Machine Learning

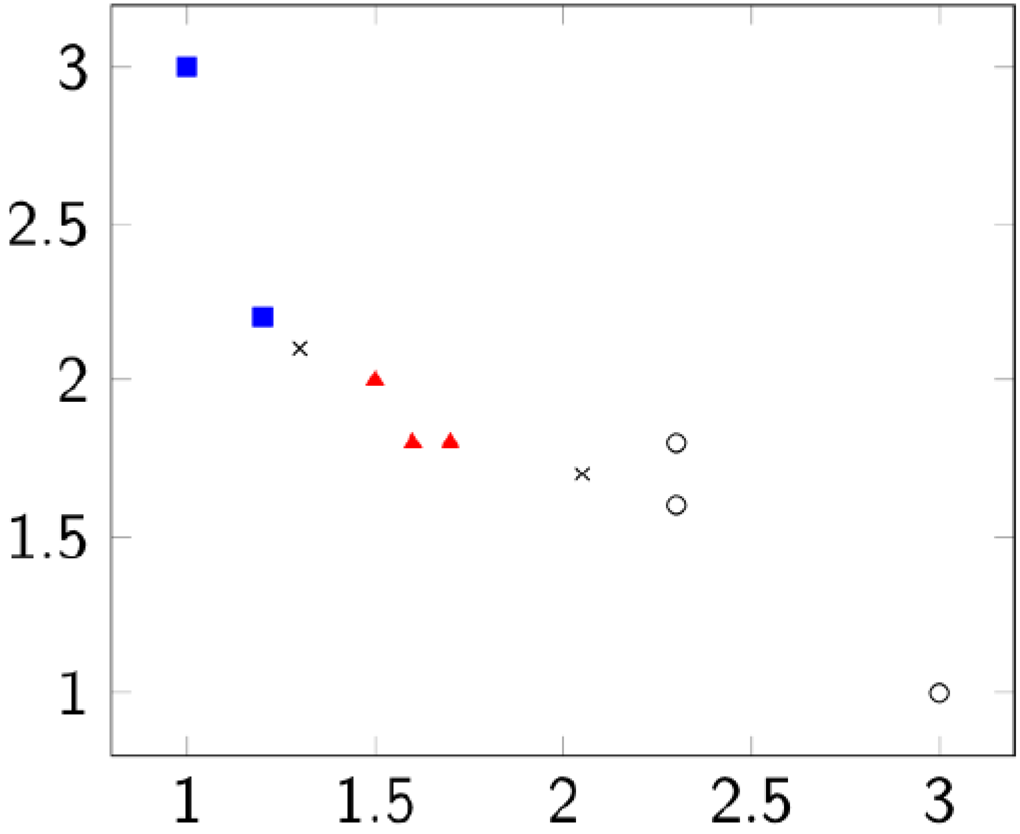
V. Adamchik CSCI 567 Fall 2019

Discussion Set 1 University of Southern California

k- NNC

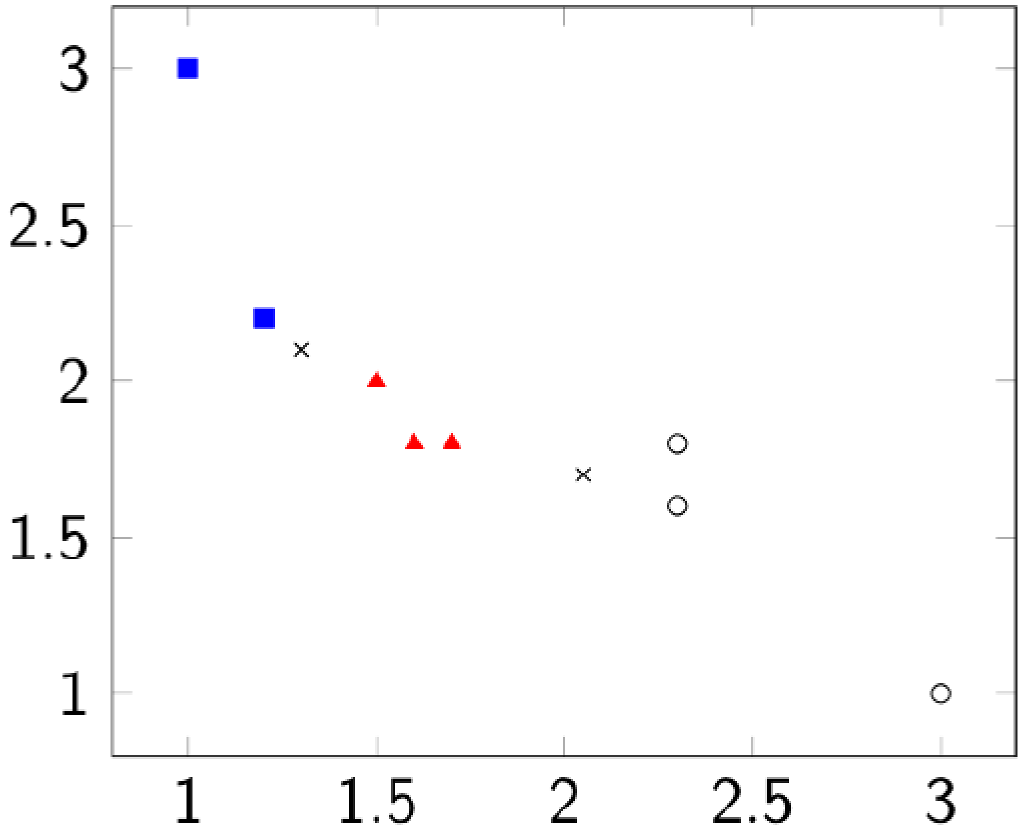
Problem 1a)

