PCSE 595 – Spring 2022 - Assignment 1 - Report

Mason Beckmeyer

Description:

In this experiment, I compared a K-Nearest Neighbors classifier to a Decision Tree classifier on their performance on the Wisconsin Breast Cancer Dataset. The results indicate that the KNN classifier performs the best, with a test set accuracy of 95%.

Methodology:

I implemented a k-nearest neighbors (KNN) classifier and a decision tree classifier. The k-nearest neighbors classifier works by storing all the training samples in memory and when classifying a sample the algorithm computes the distance between this sample and all the training samples using either Euclidean or Cosine distance. It then takes the closest k neighbors and makes a final prediction.

The decision tree classifier works by recursively splitting the training data on features that provide the highest information gain (largest reduction in entropy). I used information gain as a splitting criterion, featurized the data by finding all boundaries between classes for each feature, and used maximum tree depth to avoid overfitting.

Dataset Description:

The Wisconsin Breast Cancer Dataset consists of 699 samples with 9features each. The features were extracted from digitized images of a fine needle aspirate of a breast cell mass. These features include:

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

Each sample is labeled as benign (non-cancerous) or malignant (cancerous). There are 458 benign samples, and 241 malignant samples. The original dataset contained a few missing values. Missing values were replaced with a value of 5 which is the midpoint of possible feature values (features range in value from 1 to 10).

Experimental Details:

I withheld 20% of the data as a test set and used the remaining 80% for training. I used 5-fold cross-validation to tune classifier hyperparameters, and used average fold validation accuracy as my primary evaluation metric. The hyperparameters included k = 1->20 and euclidean or cosine distance for the KNN, and max depth 1->20 for the decision tree. Based on the results of cross validation, I found the best hyperparameters were k=3 and euclidean distance for the KNN classifier and max depth = 3 for the decision tree. Using these hyperparameter values, I trained each classifier on the full training set to generate the results shown below. I compare the classifiers against each other, and a majority class classifier as a baseline. The majority class classifier simply tags all samples as 0 (benign) which is the majority class.

Results and conclusion:
The KNN achieved a training accuracy of 95.5% and a test accuracy of 95%. The decision tree achieved a training accuracy of 94.5% and a test accuracy of 92.8%. Both classifiers outperformed the majority class baseline which achieved a test accuracy of 64%.

Based on these results, I can conclude that both classifiers were effective when compared against the majority class baseline, and that the KNN is the better of the two classifiers for this dataset.

Additional Questions:

1. Which feature (in the Wisconsin Breast Cancer Dataset) has the highest variance?
   Bare nuclei feature 5 has the highest with 13.002.

2. Which two features (in the Wisconsin Breast Cancer Dataset) has the highest covariance? What does that mean?
   Uniformity of cell size and Uniformity of cell shape have the highest covariance of 8.224. A high positive covariance between these two features means that these two features tend to increase or decrease together in a linear fashion.

3. Why is training accuracy higher (assuming it is) than test accuracy?
   Because the model can and will overfit to the training data causing it to reflect the small nuances contained within the training data that may not be replicated by the test data.

4. What was the average validation accuracy for each algorithm using optimum hyperparameters? Is it closer to the test accuracy than the training accuracy? Why?
   Using KNN k = 3 euclidean the average validation accuracy was 95.9% with the optimal hyperparameters. This is closer to the training accuracy of 95.5% this may be due to 20% of the test set not being sufficiently large enough to prevent random variations. Decision

tree average validation accuracy was 95.5%. Which is also closer to the training accuracy this may be due to the aforementioned reasons.