

Assignment # 3

SYDE 675 Winter 2019

The assignment can be done in Python or Matlab.

You need to submit both your report and the source code implementation for all questions. The report must be a single pdf and the source code must be a single .py or .m file. Please include brief comments in your code. Be sure to label all figures and include a legend where appropriate. The due date for this assignment is **April 10th, 2019**. Please also note that late submission will be subject to a penalty of 20% deduction of the assignment mark per day.

In this homework, you will be using support vector machines to gain an intuition of how SVMs work. You are allowed to use any existing implementations of SVM including MATLAB's built-in functions, OSU-SVM, LibSVM and etc. As a suggestion, you can use Lib-SVM toolbox.

Question 1

Linear SVM for Two-class Problem

Use the 'q1_classA.csv' and 'q1_classB.csv' for this question. In this part, you try different values of the C parameter of SVM. Informally, the C parameter is a positive value that controls the penalty for misclassified training examples. A large C parameter tells the SVM to try to classify all the examples correctly. Use the whole set for training purposes.

1. Load data of two classes and plot to visualizing the dataset on the same figure.
2. Train a linear SVM on the dataset. Try to use different values of C and see how the decision boundary varies. Use $C=\{0.001, 0.01, 0.1, 1\}$.
3. Plot different decision boundaries with different C and compare them beside each other on one figure in your report.
4. Which value of C is the best value for this dataset? Explain the effect of C in training of SVM.

Question 2

Adaboost

In this part you will create an adaboost classifier based on linear SVM to classify the dataset in Question 2.

1. Load and plot 'classA.csv' and 'classB.csv' and visualize them on the same figure.
2. Train a linear SVM with proper C value from the set {0.1, 1, 10, 100} and visualize the decision boundary and report the accuracy based on 10-times-10-fold cross validation.
3. Create an ensemble of classifiers based on Adaboost-M1 approach to classify the dataset again. Use a linear SVM with the selected C in part 2 as your weak learner classifier. Use $T = 50$ as the max number of weak learners.

Note:

I) For each iteration draw only 100 samples from the dataset to train each classifier.

II) If the training error is higher than 50% in one iteration, discard the classifier and re-sample the training set and train a new classifier. Continue until you have trained 50 unique SVMs.

4. Report the mean and variance of accuracy for 10-times-10-fold cross validation approach.
5. Visualize the the decision boundary of the ensemble model on the plot in part 1.