Consider the following data; triangles, squares, and circles are three classes of data in the training set; the x items are the unlabeled test data.



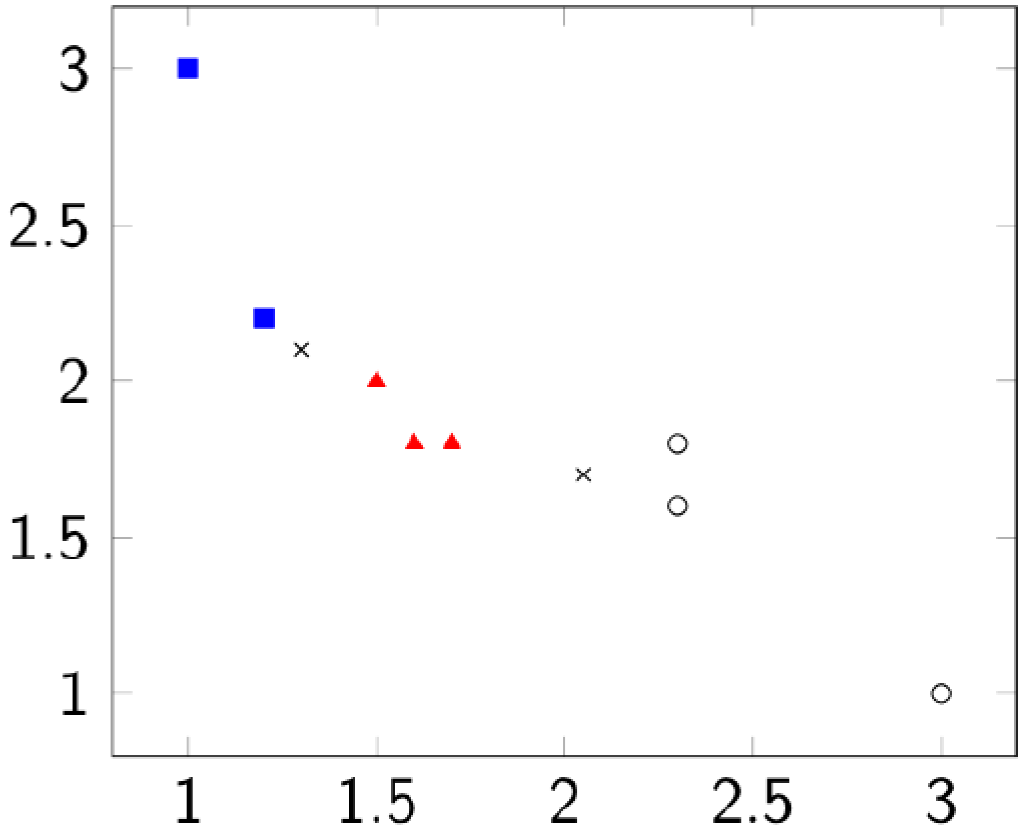
What is the test-data label x be if k = 1?

