MATH 189 Homework 7

Due March 4th, 2022

Q1. This problem involves hyperplanes in two dimensions.

(a) Sketch the hyperplane $1 + 3X_1 - X_2 = 0$. Indicate the set of points which $1 + 3X_1 - X_2 > 0$, as well as the set of points for which $1 + 3X_1 - X_2 < 0$.

(b) On the same point, sketch the hyperplane $-2 + X_1 + 2X_2 = 0$. Indicate the set of points for which $-2 + X_1 + 2X_2 > 0$, as well as the set of points for which $-2 + X_1 + 2X_2 < 0$.

Q2. We next investigate a non-linear decision boundary.

(a) Sketch the curve $(1 + X_1)^2 + (2 - X_2)^2 = 4$.

(b) On your sketch, indicate the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 > 4$, as well as the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 \leq 4$.

(c) Suppose that a classifier assigns an observation (X_1, X_2) to the blue class if $(1 + X_1)^2 + (2 - X_2)^2 > 4$, and to the red class if $(1 + X_1)^2 + (2 - X_2)^2 \leq 4$. To what class is the observation (0, 0) classified? (-1, 1)? (2, 2)? (3, 8)?

(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2 and X_2^2.

Q3. In this problem, we implement and compare support vector machines (SVMs) under various settings. We study a simulated dataset which contains a binary response variable (Y) and two continuous covariates (X_1 and X_2) over 300 observations. The dataset has been divided into a training set of sample size 200 (SVM_train.csv) and a test set of size 100 (SVM_test.csv). Analyze this data through the following steps.

(a) Draw a scatter plot of covariates (X_1 and X_2) in training set. Color the observations by their class labels. Analyze the two classes based on the plot. For example, are they visually separable? Can they be perfectly separated? Do you think the decision boundary is linear?

(b) Fit the training set by linear SVM. Select the optimal cost C by cross-validation. Use the classifier you obtained to classify the test set. Plot the classification results on both training set and test set.

(c) Fit the training set by SVM with Gaussian kernel. Select the optimal cost C and tuning parameter γ with cross-validation. Use the classifier you obtained to classify the test set. Plot the classification results on both training set and test set.

(d) Plot and compare the ROC curves of the classifiers you obtained in Step (b) and (c) for training set and test set, respectively. Analyze these two ROC curves.