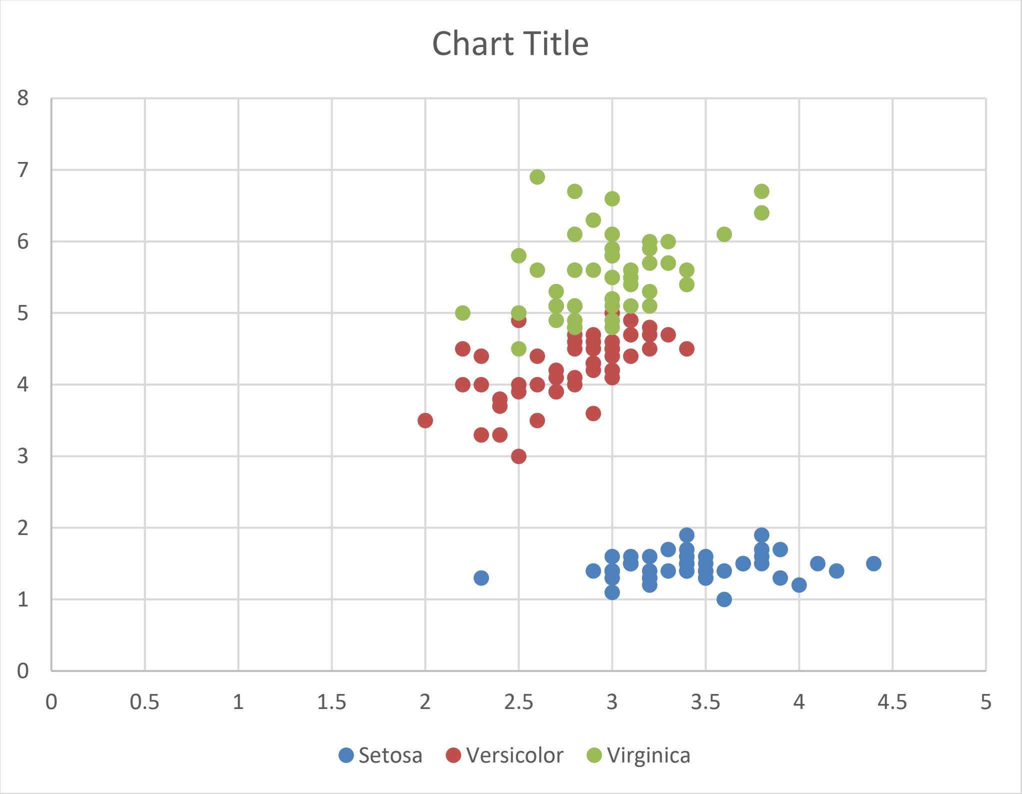
Ariel University

Machine Learning

Homework 3

For this assignment, hand in the code in one file (or via a link), and hand in a separate file with the printouts and answers.

The famous UCI Iris data set contains information on 150 flowers from three species of iris: Setosa, Versicolor and Virginica. For this assignment, we will take only the second and third features for each flower. This gives the following plot:



1. Nearest neighbor search: Implement k-nearest neighbor on the data set. For the first experiment, use only Versicolor and Virginica

A. Sample a training set with half the points. The remaining points are the test set.

B. For each of k=1,3,5,7,9 and p=1,2,∞, evaluate the k-NN classifier on the test set, under the l_p distance. (The base set of the classifier is the training set.) Compute the classifier error on both the training and test sets.

C. Repeat steps (a) and (b) 100 times.

Output the average empirical and true errors for each k and p. Also output the difference between them. Answer the following questions:

Which parameters of k,p are the best? Why is this?

How do you interpret the results? And is there overfitting?

D. Now repeat the experiment while using the net-based condensing shown in class. Answer the following questions:

   How big was the condensed set on average? Did condensing help?

E. Now run the same experiments in A,B,C,D on Setosa and Virginica, and print out the results. Analyze the difference compared to parts C&D and explain the results.

2. Write a decision tree algorithms for this data set.

The decision tree will have exactly three levels (root, children and grandchildren). At each internal node, one can choose to split based on one coordinate in the vector set. At each leaf, one can choose a label of 0 or 1.

Sample half of the flowers, create an optimal decision tree for them, and compute the empirical and true error for this decision tree. Report average error over 50 runs. Also draw the optimal decision tree for one representative sample.