4780/6780 Fundamentals of Data Science
Mockup Final

Instructor: Kiril Kuzmin

04/14/2025

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.

____

The mockup final contains 22 pages (including this cover page) and 13 problems. Total of points is 100.

Good luck and productive work!

Distribution of grades

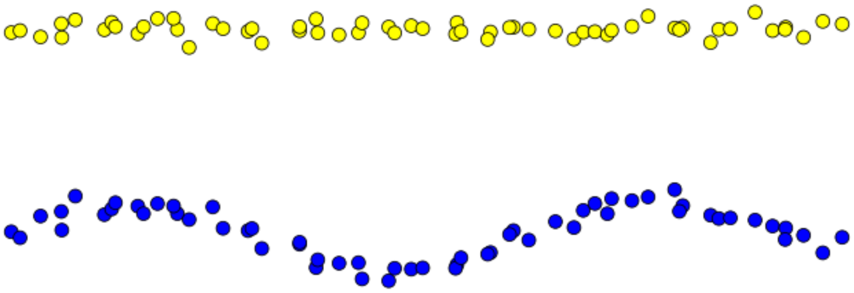
Question	Points	Score
1	8	
2	10	
3	8	
4	6	
5	8	
6	12	
7	6	
8	6	
9	12	
10	8	
11	6	
12	4	
13	6	
Total:	100	

1. Which clustering method(s) is most likely to produce the clustering shown in the figure below when $k = 2$? Choose the most appropriate method(s) from the list and briefly explain why it/they work better than the others (in at most three sentences).

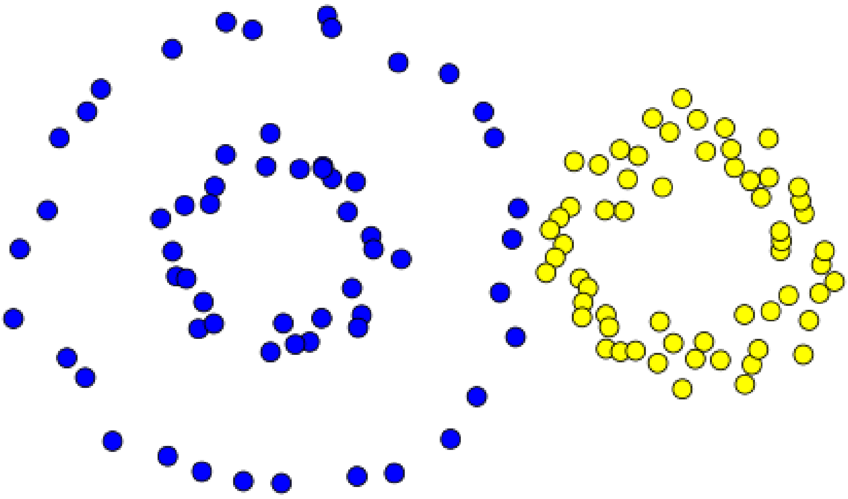
   Clustering methods to choose from:

   - Hierarchical clustering with single link
   - Hierarchical clustering with complete link
   - Hierarchical clustering with average link
   - k means clustering

   (a) (4 points) For the dataset shown below:



   (b) (4 points) For the dataset shown below:



Solution:

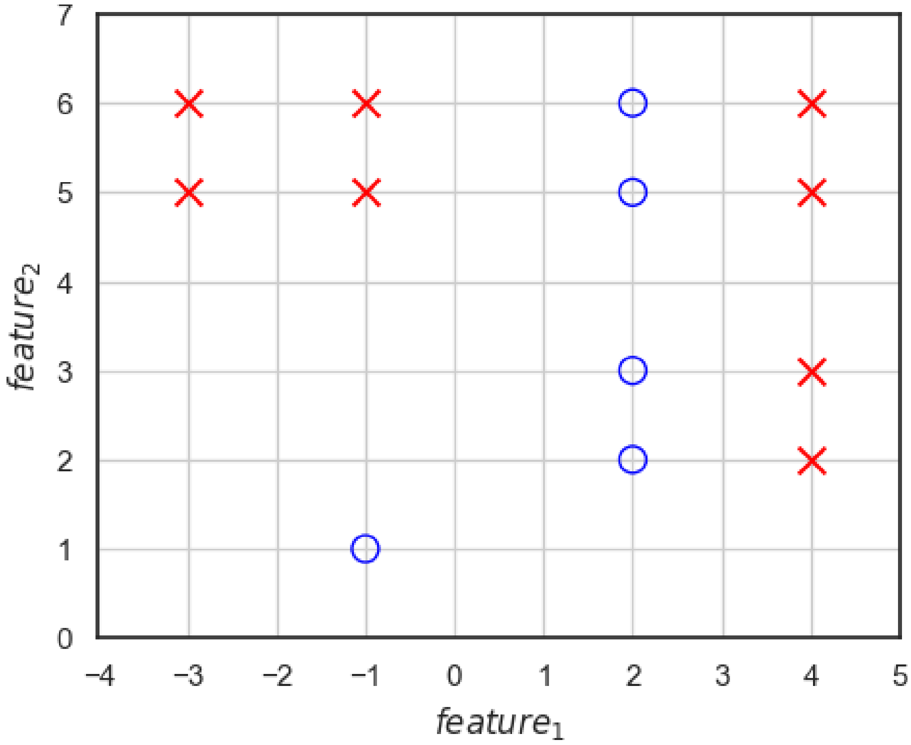
(a) **Selected method:** Hierarchical clustering with single link