Problem 1b)



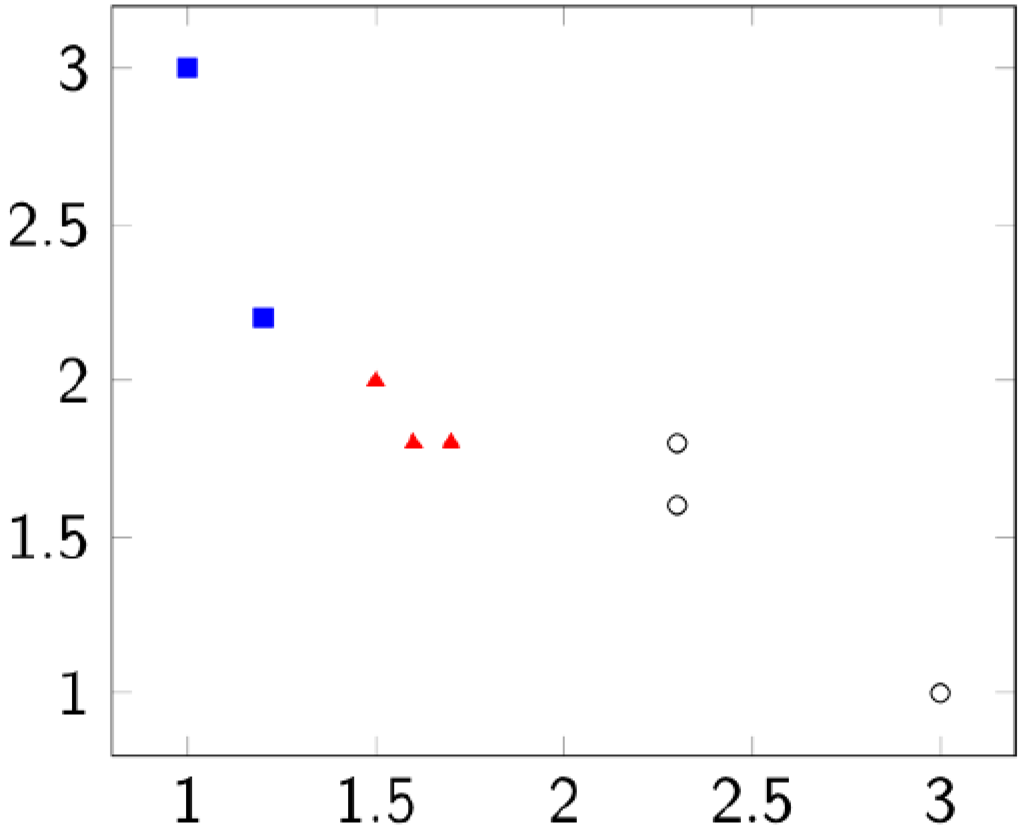
What is the test-data label x be if k = 3?

Problem 1c)



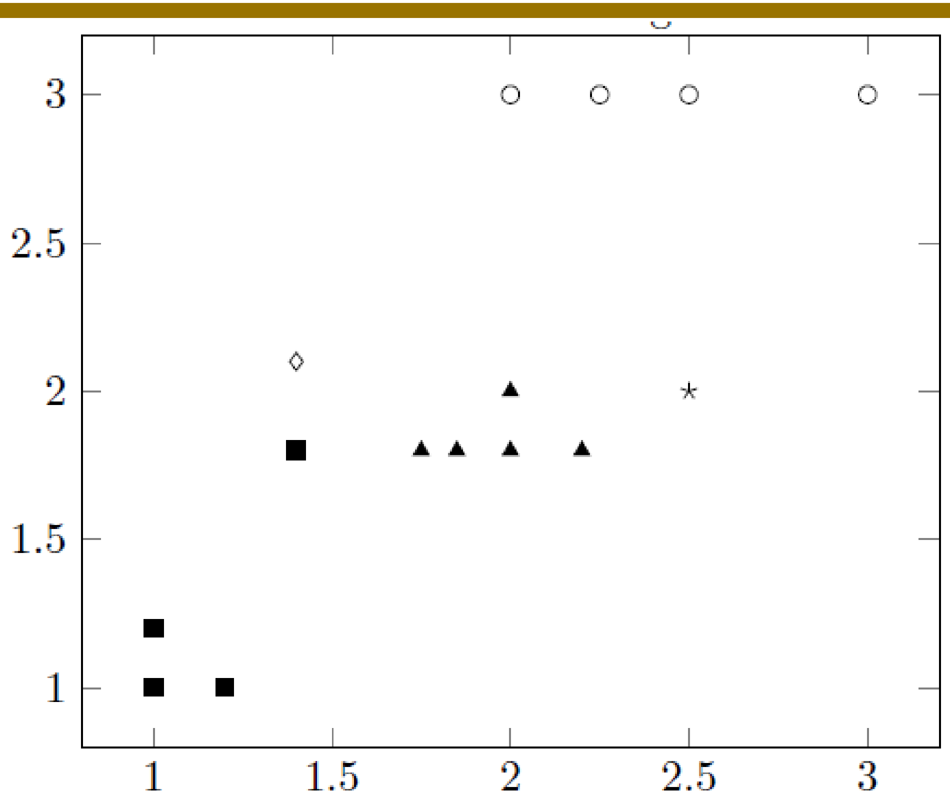
What is the test-data label x be if k = 5?

