**CPE383 Machine Learning: Quiz 9**

1. 20 points. 1 hour. RANSAC Regression. Use RANSAC to find a, b, c for the following dataset where points (xi, yi) are discrete samples from a function f(x) = ax2 + bx + c with 2 outliers.

Hint: You should get a, b, and c close to 2.2, 0.5, -4.5, respectively.

Text

Description automatically generated Text

Description automatically generated

1. Use K Means clustering on the IRIS dataset.
   1. 10 points. 0.5 hour. Using K = 3, cluster the entire dataset into 3 labels using only features 1 & 3; namely, sepal length and petal length (Note: the example in class used all 4 features for clustering). Show a scatter plot based on these 2 features using known training 3 classes using markers "<" for class 1 (Setosa), ">" for class 2 (Versicolor), and "^" for class 3 (Virginica) while also using colors based on the 3 computed clusters using colors of “pink” for cluster 1, “yellow” for cluster 2, and “cyan” for cluster 3.

Text

Description automatically generated Chart, scatter chart

Description automatically generated

* 1. 5 points. Report based on known labels what percent is misclassified when using 2 features. Text

     Description automatically generated
  2. 10 points. 0.5 hour. Plot the result of K Means clustering using all 4 features with K = 4.

Chart

Description automatically generated

* 1. 15 points. 1 hour. Reduce the 4 features (sepal length, sepal width, petal length, petal width) into 2 PCA features (an example is also provided in class). Use K = 3 to cluster the entire dataset using these 2 PCA features. Show a scatter plot like in problem 2.1 along with percent misclassified as in problem 2.2.

Text

Description automatically generated Chart, scatter chart

Description automatically generated

* 1. 20 points. Redo the example in class with all 4 features and K = 3, but using your own class or function’ my\_k\_means in Python that has initialization parameters: K, X, max\_iterations, centroid\_move\_epsilon and returns y as a 1-D array of integer labels of 1, 2, ..., K.. Each input N-dimensional data X[i] will have a 1-dimensional output label y[i] for i = 1..M where M is the number of data points. The algorithm should start by assigning K cluster centers based on random values from the (min, max) range of each dimension in the N-dimensional data X. It should stop when all centers have moved by less than the centroid\_move\_epsilon or when the max\_iterations is reached. Make sure your results are similar to the K Means library class. Text

     Description automatically generated

Text

Description automatically generated

1. 15 points. 1 hour. Decision Trees. Change the “IRIS Decision Tree.ipynb” shown in class, to use SKlearn’s Wine Recognition Dataset instead. Report the classification accuracy % for a single tree using 70%training samples and for a random forest with 100 estimators

Text

Description automatically generatedText

Description automatically generated

The Random Forest Classifier model with 100 estimators has an accuracy of 0.9814814814814815, which is approximately 2.7% better than the Decision Tree Classifier model with an accuracy of 0.96 on the wine dataset.