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

Highest Recall: Tree with depth 3

Correctly identifies the most true positives without unnecessary complexity.

Lowest Precision: Tree with depth 1

Too simple (only one split), causing many false positives (e.g., mixing different classes).

Best F1 Score: Tree with depth 3

F1 balances Recall and Precision. Depth 3 achieves the optimal trade-off.

Micro-average: Mix all data and calculate metrics directly. Best for balanced datasets.

Macro-average: Calculate metrics per class and average. Best when all classes matter equally.

Weighted-average: Weight metrics by class size. Best for imbalanced data (e.g., one class dominates).

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

E.N.D

split. What is the Information Gain? Which feature is selected for the first split, and what value determines the decision boundary?

First split feature: The 30th feature (corresponding to the 31st column in the dataset) is selected with a decision threshold of 0.5000

Parent node metrics:

Entropy: 0.9518

Gini impurity: 0.4670

Misclassification error: 0.3715

Information gain: 0.9518

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?"

If we just use 1 principal component (PC1):

Lower accuracy: F1=0.867, detects 85.7% cancers (recall)

8 false alarms per 100 tests (FPR=7.9%)
Using 2 components (PC1+PC2):

Better: F1=0.902, detects 91.4% cancers

False alarms drop to 6.3%
original full data:

Best results: F1=0.921, detects 92.9% cancers

Only 4.8% false alarms
PCA reduces data size but loses some details.The original continuous data is more beneficial for this model.

E.N.D