In The Name of Allah

Machine Learning (Fall 2017)

Instructor: Mahdi Yazdian

TA: Mrs. Farzane Nikzad

Homework#3: Support Vector Machine

Due Date: 98.02.10

In this homework, you will be using support vector machines to gain an intuition of how SVMs work and how to use a Gaussian kernel with SVMs. 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.

PART A. Linear SVM for Two-class Problem

We will begin by a 2D dataset which can be separated by a linear boundary. In this part, you will 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.

1. Load and plot data1.mat to visualizing the dataset.
2. Train a linear SVM on the dataset. Try to use different values of *C* and see how the decision boundary varies. Use *C*={0.1, 1, 10, 100).
3. Plot different decision boundaries with different C on one figure.
4. Plot training accuracy of the classifier (y-axis) in terms of different values of C (x-axis)
5. Which value of C is the best value for this dataset? Explain the effect of C in training of SVM.
6. Notice that there is an outlier positive example + on the far left. As part of this exercise, you will also see how this outlier affects the SVM decision boundary. How SVM is affected by this outlier.

PART B. Kernel SVM for Two-class Problem

In this part of the exercise, you will be using SVMs to do non-linear classification. In particular, you will be using SVM with Gaussian kernel. You can think of the Gaussian kernel as a similarity function that measures the distance" between a pair of examples, $(x_{(i)}; x_{(j)})$. The Gaussian kernel is also parameterized by a bandwidth parameter, σ, which determines how fast the similarity metric decreases (to 0) as the examples are further apart.

1. Load and plot data2.mat to visualize the dataset. From the figure, you can observe that there is no linear decision boundary that separates the positive and negative examples for this dataset. However, by using the Gaussian kernel with the SVM, you will be able to learn a non-linear decision boundary that can perform reasonably well for the dataset.
2. Train SVM with Gaussian kernel on this dataset. Find the best value of σ and C through ten-times-ten fold cross-validation. Note that it is better to test the values in multiplicative steps such as 0.01, 0.04, 0.1, 0.4, 1, 4, 10 and 40. Therefore, you should evaluate 64 different models to select the best model!
3. Plot the train and test accuracies and their corresponding variances of ten-times-ten-fold cross validation for different values of C and .
4. Plot the data and the decision boundary for the **best** C and .
5. Report the test accuracy using the best selected model and discuss about the results.

PART C. Kernel SVM for Multi-class Problem

As you know, SVM is a two-class classifier. However, it can be extended for multi-class problems using one-vs.-one or one-vs.-all strategies. Use the Vehicle dataset[1] and one-vs.one strategy to do this part.

a) Determine the best value of C and σ using ten-times-ten-fold cross validation. Plot the train and test accuracies and their corresponding variances of the ten-times-ten-fold cross validation for different values of C with and σ.
b) Plot the data and the decision boundary of the **best** C and σ.
c) Report the test accuracy using the best selected model and Discuss about the results.

Part D. Conclusion

Please itemize your findings and your conclusions regarding the overall results of this homework.

Good Luck

1 A well-known repository of machine learning benchmark datasets is called UCI datasets which is available on http://archive.ics.uci.edu/ml/datasets.html. You can download Vehicle dataset from http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29.