Single linkage is well-suited to this structure because it can chain together closely connected points, even in non-convex shapes. Other linkage methods like complete or average may incorrectly merge points across true clusters due to larger inter-cluster distances. k means clustering fails because it assumes convex, spherical clusters.

(b) **Selected method:** k means

k means clustering works well here since the clusters are compact, convex, and well enough separated. Hierarchical methods, especially early in the process, may incorrectly group nearby points from opposite sides of the true boundary due to how distances between intermediate clusters are computed.

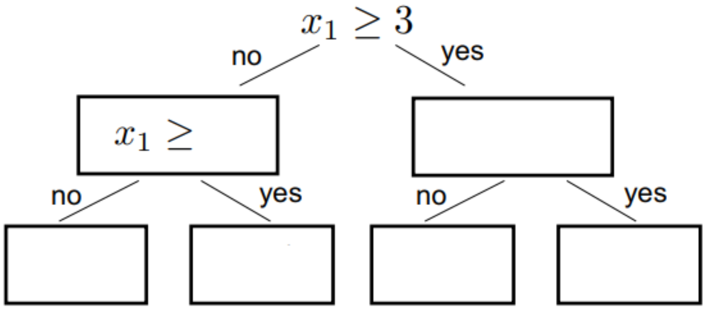
2. We seek to learn a classifier on the dataset shown below, consisting of 13 data points labeled as red crosses (class -1) and blue circles (class $+1$):



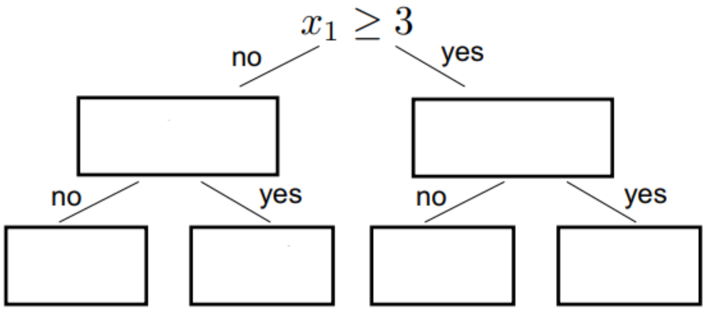
(a) (2 points) We first train a linear SVM model on this dataset. Will it achieve perfect training accuracy? Briefly explain.

(b) (6 points) We now train a depth-2 decision tree on this dataset using the **entropy** criterion, with `min_samples_split = 2`. The partially completed tree is shown below, where the root node splits on `feature1 >= 3`, and the second-level splits are also on `feature1`.

Complete the tree by filling in the missing split conditions and class labels at the leaves. Leave any unused boxes empty.



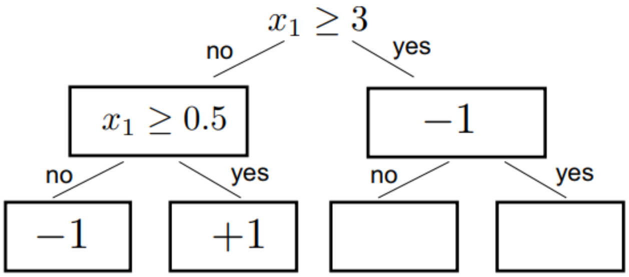
(c) (2 points) Now suppose we set `min_samples_split = 10`. Again, complete the tree shown above under this constraint. Fill in the leaf nodes where applicable, and leave other boxes empty if a split does not occur.