Problem 2



What is the accuracy and error with leave-one-out, assuming k = 1?

Problem 3

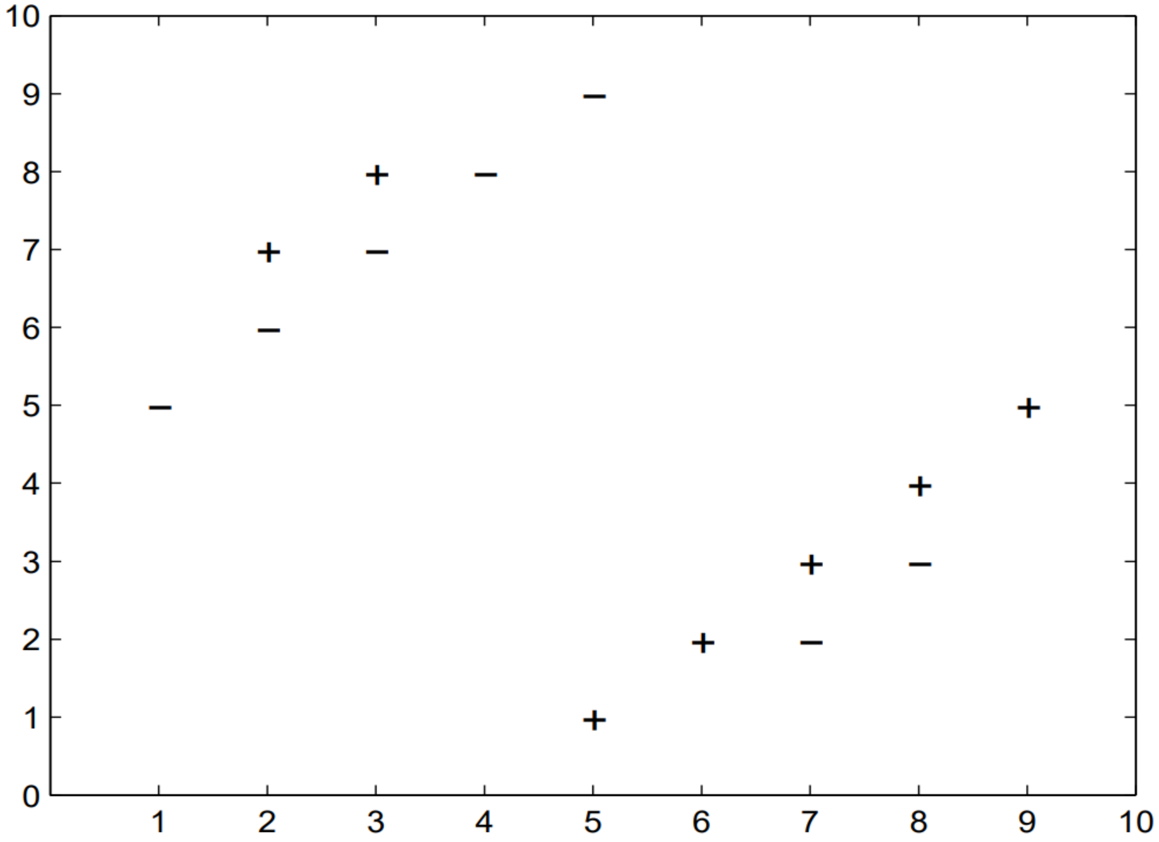


How many training data points will be misclassified with leave-one-out (k = 1)?

What is the smallest value of k to always classify the diamond as class triangle?

What is the smallest value of k to classify the star as open circle?

Problem 4



What value of k minimizes leave-one-out error for this dataset?

Problem 5

Given a k-nearest neighbour model that is trained on N training points using the Euclidean distance where each sample is d dimensional. What is time complexity of testing the model? (There are M samples in test data).

(a) $O(M\ N\ d^2\ k)$
(b) $O(M\ N^2\ d\ k)$
(c) $O(M\ N\ d\ k)$
(d) $O(M^2\ N\ d\ k)$

Problem 6

Increasing k in k-nearest neighbor models will:

(a) Increase bias, increase variance
(b) Increase bias, decrease variance
(c) Decrease bias, increase variance
(d) Decrease bias, decrease variance

Problem 7

If the complexity of a model increases, which of the following is expected to increase?

(a) Bias
(b) Variance

ML Concepts

What is a *parametric* model?

