Computational Geometry

Homework Assignment

Instructors: Ioannis Z. Emiris and Anna Karasoulou

Spring 2020

Exercise 1. Implement from scratch the k-NN algorithm in python programming language. Present your method and how you worked, conclude by discussing the disadvantages of the k-NN algorithm, if any.

Exercise 2. Using the code provided at the class (or implementing your own) explain what is the curse of dimensionality and how is related to the k-NN algorithm.

Exercise 3.

1. Suppose there is a set of points on a two-dimensional plane from two different classes. Points in class Red are (0, 1), (2, 3), (4, 4) and points in class Blue are (2, 0), (5, 2), (6, 3). Draw the k-nearest-neighbor decision boundary for k = 1 as we discussed in the lecture. Experiment yourself with two or more differrent distance metrics. Present your results.

2. If the y-coordinate of each point was multiplied by 5, what would happen to the k = 1 boundary? Draw a new picture. Explain whether this effect might cause problems in practice.

3. Can you draw the decision boundary for k=3?

4. Suppose now we have a test point at (1, 2). How would it be classied under 3-NN? Given that you can modify the 3-NN decision boundary by adding points to the training set in the diagram, what is the minimum number of points that you need to add to change the classication at (1, 2)? Provide also the coordinates for these new points and justify your answer.

Exercise 4. How long does it take for k-NN to classify one point? Or in other words what is the testing complexity for one instance? Assume your data has dimensionality d, you have n training examples and use Euclidean distance. Assume also that you use a quick select implementation which gives you the k smallest elements of a list of length m in O(m).

Exercise 5. To see an application of the k-NN algorithm in a real world classification problem consider the data found at https://www.kaggle.com/uciml/iris. Download from there the Iris.csv file. Ignore the id column and consider the columns: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm as point coordinates in a four dimensional space. Consider the column Species as the class/label column. The file contains 150 rows. Run the algorithm on the first 100 rows and make predictions for the rest 50. How your predictions are compared to the actual? Explain your methodology.