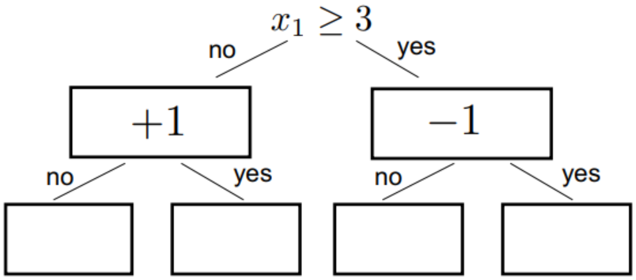
Solution:

(a) No, the dataset is not linearly separable.

(b) The completed decision tree using `min_samples_split = 2` and the entropy criterion is shown below:



(c) The completed decision tree using `min_samples_split = 10` is shown below. Some branches remain unexpanded due to the minimum split requirement:



3. You are training a logistic regression model and observe that it performs poorly on the test data.

   (a) (4 points) Could this poor performance be due to *underfitting*? Justify your answer.

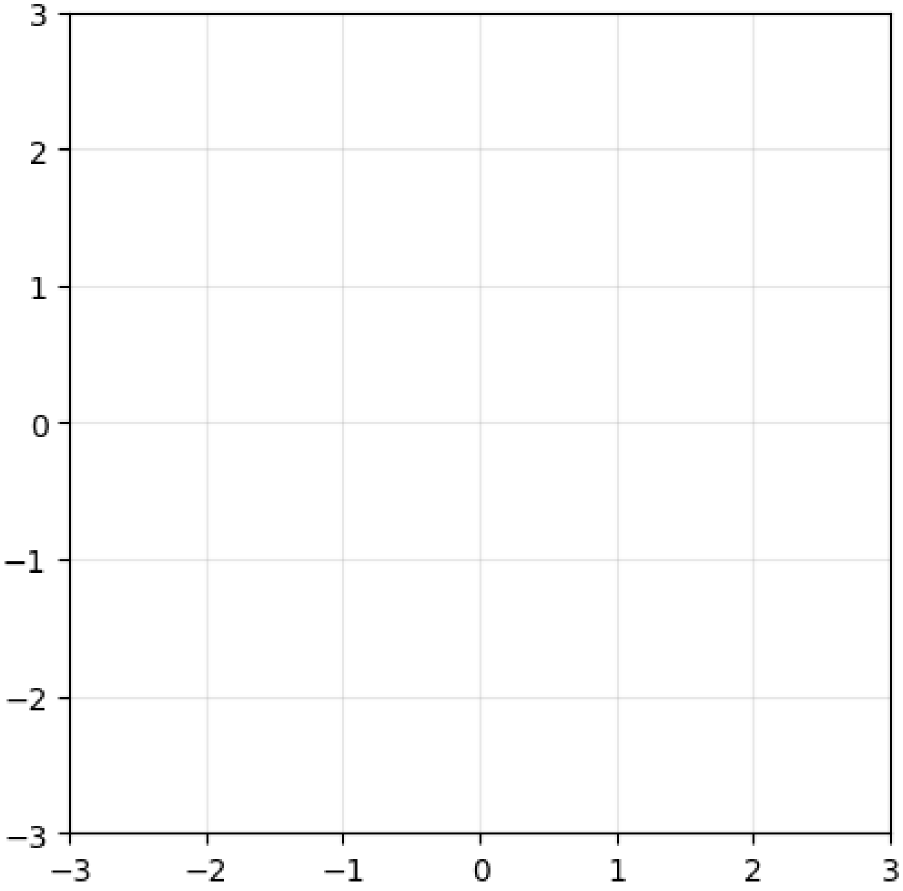
   Solution: Yes, logistic regression separates the two classes using only a plane (linear object!). This might be too simple to capture the variations in the data. Consider, for example, the case of two classes that are separable by a curved boundary. Logistic regression would be unable to model this non-linear relationship, leading to underfitting.

   (b) (4 points) Could this poor performance be due to *overfitting*? Justify your answer.

   Solution: Yes, if there are too many features, the data could appear to be linearly separable as a mathematical artifact. This could result in overfitting the training data, where the model fits noise or irrelevant patterns instead of generalizing well to unseen data.

