61011 Lab - Week 3

David Wong and Christopher Yau

October 2020

1 Introduction

The targets for this week are listed below. Unlike last week, we do not go through the detailed implementation of SVMs and nearest neighbour methods. Instead, the purpose of this lab is to help you to gain an intuitive understanding of how various hyperparameters affect the classification performance of the two methods.

During (or before) the live Wednesday lab sessions, you may request help by entering your name on the following form:
https://forms.gle/RpBM8C6m9QAJVWGHA.

When you join the weekly blackboard collaborate session, a teaching assistant will move you to a break out room to discuss your issue.

Level 1

On a data set of your choice:

- Use a nearest neighbour classifier on a data set. Supply a random half of your data examples as training data, and test of the remainder. Plot the value of the test error as you increase the value of k from 1 to 15. What do you notice about the even numbered values of K?
- Repeat the target above, but over several random train/test splits - plot an average across random variation, including error bars for the standard error. What do you notice?
- Train and test a Support Vector Machine on a data set.
- vary the γ parameter for the RBF SVM and plot the error as you vary it between 0.01 and 10 (in sensible step sizes). Why does it not make sense to use *negative* values?

Level 2

- Try several different SVM parameters (perhaps C;γ for the RBF kernel) plotting the testing error for each.
- Use the knn model and analyse how much time it takes to classify data points with different numbers of features, and different numbers of examples in the training data. (you may need to generate your own training data to increase the data set size. One way to do this is to select random numbers from a multivariate normal distribution)

Level 3

Find good sets of SVM parameters (perhaps C; γ for the RBF kernel) via a cross validation grid search, splitting the data up as described at the end of section 4.2 in the course handbook. Be sure to plot confidence intervals.