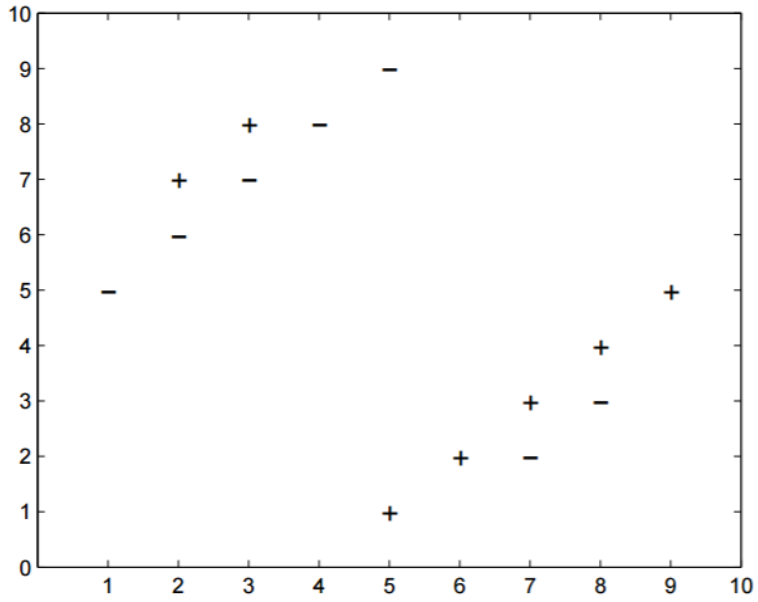
BUSINESS DATA MINING (IDS 572)

HOMEWORK 8
DUE DATE: WEDNESDAY NOVEMBER 16 AT 3:00 PM

- Please provide succinct answers to the questions below.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.

Problem 1. In the following questions you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. Note that a point can be its own neighbor.

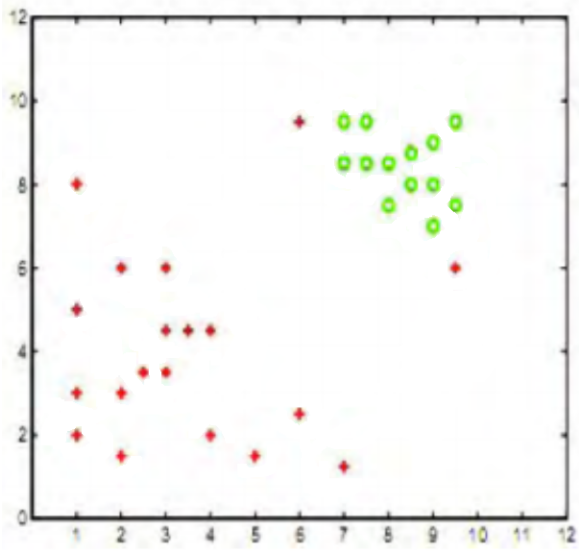


(a) What value of k minimizes the training set error for this data set? What is the resulting training error? Explain.

(b) Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?

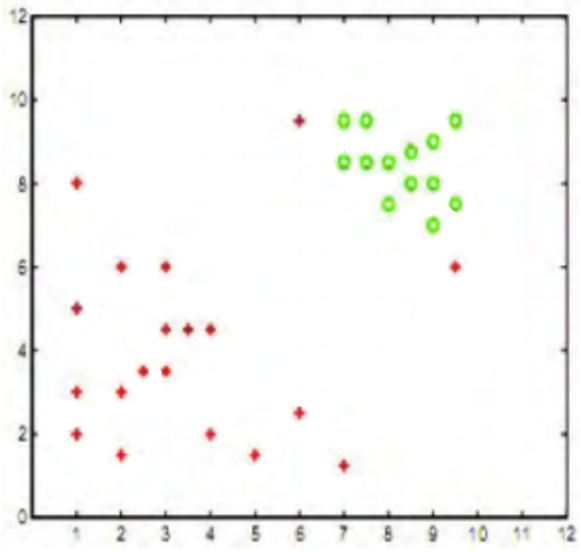
(c) What value of k minimizes leave-one-out cross-validation error for this dataset? What is the resulting error?