4. (6 points) Draw a dataset consisting of 4 unique points in $\mathbb{R}^2$, each with *integer* coordinates. Two of the points must have positive labels (drawn as "×") and 2 of the points have negative labels (drawn as "○") such that the 1-nearest neighbor decision boundary is the same as an optimal decision tree boundary but different than the optimal logistic regression boundary.

   Also, draw the resulting decision tree/1-NN boundary.

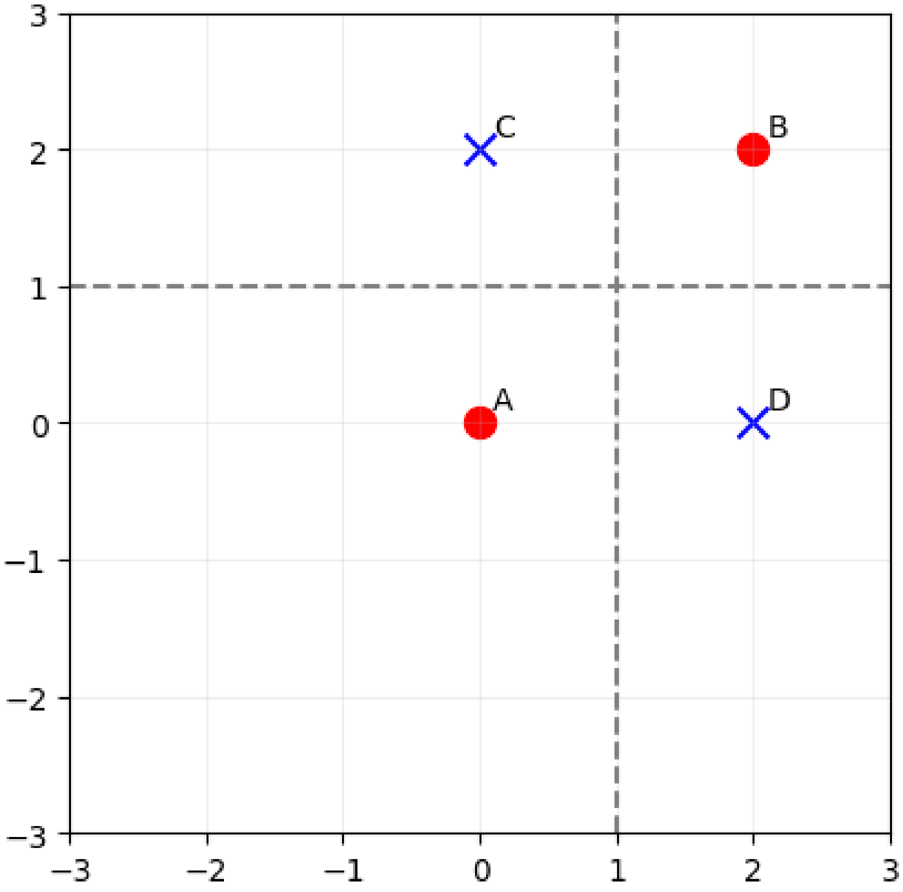


Solution: One dataset that would work is:

- A = (0, 0), label 0 (○)
- B = (2, 2), label 0 (○)
- C = (0, 2), label 1 (×)
- D = (2, 0), label 1 (×)

Both 1-NN and an axis-aligned decision tree can split the space at $x = 1$ and $y = 1$ to perfectly separate the classes.

However, logistic regression will attempt to find a single linear boundary – such as a diagonal – which cannot cleanly separate this dataset. It will misclassify some points or regions.



5. For a probabilistic classifier in a binary classification setting, consider the decision rule that predicts class 1 if $P(y = 1 \mid x) > t$, and predicts class 0 otherwise. This rule depends on a threshold $t \in [0, 1]$.

   Suppose the threshold t is increased.

   (a) (4 points) Does the **precision** tend to increase, decrease, or stay the same? Explain your answer briefly.

   Solution: As t increases, the classifier becomes more conservative in predicting class 1. Fewer predictions are made for class 1, and the ones that are made are more likely to be correct. Therefore, precision tends to **increase**.

   (b) (4 points) Does the **recall** tend to increase, decrease, or stay the same? Explain your answer briefly.

   Solution: As t increases, the classifier predicts class 1 less often, so it misses more actual positives. Therefore, recall tends to **decrease**.

6. (12 points) Suppose we are evaluating a binary classifier using an ROC curve. To construct the ROC curve, we compute the True Positive Rate (TPR) and False Positive Rate (FPR) at various classification thresholds t. The table below shows the TPR and FPR values at different thresholds, with a few entries left as variables A, B, C, D.

