Homework 2
Kernel SVM and Perceptron

Dana Van Aken

Problem 1: SVM decision boundaries

1. (a) $\longrightarrow 4$
   Since this a linear SVM, it has to be either graph 3 or 4. The correct graph is 4 because there are 1-2 slight margin errors, so C must be small since the data is linearly separable and would try to fit the training data exactly if the cost was high.

   (b) $\longrightarrow 3$
   This corresponds to graph 3 because it's linear and is trying to fit the data more precisely since the cost is higher. The decision boundary in graph 3 is less general than the one in graph 4.

   (c) $\longrightarrow 5$
   This equation is quadratic and it's actual graph should look similar to graph 5. I do /textitnot think that this equation corresponds to graph 2 because graph 2 is noisy and I would expect it to use something like a Laplace kernel.

   (d) $\longrightarrow 1$
   Graphs 1 and 6 are similar in shape and appear to be RBF kernels. This equation specifically corresponds to graph 1 because the variance is larger. A larger variance helps generalize the decision boundary since the data points are going to deviate further from the mean.

   (e) $\longrightarrow 6$
   This equation corresponds to graph 6 since the variance is small. This small variance is causing results in overfitting the data points (when compared to the more general fit in graph 1).

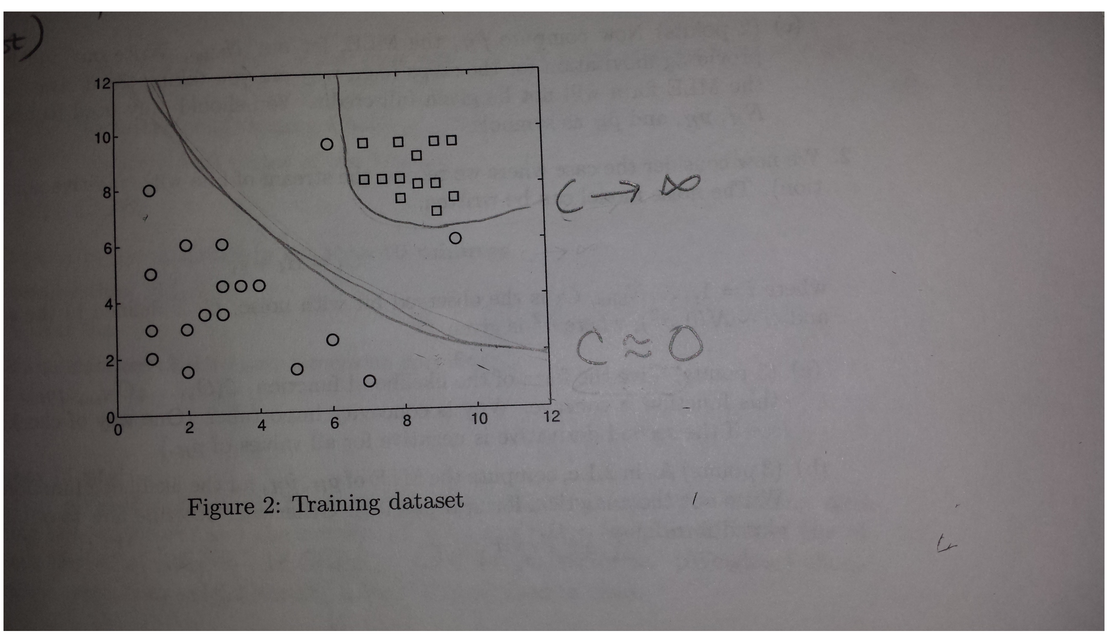

Figure 2: Training dataset

2. (a) As C approaches infinity, the cost of misclassification is extremely high so the model will try to separate the data exactly. In this case, it is successful since the training data is linearly separable, but the resulting data overfits the boundary.

(b) If C is near 0, then there is no cost to overfitting so 1 or more outlier data points do not determine the decision boundary. This means that we can permit the outliers to be margin errors and the resulting decision boundary is much more general compared to the one produced when C is near infinity.

(c) Case (b) would work a lot better in the classification task. Given that the training data is error-prone, using case (a) would be highly inappropriate because it assumes that the training data is noise-free. Case (a) will produce a more general decision boundary that will likely classify the data more successfully than the overfit decision boundary produced by case (b).