Problem 2. The original SVM proposed was a linear classier. As discussed in class, in order to make SVM non-linear we map the training data on to a higher dimensional feature space and then use a linear classier in the that space. This mapping can be done with the help of kernel functions. For this question assume that we are training an SVM with a quadratic kernel - i.e. our kernel function is a polynomial kernel of degree 2. This means the resulting decision boundary in the original feature space may be parabolic in nature. The dataset on which we are training is given below:

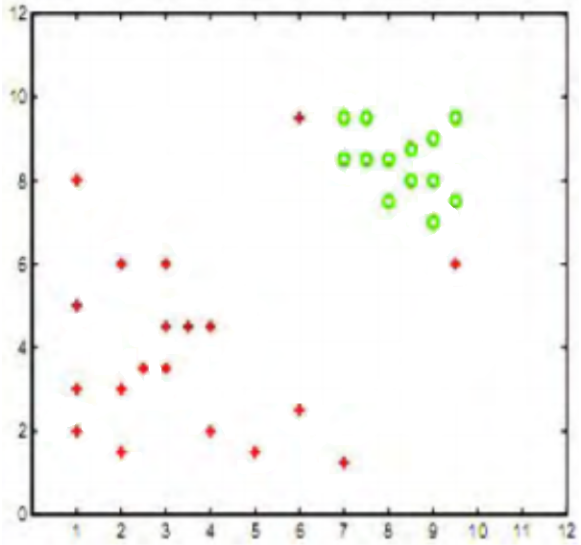


The slack penalty C will determine the location of the separating parabola. Please answer the following questions qualitatively.

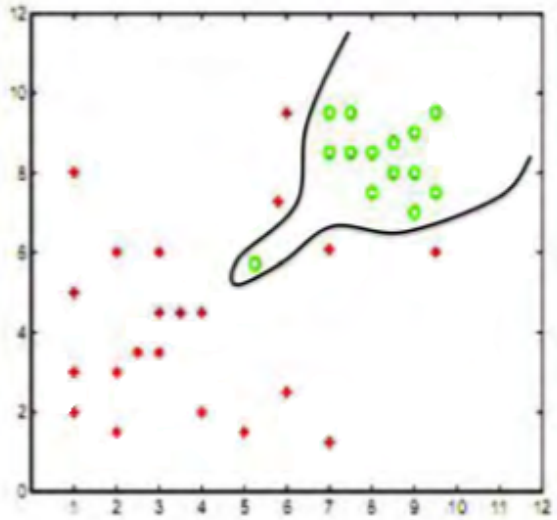
(a) Where would the decision boundary be for very large values of C? (Remember that we are using a quadratic kernel). Justify your answer in one sentence and then draw the decision boundary in the figure below.



(b) Where would the decision boundary be for C nearly equal to 0? Justify your answer in one sentence and then draw the decision boundary in the figure below.



(c) Now suppose we add three more data points as shown in figure below. Now the data are not quadratically separable, therefore we decide to use a degree-5 kernel and find the following decision boundary. Most probably, our SVM suffers from a phenomenon which will cause wrong classification of new data points. Name that phenomenon, and in one sentence, explain what it is.



Problem 3. Please do the "Predicting Customer Churn at QWE Inc" and answer the following questions.

(a) Is Wall's belief about the dependence of churn rates on customer age supported by the data? To get some intuition, try visualizing this dependence (Hint: no need to run any statistical tests).

(b) I want you to specifically run a logistic regression model that best predicts the probability that a customer leaves. (1) What is the predicted probability that Customer 672 will leave between December 2011 and February 2012? Is that high or low? Did that customer actually leave? (2) What about Customers 354 and 5,203?

(c) Answer Well's "ultimate question": provide the list of 100 customers with highest churn probabilities and the top three drivers of churn for each customer.