ECE/CS/ME 539 – Fall 2024 — Homework 4

1.

(4 points) Apply a kNN classifier to the `iris.csv` dataset (which is given to you). Use the file "knn.ipynb" to complete missing parts of the code.

(a) Perform stratified data partition at a 70/15/15 ratio to yield a training/validation/testing partition: $\mathbf{X_{train}}$, $\mathbf{y_{train}}$ (label), $\mathbf{X_{val}}$, $\mathbf{y_{val}}$, and $\mathbf{X_{test}}$, and $\mathbf{y_{test}}$ using scikit-learn's package `train_test_split()`.

(b) Let the number of neighbors be either 1, 3, 5, 10, 20, 30, 40, 50, 75, or 100. For each number of neighbors, train a kNN model using scikit-learn package `NeighborsNeighbors`. Evaluate the correct classification rate of each model on the validation set.

(c) What model parameter (# of neighbors) yields the highest classification rate in part (b)? Which one do you choose for the final (optimal) model?

(d) Train a new kNN model with the number of neighbors selected in part (c). Use both training and validation data to fit the new model. Apply the model to the test set $\mathbf{X_{test}}$ and compute the corresponding classification rate and the confusion matrix.

2.

(3 points) Use the starter code in "knn.ipynb" to re-implement the kNN classifier. Both `fit()` and `predict()` need to be updated. After reimplementing the classifier, train a model similar to that of (1d) and apply it to the test set. Confirm that you obtain the same results as the sklearn version.

3.

(3 points) Perform decision tree classification on the dataset `winequality-red.csv`. Use the file "DecisionTreeStarter.ipynb". Print the unique class labels. Use 80/20 stratified data partitioning. Provide a figure of the resulting decision tree, the classification accuracy rate, and the confusion matrix when tested with the testing dataset.