ECEN 689, Challenge 5

Brandon Thayer

November 8, 2018

The details of Challenge 5 can be found here. In short, we're tasked with using support vector machines (SVM) to classify a data set with two classes and two features. The purpose of this report is to explain my choice of kernel for this challenge, and present the SVM decision regions. Before even considering kernel selection, it's important to explore the data. In this simple two feature data set, the data can be explored by plotting it: see Figure 1. It's clear the data are not linearly separable.

Scikit-learn has four available kernels: linear, polynomial, radial basis function (RBF), and sigmoid. All four kernels were tried, and their corresponding decision regions examined. For this data, RBF gives the highest accuracy, and what appears to be the best decision region when considering both the training and testing data. The RBF kernel is defined as: $\exp\left(-\gamma\|x - x\prime\|^2\right)$. Additionally, the kernel accepts a regularization parameter, C.

A grid search with cross validation over logarithmically spaced values of C and γ was performed, and the decision regions for the best parameter combination are shown in Figure 1a. Manual exploration with $C = 1$ and a variable γ was performed, and the decision region for $\gamma = 10$ is shown in Figure 1b.

For the actual classification, I chose to use the manually determined parameters in Figure 1b. The grid search (Figure 1a) results in a region with a boundary which is very close to the lower right linear portion of purple circles, so small noise in testing data could result in incorrect classification. Using higher values of γ result in a "shrunken" version of the region in Figure 1b. While this results in higher accuracy in the training data, it could be over-fitting.
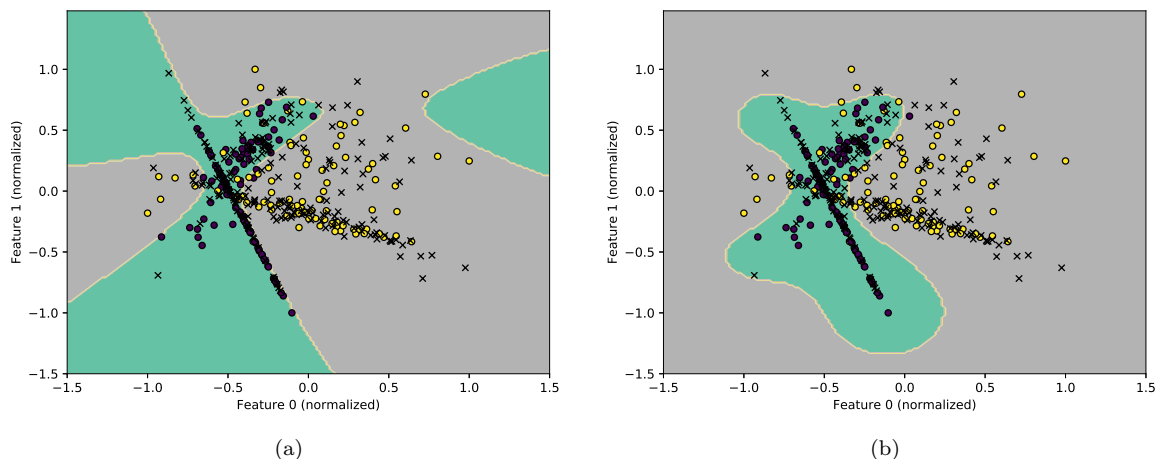

(a)

(b)

Figure 1: Scatter plots of normalized training data. Circles are training data, x's are testing data. Circle colors represent the two classes. SVM decision regions are plotted in the background: in the green area the classifier will predict a purple circle, and in the gray area the classifier will predict a yellow circle. (a): Result of grid search, $C = 10000$, $\gamma = 1$ (b): Manual exploration: $C = 1$, $\gamma = 10$