CS 6140 Machine Learning - Assignment 1
(200 points)

1 Classification Trees With Numerical Features

For this problem you will be working with two datasets:

- **Iris**: has three classes and the task is to accurately predict one of the three sub-types of the Iris flower given four different physical features. These features include the length and width of the sepals and the petals. There are a total of 150 instances with each class having 50 instances.
- **Spambase**: is a binary classification task and the objective is to classify email messages as being spam or not. To this end the dataset uses fifty seven text based features to represent each email message. There are about 4600 instances

Since, both datasets have continuous features you will **implement** decision trees that have **binary splits**. For determining the optimal threshold for splitting you will need to search over all possible thresholds for a given feature(refer to class notes and discussion for an efficient search strategy). Use **information gain** to select the splitting features and values and measure node impurity using entropy in your implementation.

1.1 Growing Decision Trees (*50 points*)

Instead of growing full trees, you will use an early stopping strategy. To this end, we will impose a limit on the minimum number of instances at a leaf node, let this threshold be denoted as η_{min}, where $0 \leq \eta_{min} \leq 1$ is described as a ratio relative to the size of the training dataset. For example if the size of the training dataset is 150 and $\eta_{min} = 0.05$, then a node will only be split further if it has more than eight instances.

(a) For the Iris dataset use $\eta_{min} \in \{0.05, 0.10, 0.15, 0.20\}$, and calculate the accuracy using ten fold cross-validation for each value of η_{min}

(b) For the Spambase dataset use $\eta_{min} \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and calculate the accuracy using ten fold cross-validation for each value of η_{min}

You can summarize your results in two separate tables, one for each dataset (report the average accuracy and standard deviation across the folds).

1.2 Interpreting the results

(a) Select the best value of η_{min} for the Iris dataset, and create a class confusion matrix using ten-fold cross validation(*use only the test set for populating the confusion matrix*). How do you interpret the confusion matrix, and why?

(b) Select the best value of η_{min} for the Spambase dataset, and create a class confusion matrix using ten-fold cross validation(*use only the test set for populating the confusion matrix*). How do you interpret the confusion matrix, and why?

(c) How does different values of η_{min} impact classifier performance for both datasets and why? Support your claims/insights through your results.

(d) (*Optional Exercise - no points*) Compare the average (taken across the k-folds) training and test accuracy for different values of η_{min} for both datasets. You can use any graphing tool/API to plot the training and test accuracy against η_{min}. Do you see any evidence of overfitting or underfitting in either or both cases? Justify your conclusions by describing the probable cause(s).

2 Classification Trees With Categorical Features

In this problem you will be working with a new dataset:

- **Mushroom**: is binary classification dataset and the task is to accurately predict whether a mushroom is poisonous or edible given 21 different categorical (ordinal) features for each mushroom. These features describe various physical properties of the mushrooms such as length, diameter, etc. There are a total of 8124 instances.

As this dataset has only ordinal features we will **implement** both binary and multiway decision trees. Use **information gain** to select the splitting features and values and measure node impurity using entropy in your implementation.

2.1 Multiway vs Binary Decision Trees (***50 points***)

In this problem we will grow both multiway and binary decision trees using categorical features. Similar to the last problem we will be using an early stopping strategy to keep our decision tree from overfitting.

(a) Grow a multiway decision tree using $\eta_{min} \in \{0.05, 0.10, 0.15\}$, and calculate the accuracy using ten fold cross-validation for each value of η_{min}

(b) Replace each categorical feature $F \in \{f_1, f_2, \ldots, f_v\}$ with v binary features, corresponding to each distinct value of the feature. For example given the second feature *cap-surface*, that takes values {fibrous=f, grooves=g, scaly=y,smooth=s}, we will replace the original feature with four new boolean features: {*is cap-surface=f*}, {*is cap-surface=g*}, {*is cap-surface=y*} and {*is cap-surface=s*}. Only one of these features will be 1 for a given instance and the rest will be 0. Grow a binary decision tree with this new extended feature set, using $\eta_{min} \in \{5, 10, 15\}$, and calculate the accuracy using ten fold cross-validation for each value of η_{min}.

You can summarize your results in two separate tables, one for each dataset (report the average accuracy and standard deviation across the folds). Do you see a difference between the two approaches? Explain your answer briefly based on the results.

2.2 Interpreting the results

(a) Select the best value of η_{min} for the the above two cases i.e., multiway and binary splits, and create a class confusion matrix using ten-fold cross validation(*use only the test set for populating the confusion matrix*). How do you interpret the confusion matrix, and why?

(b) Is there a difference in the optimal value of η_{min} for mulitway vs binary splits? Please explain your finding using your results i.e., if there is a difference what are the probable causes? on the other hand if the optimal values are similar what does this tell you about binary vs multiway splitting in decision trees?

(c) (*Optional Exercise - no points*) Compare the average (taken across the k-folds) training and test accuracy for different values of η_{min} for both multiway and binary splits. You can use any graphing tool/API to plot the training and test accuracy against η_{min}. Do you see any evidence of overfitting or underfitting in either or both cases? Justify your conclusions by describing the probable cause(s).

3 Entropy (*20 points*)

Consider training a binary decision tree using entropy splits.

(a) Prove that the decrease in entropy by a split on a binary yes/no feature can never be greater than 1 bit.

(b) Generalize this result to the case of arbitrary multiway branching.

4 Gain and Impurity Measures (*20 points*)

Consider an arbitrary impurity measure $\iota(q)$, where q is a node in the decision tree and a feature V that takes $|V|$ distinct values. For a split resulting in $|V|$ children, the gain for the impurity measure is:

$$Gain(q, V) = \iota(q) - \sum_{i=1}^{|V|} \frac{N_i}{N_q} \iota(i)$$

where N_i is the number of instances in the i^{th} child node. Show that maximizing the $Gain(.,.)$ is equivalent to minimizing the impurity measure $\iota(.)$ over the $|V|$ children.

5 Gini Index (*20 points*)

Gini index another node impurity measure that is used to guide the splitting process for growing decision trees for classification. For a node q in the decision tree, Gini index is given as:

$$Gini(q) = \sum_{k=1}^{M} p_{qk}(1 - p_{qk})$$

for a classification problem with $M > 2$ classes. Show, that we can equivalently represent Gini index as

$$Gini(q) = \sum_{k \neq k'} p_{qk} p_{qk'}$$

6 Regression Trees (*40 points*)

In this problem you will **implement** regression trees using a new dataset:

- **Housing**: This is a regression dataset where the task is to predict the value of houses in the suburbs of Boston based on thirteen features that describe different aspects that are relevant to determining the value of a house, such as the number of rooms, levels of pollution in the area, etc.

(a) As this dataset has only numerical features we will be growing decision trees using only binary splits. Use the drop in **sum of squared errors (SSE)** to define the splits (please consult the regression tree notes posted in module 1 lesson 3). Use an early stopping strategy similar to the previous decision tree problems and use $\eta_{min} \in \{0.05, 0.10, 0.15, 0.20\}$. Calculate the SSE using ten fold cross-validation for each value of η_{min} and report the average and standard deviation across the folds (summarize your results in a table) (*30 points*).

(b) Does η_{min} impact the results significantly? Explain your answer i.e., if your results indicate that η_{min} has a significant impact on tree performance, what are the probable causes? (*10 points*).

(c) (*Optional Exercise - no points*) Compare the average (taken across the k-folds) training and test SSE for different values of η_{min}. You can use any graphing tool/API to plot the training and test error against η_{min}. Do you see any evidence of overfitting or underfitting? Justify your conclusions by describing the probable cause(s).