   Use the labeled dataset and predicted probabilities provided below to compute the missing values.

   Note: Multiple letters may correspond to the same numerical value.

   ROC curve table:

Threshold t	TPR	FPR
0.95	0	0
0.90	$\frac{1}{4}$	0
0.85	$\frac{1}{2}$	0
0.75	A	B
0.70	C	D
0.65	1	0.5
0.45	1	1

   Dataset with predicted probabilities:

Data point #	x_1	x_2	True label y	Predicted probability
1	1.00	1.00	1	0.65
2	0.50	2.50	1	0.90
3	0.75	0.50	1	0.75
4	0.00	0.00	1	0.85
5	0.50	1.50	0	0.70
6	1.00	0.00	0	0.45

   Compute the values of A, B, C, D and explain briefly how you obtained them.

Solution: We are given 6 data points:

- Positives (label 1): points 1, 2, 3, 4;

- Negatives (label 0): points 5, 6.

We compute predictions at various thresholds t, marking a point as positive if its predicted probability is $\geq t$.

At $T = 0.75$:

- Predicted positive: Points 2 (0.90), 4 (0.85), 3 (0.75)
- Predicted negative: Points 1 (0.65), 5 (0.70), 6 (0.45)

From these:

- TP = points 2, 4, 3 $\Rightarrow$ 3 true positives
- FP = 0 (none of the predicted positives are negatives)

Therefore,

$$A = \frac{3}{4}, \quad B = \frac{0}{2} = 0$$

At $T = 0.70$:

- Predicted positive: Points 2 (0.90), 4 (0.85), 3 (0.75), 5 (0.70)
- Predicted negative: Points 1 (0.65), 6 (0.45)

From these:

- TP = points 2, 4, 3 $\Rightarrow$ 3 true positives
- FP = point 5 $\Rightarrow$ 1 false positive

Therefore,

$$C = \frac{3}{4}, \quad D = \frac{1}{2}$$

Answer. $A = \frac{3}{4}, B = 0, C = \frac{3}{4}, D = \frac{1}{2}$

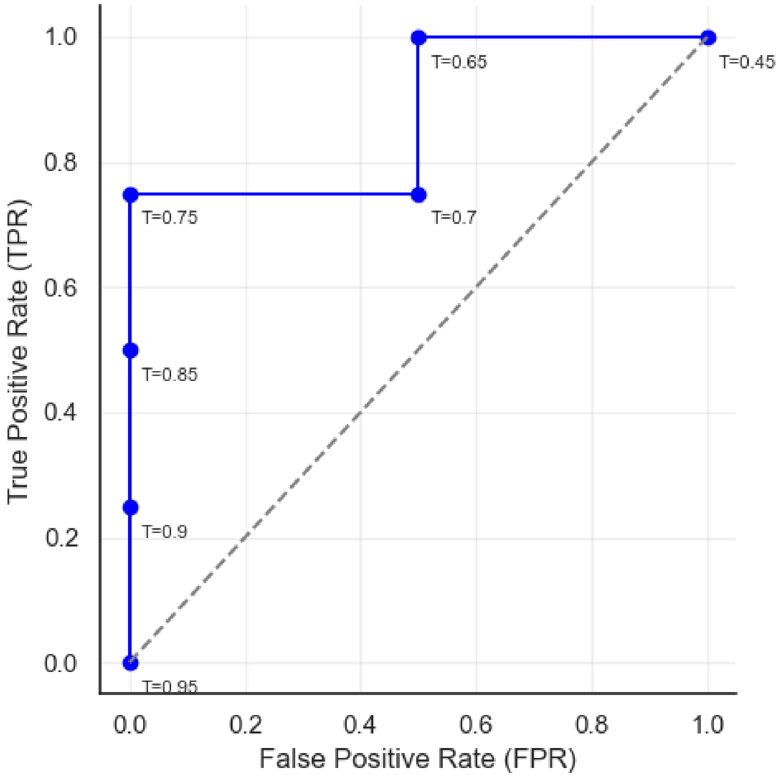
7. For the ROC curve described in the previous problem:

   (a) (4 points) Draw the ROC curve. Label both axes.

   (b) (2 points) Calculate the Area Under the Curve (AUC).

Solution:

(a) The ROC curve is plotted below using the given FPR and TPR values at various thresholds. Each point corresponds to a threshold, and the curve is formed by connecting these points. The diagonal line represents the performance of a random classifier.



