CS 422 Due:23h00
Data Mining Homework 2 March 23rd, 2025

Practicum Problems

These problems will primarily reference the lecture materials and examples provided in class using Python. It is recommended that a Jupyter/IPython notebook be used for the programmatic components. Students are expected to refer to the prescribed textbook or credible online resources to answer the questions accurately.

Problem 1

Load the Iris sample dataset from sklearn (using load_iris()) into Python with a Pandas DataFrame. Induce a set of binary decision trees with a minimum of 2 instances in the leaves (min_samples_leaf=2), no splits of subsets below 5 (min_samples_split=5), and a maximum tree depth ranging from 1 to 5 (max_depth=1 to 5). You can leave other parameters at their default values. Which depth values result in the highest Recall? Why? Which value resulted in the lowest Precision? Why? Which value results in the best F1 score? Also, explain the difference between the micro, macro, and weighted methods of score calculation

Depth 3, 4, and 5 achieve perfect model performance with the highest recall and F1 scores. Depth 1 results in the lowest precision due to an overly simplistic model. Micro focuses on overall performance, influenced by larger classes, and is suitable for balanced datasets. Macro treats each class equally, easily affected by smaller classes. Weighted averages by sample size, dominated by larger classes, and is suitable for imbalanced datasets.

Problem 2

Load the Breast Cancer Wisconsin (Diagnostic) sample dataset from the UCI Machine Learning Repository (the discrete version at: breast-cancer-wisconsin.data) into Python using a Pandas DataFrame. Induce a binary Decision Tree with a minimum of 2 instances in the leaves, no splits of subsets below 5, and a maximum tree depth of 2 (using the default Gini criterion). Calculate the Entropy, Gini, and Misclassification Error of the first split. What is the Information Gain? Which feature is selected for the first split, and what value determines the decision boundary?

A split was performed using the feature “Uniformity of Cell Size” with a threshold of 3.5. The entropy, Gini impurity, and misclassification rate all decreased significantly after the split, and the information gain was also high, indicating that this split was effective.

E.N.D

Problem 3

Load the Breast Cancer Wisconsin (Diagnostic) sample dataset from the UCI Machine Learning Repository (the continuous version at: wdbc.data) into Python using a Pandas DataFrame. Induce the same binary Decision Tree as above (now using the continuous data), but perform PCA dimensionality reduction beforehand. Using only the first principal component of the data for model fitting, what are the F1 score, Precision, and Recall of the PCA-based single factor model compared to the original (continuous) data? Repeat the process using the first and second principal components. Using the Confusion Matrix, what are the values for False Positives (FP) and True Positives (TP), as well as the False Positive Rate (FPR) and True Positive Rate (TPR)? Is using continuous data beneficial for the model in this case? How?"

The F1 scores of the models after PCA dimensionality reduction are all lower than when using the original continuous data. This indicates that PCA dimensionality reduction lost some information useful for model prediction in this case. After increasing to two principal components, the precision slightly improved, but the recall decreased, and the F1 score also decreased. In this specific case, using the original continuous data directly is better than first performing PCA dimensionality reduction and then modeling. Although PCA can reduce dimensionality and simplify the model, it led to information loss and reduced the overall performance of the model in this breast cancer diagnosis case. Using the original continuous data directly can achieve better prediction results.

E.N.D