Machine Learning *COMP 5/6630*

HW02: Nearest neighbor classifier

Remember that only PDF submissions are accepted.

1. Give an example of a low dimensional (approx. 20 dimensions), medium dimensional (approx. 1000 dimensions) and high dimensional (approx. 100000 dimensions) problem that you care about.

   Low dimensional dataset: Parkinson Dataset

   I chose this one because unfortunately my grandfather has Parkinson. It is a disease that affects the nervous system and movement. It is measured by looking at 23 different attributes which contain biomedical measurements. Scientists can use this data to separate healthy people from people with Parkinson.

   Source: https://archive.ics.uci.edu/ml/datasets/Parkinsons

   Medium dimensional data: Malware static and dynamic features VxHeaven and Virus Total Data Set

   I chose this dataset, because I've always had a keen interest in cyber security. This dataset has 1087 attributes and it can be used to check if something is/contains this form of malware, or not.

   Source: https://archive.ics.uci.edu/ml/datasets/Malware+static+and+dynamic+features+VxHeaven+and+Virus+Total

   High dimensional data: Deepfakes: Medical Image Tamper Detection Data Set

   I chose this dataset because again it relates to cybersecurity, and its also very interesting. This dataset is designed to see if medical evidence has been tampered with. It can even differentiate between a fake cancer that has been injected into the imagery, and real cancer. According to the study, 3 radiologists have evaluated the dataset and they could not reliably tell the difference between the fake and real cancers. This means that this is a very challenging task.

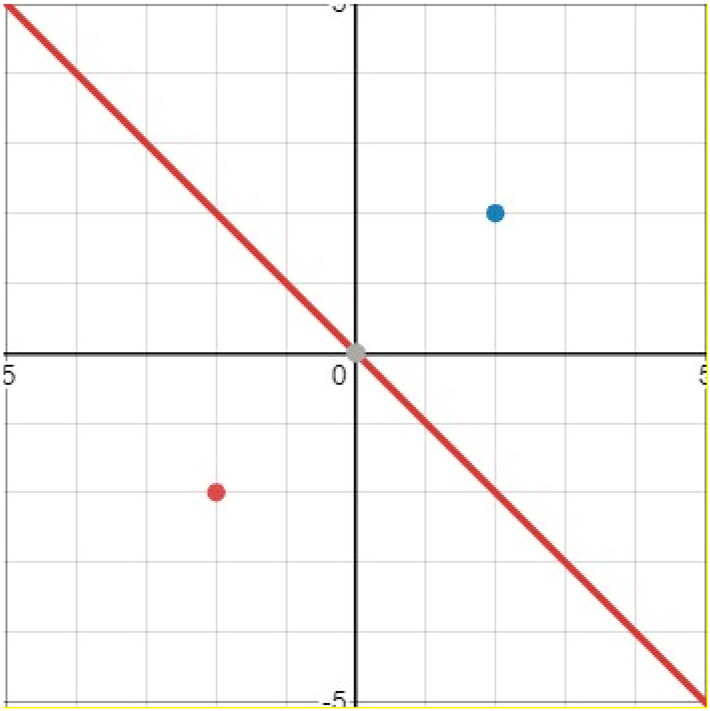
   Source: https://archive.ics.uci.edu/ml/datasets/Deepfakes%3A+Medical+Image+Tamper+Detection

2. What does the decision boundary of 1 nearest neighbor classifier for 2 points (one positive, one negative) look like?

I think this concept can be explained best, by using a visual example.

Below we see 2 points and a function representing our decision boundary.

Point A: (-2,-2) Point B: (2,2) Decision boundary: f(x) = -x



The decision boundary, would basically be an axis of symmetry between the two points. In other words, if we pick any point on the line of our decision boundary, that point will have the same distance to both the the negative point A and the positive point B

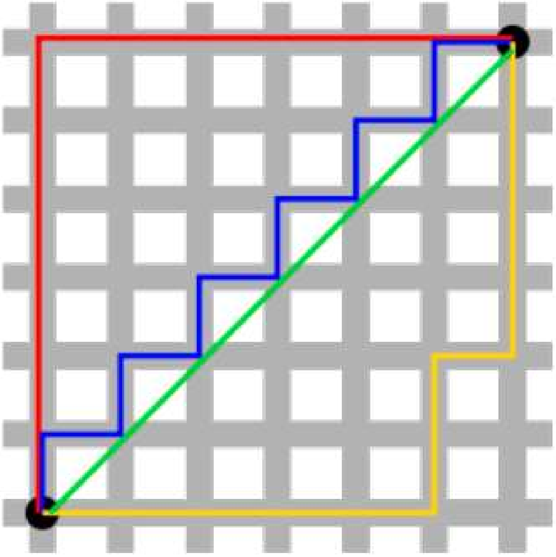
3. Clustering was introduced as a way to speed up k nearest neighbor (kNN) classification. Is it possible that clustering can lead to a better classifier? Briefly explain why.

Clustering can indeed be used as a tool to achieve a better classification. It has been proven by machine learning practitioners, that clustering by itself is not a good way to perform a classification task. However, when the clusters that are obtained from the clustering process are used as features in another machine learning classifier, then it will greatly improve the classification performance and accuracy.

4. Give two examples of data where the Euclidean distance is not the right metric.

- High dimensional data
- Sparse data

5. Does the accuracy of a kNN classifier using the Euclidean distance change if you (a) translate the data(b) scale the data (i.e., multiply the all the points by a constant), or (c) rotate the data? Explain. Answer the same for a kNN classifier using Manhattan distance[1].
Euclidian distance: green line
Manhattan distance: red, yellow, and blue line.



a) Translating the data
If we translate the data, that would mean that we just shift all the points a certain way. In this case, all the distances stay the exact same, which means that this will not affect the accuracy using Euclidian or Manhattan distance.

b) Scaling the data (multiply ALL the point by a constant)
I believe that if we multiply ALL the points by the same constant, there will be no change in accuracy for either the Euclidian or the Manhattan distance. (With the exception of multiplying all our data by zero as a constant. That would mean all our data becomes zero and therefore all our data would be the same. Which means all our data would become useless with either Euclidian distance or Manhattan distance)

c) Rotating the data
Euclidian distance: the accuracy does not change since the data changes uniformly.
Manhattan distance: the coordinates will get transformed and therefore changing the accuracy.

[1] http://en.wikipedia.org/wiki/Taxicab_geometry