(b) Area Under the Curve (AUC) is:

$$\text{AUC} = 1 - 0.5 \cdot 0.25 = 1 - 0.125 = 0.875.$$

8. (6 points) Consider training and predicting with a Naive Bayes classifier for two document classes. Assume we are not using any *pseudocounts*[1].

   The word "Atlanta" appears **once** in documents from class 1, and **never** in documents from class 0.

   Suppose we encounter a new document that contains the word "Atlanta".

   What is the posterior probability that the document belongs to class 1?

Solution: Without pseudocounts, the probability of a word that was never seen in a class becomes exactly zero. In this case:

- $P(\text{“Atlanta”} \mid \text{class } 1) > 0$ (since it appeared once)
- $P(\text{“Atlanta”} \mid \text{class } 0) = 0$ (since it never appeared)

During prediction, Naive Bayes computes the posterior probability using:

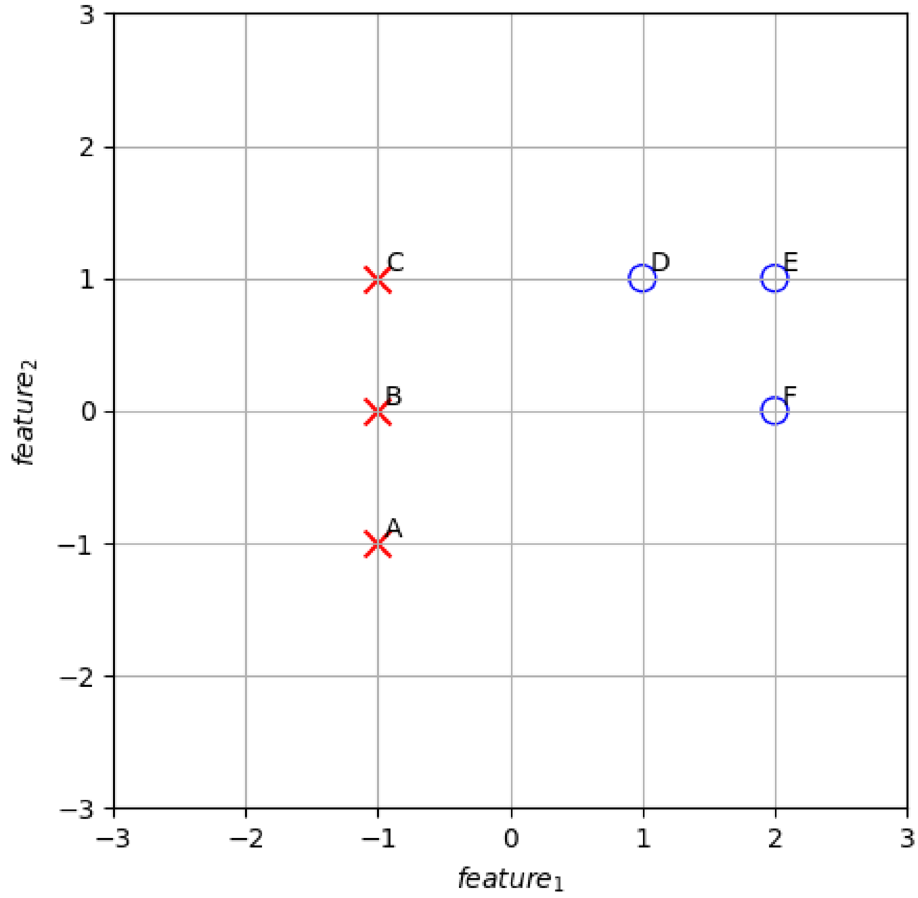
$$P(\text{class } i \mid \text{“Atlanta”}) \propto P(\text{class } i) \cdot P(\text{“Atlanta”} \mid \text{class } i)$$

It means that $P(\text{class } 0 \mid \text{“Atlanta”}) = 0$. Thus, $P(\text{class } 1 \mid \text{“Atlanta”}) = 1$.

Answer. $P(\text{class } 1 \mid \text{“Atlanta”}) = 1$.

[1] **Pseudocounts** are small artificial counts (usually 1) added to observed word frequencies to avoid zero probabilities when a word does not appear in a given class. This technique is also called *Laplace smoothing.*

9. (12 points) Consider the dataset shown below, consisting of 6 points in $\mathbb{R}^2$:



We apply the k-Nearest Neighbors (KNN) model with Euclidean distance to perform classification. We will evaluate the model using *leave-one-out cross-validation* (LOOCV). In the event of a tie in the majority vote, the classifier defaults to predicting the *negative* class – blue circles.