Which of the following phenomenon is called *overfitting*?

(A) low training error, low test error

(B) low training error, high test error

(C) high training error, low test error

(D) high training error, high test error

What is the popular solution for overfitting?

ML Concepts

What is a S-fold *cross-validation* in your own words?

How do you partition the training data into S sets?

Is there is an a priori good choice of S? What is the rule of thumb choice of S?

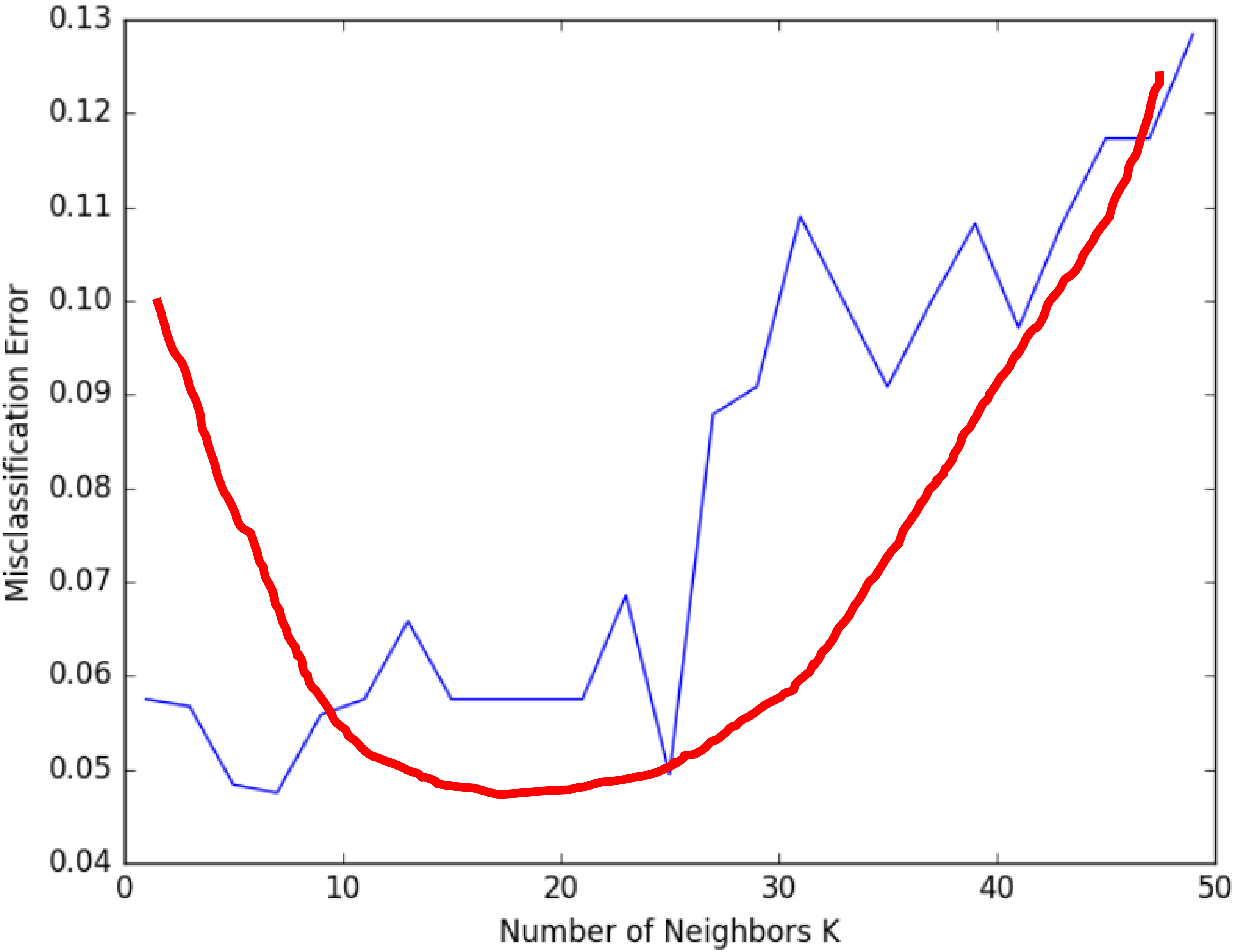
Hyperparameter k

The k-nearest neighbor classifier requires a setting for *k*. But what number works best?

Can we use the training data to find k?

How do you find k if there is no validation set?

Hyperparameter k



Digit Recognition

7291 train ponts, 2007 test points

Error rates:

- Neural net: 0.049
- 1-NN/Euclidean distance: 0.055
- 1-NN/tangent distance: 0.026