	4	9	0.52
3	5	0.44	5	9	0.47
4	5	0.26	7	9	0.28
1	6	0.43	8	9	0.15
2	6	0.53	1	10	0.33
5	6	0.22	3	10	0.52
1	7	0.40	7	10	0.36
2	7	0.52	9	10	0.14
3	7	0.52			
4	7	0.46			
5	7	0.35			
6	7	0.19			

Table 1: Values for x, y, and $-\frac{x}{y}\log\frac{x}{y}$ to calculate the entropy.

This page is intentionally left blank to accommodate work that would not fit elsewhere and/or scratch work.