For each value $k \in \{3,\ 4,\ 5\}$, compute the LOOCV accuracy. Which value of k yields the highest average accuracy for this dataset?

Solution: **Case $k = 3$.** Each point is classified using its 3 nearest neighbors.

- All 3 red points (A, B, C) have the following 3 nearest neighbors: 2 red and 1 blue ⇒ points A, B, C are correctly classified.
- All 3 blue points (D, E, F) have the following 3 nearest neighbors: 2 blue and 1 red ⇒ points D, E, F are correctly classified.

$$\text{Correct predictions: } 6 \quad \Rightarrow \quad \text{Average accuracy: } \frac{6}{6} = \boxed{1.00}$$

Case $k = 4$**.** Each point is classified using its 4 nearest neighbors.

- Red points: each has 2 red and 2 blue neighbors ⇒ tie ⇒ predicted as negative ⇒ misclassified.
- Blue points: each has 2 red and 2 blue neighbors ⇒ tie ⇒ predicted as negative ⇒ correctly classified.

$$\text{Correct predictions: } 3 \quad \Rightarrow \quad \text{Average accuracy: } \frac{3}{6} = \boxed{0.50}$$

Case $k = 5$**.** Each point is classified using all 5 other points.

- Each point sees 3 from the opposite class and 2 from its own.
- All predictions follow the majority vote ⇒ all points misclassified.

$$\text{Correct predictions: } 0 \quad \Rightarrow \quad \text{Average accuracy: } \frac{0}{6} = \boxed{0.00}$$

Answer. The best LOOCV average accuracy is achieved when $k = 3$.

10. For each of the following True/False questions, please provide a brief explanation following your answer.

   (a) (2 points) If we take any linearly separable dataset and add a new feature, it is still guaranteed to remain linearly separable.

      A. True

      B. False

   Solution: True. We can assign a weight of $\theta_{d+1} = 0$ to the new feature in the separating hyperplane. This effectively ignores the added feature and preserves the original decision boundary. Thus, the data remains linearly separable.

   (b) (2 points) If we take any linearly separable dataset and remove a feature, it is still guaranteed to remain linearly separable.

      A. True

      B. False

   Solution: False. Removing a feature may eliminate the information necessary to maintain separability.

   For example, suppose the data is only separable because of one specific feature. If that feature is removed, the resulting dataset could contain points with identical features but different labels, making it impossible to separate them with a linear decision boundary.

   Therefore, linearly separable data may become non-separable after dropping a feature.

   (c) (2 points) If we take any dataset that is not linearly separable and remove a feature, it is still guaranteed to not be linearly separable.

      A. True

      B. False

   Solution: True. Removing a feature is equivalent to setting its corresponding weight to zero in the original model.

   If the dataset was not linearly separable to begin with, then no setting of weights – including one where the removed feature has weight zero

– can make it separable. Therefore, removing a feature cannot make a non-separable dataset linearly separable.

(d) (2 points) If we take any dataset that is not linearly separable and remove a *data point*, it is still guaranteed to not be linearly separable.

A. True

B. False

Solution: False. Removing a data point can change the structure of the dataset and potentially make it linearly separable.

