CSCE 509

Assignment 1

Part 0: Data creation

Generate a 2D, two-class data set. Class 0 are uniformly distributed in the (union of the) 3 squares [-1, 0]x[-1, 0] + [-1, 0]x[0, 1] + [0, 1]x[-1, 0]. Class 1 is Gaussian distributed centered at (0.5, 0.5) with variance (0.5, 0.5).

Generate 150 points in Class 0 and 50 points in Class 1. Plot the points in both classes, using a different color or shape for each class.

Part 1: Perceptron

Implement a Perceptron classifier to separate the two classes. Use a test set of 50 Class 0 and 50 Class 1 points to evaluate the performance of your classifier. Use 4-fold cross validation to report the accuracy, precision, recall, AUROC metrics. What is the variance of the accuracy across the 4 runs?

Part 2: Linear SVM

A. Repeat Part 1 with a Linear SVM. Use a very small *C* hyperparameter.
B. Repeat Part 1 with a Linear SVM. Use a very large *C* hyperparameter

Part 3: Nonlinear SVM

A. Repeat Part 1 with a nonlinear SVM. Use a very small *C* hyperparameter.
B. Repeat Part 1 with a nonlinear SVM. Use a very large *C* hyperparameter.
C. Increase the Class 1 variance to (1.5, 1.5). Plot the data. Repeat Part 1 with a reasonable choice of *C*.

Part 4: Mismatch between training and test data

For test data, use the same distribution of Class 0 as before, such as in Part 3C. Similar to Part 3C, set the Class 1 variance to (1.5, 1.5).

For training data, change the Class 0 samples to be uniformly distributed in the (union of the) 3 squares [-0.5, 0.5]x[-0.5, 0.5] + [-0.5, 0.5]x[0.5, 0.5] + [0, 1]x[-0.5, 0.5]. Class 1 is Gaussian distributed centered at (0.5, 0.5) with variance (0.5, 0.5).

Implement a nonlinear SVM. Train using the training data. Use 4-fold cross validation and the test data to report the accuracy, precision, recall, AUROC metrics. What is the variance of the accuracy across the 4 runs?

Part 5: High variance

Generate the data sets as in Part 3C (i.e., no mismatch between training and test data sets). Reduce the number of samples in the training set to 60 samples in Class 0 and 20 samples in Class 1. Implement a nonlinear SVM. Train using the training data. Use 4-fold cross validation and the test data to report the accuracy, precision, recall, AUROC metrics. What is the variance of the accuracy across the 4 runs?