Solutions

Yongxing NIE
College of Engineering
Northeastern University
Toronto, ON
nie.yo@northeastern.edu

Q2.

A: I will choose decision tree. Because the dataset has both positive and negative points, but points with different signs might be close in space/coordinates. KNN are based on the nearest neighbors, the nearest neighbor in space might have different signs. While decision tree splits on the best attribute, therefore it's more suitable for this problem.

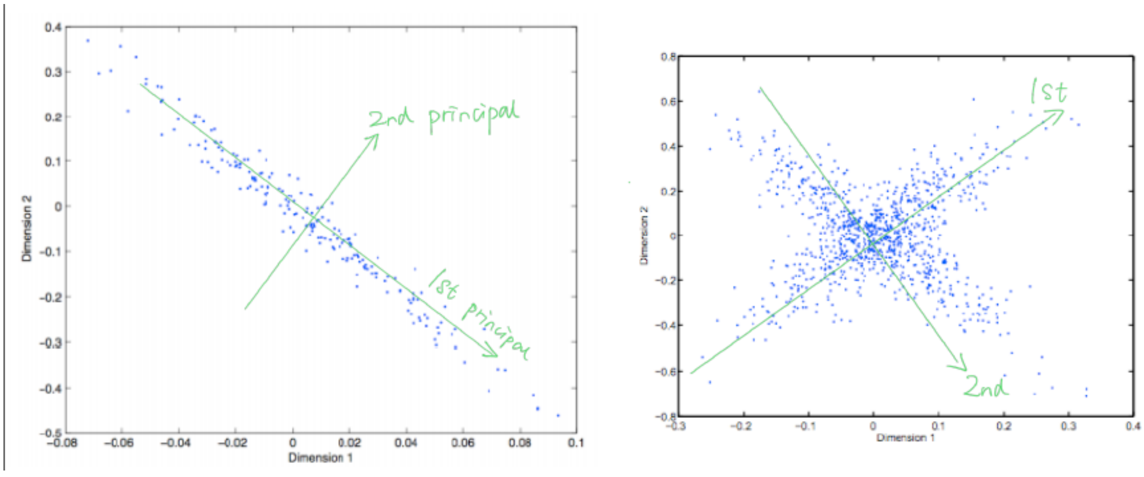
Q3.

(1) A: 2. In a 2D space, with any set of three points, their distance to the origin would be d1, d2, d3. The distance would follow the order d1<=d2<=d3. We cannot include points with radius d1 and d3 inside while excludes point with radius d2. Therefore, the VC dimension of an origin centered circle is two.

(2) A:2. The reason is the same as the origin centered circle, and we consider radius r1, r2, r3 in this case.

Q4.

A:



For the graph in the right, the data points are almost symmetrical, there are two versions of 1st and 2nd principal components.

Q5.

A:

(a) **Hierarchical clustering with single link** will produce the result. Because the data points are arranged in two rows, the distance between two closest points in two rows will be larger than the that in one row, therefore single link will work well here.

(b) **Hierarchical clustering with complete link** will produce the result. The blue points next the yellow points are less distant to the yellow cluster than to points in its own cluster, that is the trait of complete link.

(c) **Hierarchical clustering with average link** will produce the result. The pattern shows a balance between chaining and crowding, especially for the part where blue points and yellow points are mixed.

Q6.

A: I have applied decision tree to perform Boston house price regression in my midterm. Since the house prices are continuous value, I apply decision tree regressor to predict the house price. And I calculate the R^2 and mean squared error (MSE) to measure the fitting performance.

Since we are predicting continuous variables, we cannot calculate the entropy and information gain as done in decision tree classification. MSE can tell us how much our predictions deviate from the original targets. We only care about how much the deviations are. Not in which path or direction is better. Therefore, we square the deviations and divide the entire sum by the total number of records. And the expected value at any leaf are all the same.

Q7.

A: Lazy learning stores data set without learning from it, and it start classifying only after it receives test data, while eager learning does the opposite, and it takes longer time in learning and less time in classifying data.

The advantages of lazy decision tree include:

- Compared with eager decision tree which create a single decision tree for classification, lazy decision tree creates a path in a tree that would be best for a given test instance. If many features are envolved, we may still have enough data to make the nodes. In other words, it is capable of handling problems with many attributes.
- Since lazy learning is creating a possible path, it can handle missing values well by avoiding splits on such values.
- It is simpler since we don’t need to build all possible trees but only a possible path.

The disadvantages of lazy decision tree include:

- There is no pruning.