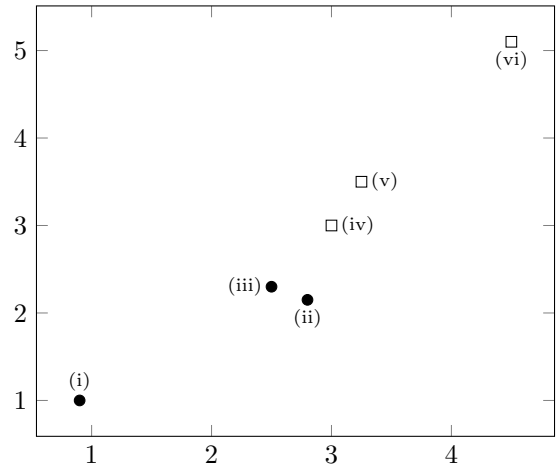
CSCI 467: Introduction to Machine Learning Spring 2022

Homework 1

Due: September 15th, 2025

Problem 1: KNN Classification and Cross Validation (30 points)

Consider the following dataset with two possible class labels (square and circle) assigned to points in the (x, y) plane. In this problem, we will consider the K-Nearest-Neighbor (KNN) classifier with the Euclidean distance, and use Leave-One-Out Cross Validation to choose the value of the hyperparameter K. Luckily, for KNN classification, Leave-One-Out is relatively simple since we do not need to re-learn a model for each setting of K. Clarification: For any distances that are ambiguous (where it is unclear which datapoint is closer), you can choose arbitrarily between the closest points and we will take any resulting answer.



(a) (7 points) Perform Leave-One-Out Cross Validation with $K = 1$. For each split of the dataset (i.e. using each training data point (i)-(vi) as a validation point), write down the prediction made by the KNN classifier. What is the classification error, averaged across splits?

Validation Point	Classification
i	
ii	
iii	
iv	
v	
vi	

Average Error:

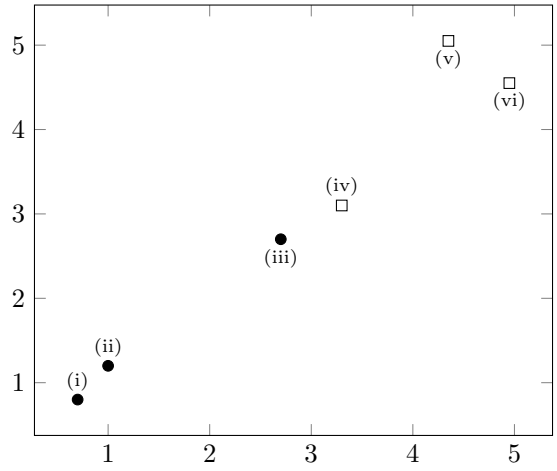
(b) (7 points) Perform Leave-One-Out Cross Validation with $K = 3$. For each split of the dataset (i.e. using each training data point (i)-(vi) as a validation point), write down the prediction made by the KNN classifier. What is the classification error, averaged across splits?

Validation Point	Classification
i	
ii	
iii	
iv	
v	
vi	

Average Error:

(c) (2 points) Our choice of hyperparameter K must be made *before* receiving test data. Based on the cross validation procedure above, which value of K should we use to evaluate unseen test data?

Now, consider the following test dataset. The goal of cross validation in parts (a)-(c) was to choose a hyperparameter that would yield good generalization performance on unseen data, but this procedure isn't always perfect!



(d) (7 points) For each test data point, write down the prediction made by the KNN classifier with $K = 1$. What is the average classification error?

Test Point	Classification
i	
ii	
iii	
iv	
v	
vi	

Average Error:

(e) (7 points) For each test data point, write down the prediction made by the KNN classifier with $K = 3$. What is the average classification error?

