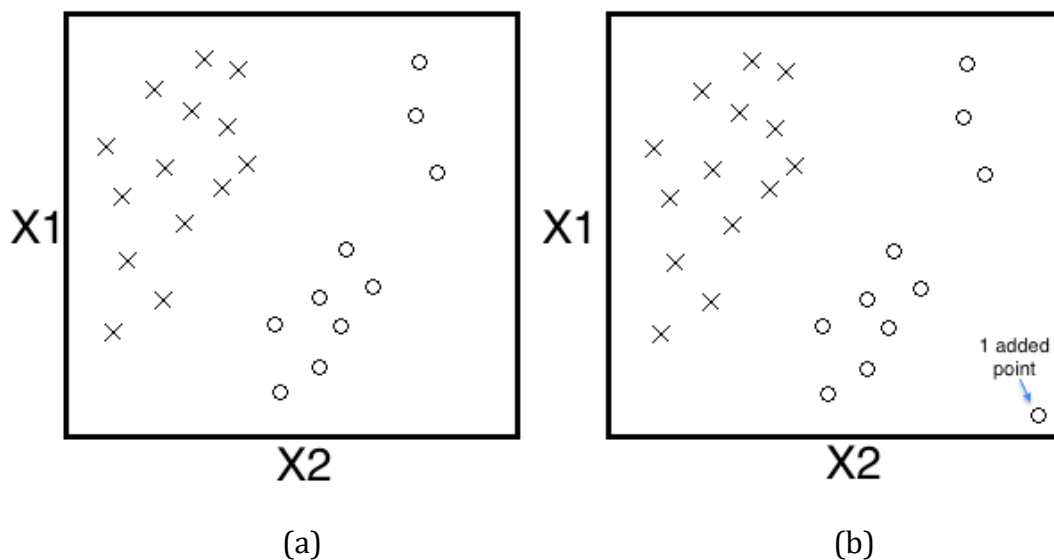


EE4146 Assignment 2024-25: Due day Week 9, March 22 Saturday before 11:50pm (submit to Canvas in pdf format). Please check Canvas.

We absolutely do not ACCEPT late submission after canvas window closed via email.

1. When a decision tree is grown to its full depth (bottom branches), it is more likely to fit the noise of the dataset.
A. True
B. False
C. Noise has no relation with leaves, question does not stand. **(10 marks)**
2. When the feature space is larger, over-fitting is more likely.
A. True
B. False
C. over-fitting has no relation with feature space **(10 marks)**
3. As the number of data goes to infinity, the trained model on the dataset will have
i A. Lower Variance B. Higher Variance C. Same Variance D. No relation as it is dataset dependent **(10 marks)**
ii A. Lower Bias B. Higher Bias C. Same Bias D. No relation as it is dataset dependent **(10 marks)**



4. (a) The above figure (a) shows a subset of 2 classes training dataset with 2 features X1, X2. Draw a decision boundary for an LDA classifier. Give less than 10 words to explain the most important feature of your proposed decision boundary. **(10 marks)**
b. If we add a new data point shown in Figure (b). How does the decision boundary be affected by the newly added data point? Answer by using less than 10 words. **(10 marks)**

5. Why is the computational complexity of KNN is often larger especially for high dimensional datasets. **(10 marks)**

6. Which of the following statements are true for k-NN classifiers (circle all answers that are correct). **(10 marks)**

1. The classification accuracy is better with larger values of k.
2. The decision boundary is smoother with smaller values of k.
3. k-NN is a type of instance-based learning.
4. k-NN does not require an explicit training step.
5. The decision boundary is linear.
6. The number of features of the dataset has no effect on the computational time.

7. Assume we need to cluster buildings around Kowloon without using their exact addresses. For each building, we map properties of the building to a numeric value. For instance, the building's location is mapped as Kowloon City = 0, MonKok = 1, Kowloon Tong = 2, etc., the exterior material is brick-look = 0, Concrete-look = 1, tile-look = 2, glass = 4, etc., the Building color is white = 0, green = 1, white = 2, etc. We have 50 such features so each building can be represented as a vector in $\mathbb{R}^{50}$. Which of the two clustering algorithms learned in class (hierarchical clustering, k-means) would be most appropriate for this task? Explain briefly the major point for such a selection.

(10 marks)

8. List 2 major disadvantages of using Naïve Bayes classifier. **(10 marks)**