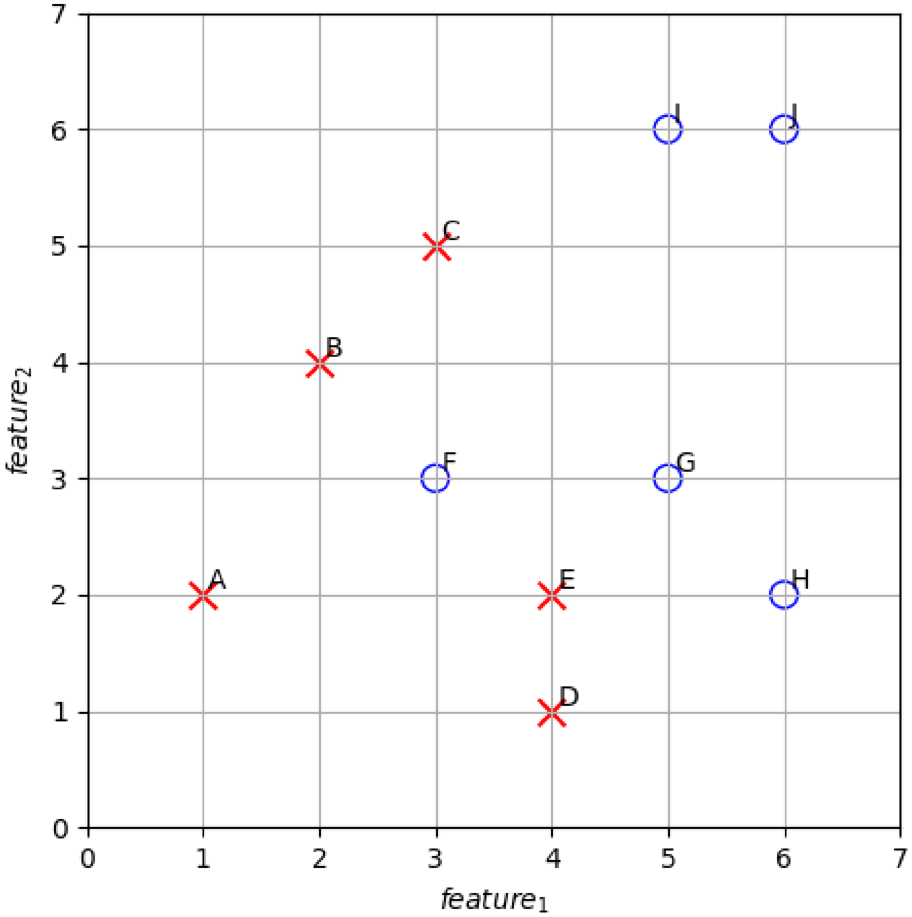
For example, consider the following dataset:

- $(1, 1)$: label $+1$
- $(1, -1)$: label -1
- $(-1, -1)$: label $+1$
- $(-1, 1)$: label -1

This dataset is not linearly separable (it's the classic XOR pattern). However, if we remove the point $(1, 1)$, the remaining three points become linearly separable.

Therefore, it is not guaranteed that removing a data point preserves non-separability.

11. Consider the 2D dataset shown in the plot. Your task is to draw a linear decision boundary (hyperplane) that achieves the highest possible classification accuracy on the dataset.



(a) (4 points) Sketch a straight-line decision boundary that separates the two classes with the best accuracy. Clearly indicate which side of the line corresponds to each class.

(b) (2 points) What is the resulting classification accuracy of your boundary?

Solution:

A good linear decision boundary can be drawn approximately along the line:

$$feature_1 = 4.5.$$

This line roughly separates the red crosses and blue circles with just one misclassification. Out of 10 total points, 9 are classified correctly and only point F is misclassified.

$$\text{Accuracy} = \frac{9}{10} = 90\%.$$

12. Consider learning a classifier on a dataset with 1000 features in total. The features are distributed as follows:

- 50 features are truly informative about the class,
- 50 features are exact copies (duplicates) of the informative features,
- The remaining 900 features are not informative.

Assume there is enough training data to reliably assess the usefulness of features, and that the feature selection methods apply appropriate thresholds.

(a) (2 points) How many features would be selected by a mutual information filter?

(b) (2 points) How many features would be selected by a wrapper based method?

Solution:

(a) A mutual information filter ranks features based on their individual relevance to the target. Since it cannot distinguish between duplicated features, it would assign high scores to both the 50 original informative features and their 50 exact copies. Therefore, it would select 100 features.

(b) A wrapper method evaluates feature subsets based on model performance. Since duplicated features provide redundant information, a good wrapper would include only one copy per informative feature. Therefore, it would select 50 features.

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

    (a) (2 points) ______________ is defined as the percentage of correct predictions out of all the observations.

    A. Confusion matrix

    B. Accuracy

    C. Precision

    D. Recall

    E. F_1 score

    (b) (2 points) You train a classifier on 10,000 training points and obtain a training accuracy of 99%. However, when you submit to Kaggle, your accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your performance on Kaggle?

    A. Set your regularization value (λ) to 0

    B. Use validation to tune your hyperparameters

    C. Train on less data

    D. Train your model more so that it would archive a training accuracy of 100%

    E. None of the above

    (c) (2 points) Which of the following is true of support vector machines with a linear kernel?

    A. Increasing the hyperparameter C tends to decrease the sensitivity to outliers.

    B. The default value for the hyperparameter C is scikit-learn is 0

    C. It can separate XOR data

    D. Increasing the hyperparameter C tends to increase the margin

    E. Increasing the hyperparameter C tends to decrease the margin