Test Point	Classification
i	
ii	
iii	
iv	
v	
vi	

Average Error:

Problem 2: Understanding Entropy (28 points)

The goal of this problem is to gain an intuitive understanding of entropy and why it is a good criteria for growing decision trees. Recall from lecture that the entropy of a discrete distribution P over C classes is defined as:

$$H(P) = -\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$$

(a) Consider a node where all samples belong to the same class i. We can write the probability distribution as follows,

$$P(Y = k) = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases}$$

We will prove that P is a minimum entropy distribution.

(i) (5 points) Prove that for any distribution P, the minimum possible entropy is 0. In other words, $H(P) \geq 0$.

(ii) (5 points) Show that a node where all samples belong to the same class achieves minimum entropy.

(iii) (4 points) Explain why this is desirable for a decision tree leaf.

(b) Next consider a node with an equal number of samples from all classes. Here the probability of each label is the same.

$$P(Y = k) = \frac{1}{C}$$

We'll prove that P is a maximum entropy distribution for the 2-class case ($C = 2$). Clarification: For the remainder of the problem, consider an arbitrary 2-class distribution P.

(i) (5 points) Define $P_1 = P(Y = 1)$ and $P_2 = P(Y = 2)$. Rewrite $H(P)$ as a function of just P_1.

(ii) (5 points) We want to find P such that $H(P)$ is maximized. In this case we can just find a local maximum of $H(P)$. Calculate the values of P_1 and P_2 that maximize entropy.

(iii) (4 points) Explain why nodes with high entropy are not desirable for decision trees.

Problem 3: Decision Tree Branching (20 points)

In the process of constructing a decision tree to predict a class label Y (either $Y = a$ or $Y = b$), you are faced with the choice to branch on one of two binary features, X_1 or X_2.

For feature $X_1 = 1$, there are 25 examples with $Y = a$ and 5 examples with $Y = b$, and for $X_1 = 0$, there are 0 examples with $Y = a$ and 5 examples with $Y = b$.

For feature $X_2 = 1$, there are 2 examples with $Y = a$ and 10 examples with $Y = b$ and for $X_2 = 0$, there are 20 examples with $Y = a$ and 3 examples with $Y = b$.

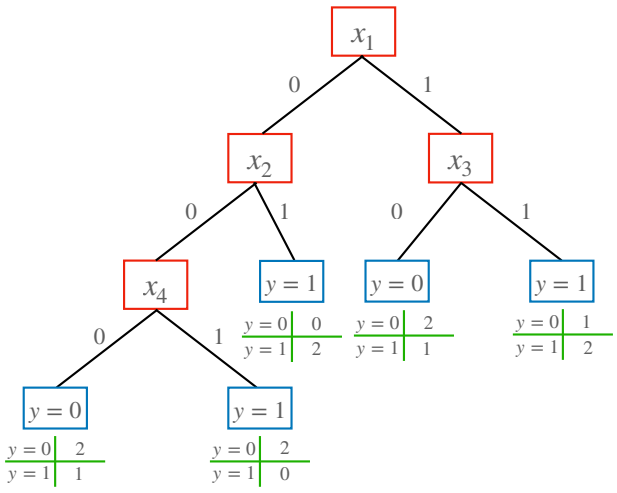
(a) (10 points) Calculate the conditional entropy for splitting on X_1 and splitting on X_2. According to this splitting strategy, on which feature should your decision tree branch?

(If possible, please calculate logarithms with base 2)

(b) (10 points) Calculate the Gini impurity for splitting on X_1 and splitting on X_2. According to this splitting strategy, on which feature should your decision tree branch?

Problem 4: Decision Tree Pruning (22 points)

Consider the following decision tree with features x_1, x_2, x_3 taking on values $\{0, 1\}$ (labelled in red) and predictions y at each leaf node (labelled in blue). There are 13 validation examples. The number of validation examples within each class are noted in the table below each leaf (labelled in green).



(a) (5 points) Calculate the classification error over the validation set.

(b) (5 points) Calculate the classification error if you were to prune the x_2 node.

(c) (5 points) Calculate the classification error if you were to prune the x_3 node.

(d) (5 points) Calculate the classification error if you were to prune the x_4 node.

(e) (2 points) Based on classification error, should this tree be pruned? If so, which node should be pruned?