DM Assignment 01

10.1.2016

Adithya Abraham Philip

1PI13CS008

Overview

The objective of the assignment is to predict the class of an Iris plant based on certain physical characteristics. The data set used contained petal and sepal lengths and widths of 150 Iris plants. 10 were randomly chosen to be the test set while the remaining were used as training data.

Goals

1. Obtain estimate of classifier accuracy for each distance metric (Euclidean, Manhattan, Supremum) using k=1..5 (th) closest neighbours.
2. Observe and explain variation in accuracy for different distance metrics and k values.
3. Properly formatted tables are not included in this document, since they are automatically generated by the program code, as can be seen in **output.txt** (preferably opened in a text editor with **text-wrap disabled**).

Observations - Accuracy

K = 1; Euclidean = 0.9; Manhattan = 0.9, Supremum = 0.9

K = 2; Euclidean = 0.9; Manhattan = 0.9, Supremum = 0.8

K = 3; Euclidean = 0.9; Manhattan = 0.9, Supremum = 0.9

K = 4; Euclidean = 1.0; Manhattan = 1.0, Supremum = 1.0

K = 5; Euclidean = 1.0; Manhattan = 0.9, Supremum = 1.0

Two sets of analysis are presented - variation of accuracy with **distance measure** and variation with **k (th closest neighbours)**.

Variation with Distance Measure

It can be seen that Euclidean Distance seems to be the most accurate for k=1..5. Supremum catches up with the movement of k towards higher values (k=4, 5), while Manhattan reaches a peak at k=4 but falls behind at k=5. Considering the fact that the test data is relatively small, this observation cannot be taken as conclusive, and only as a tentative hypothesis. The validity of the observation could be strengthened by using a larger range of k values and test data set.

Variation with K (th Closest Neighbours)

The general trend is a rise in accuracy towards higher values, but with an apparent limit. k=4 seems to be the "sweet spot" with maximum accuracy (1.0) for all distance measures, while k=5 sees Manhattan Distance fall behind in accuracy (0.9).

Explanation:

k=1: Picking only one closest neighbour has a higher probability of resulting in an inaccurate prediction, as that neighbour could be an **outlier** in its class, or the result of an **incorrect entry**, which simply happens to be close to the test object in terms of distance.

k=2: The situation is barely an improvement over k=1, since even if the second closest neighbour was the correct class, the first closest neighbour being the wrong class would result in a **50-50 tie**, which must be broken **randomly** or **arbitrarily**. This does not contribute to accuracy.

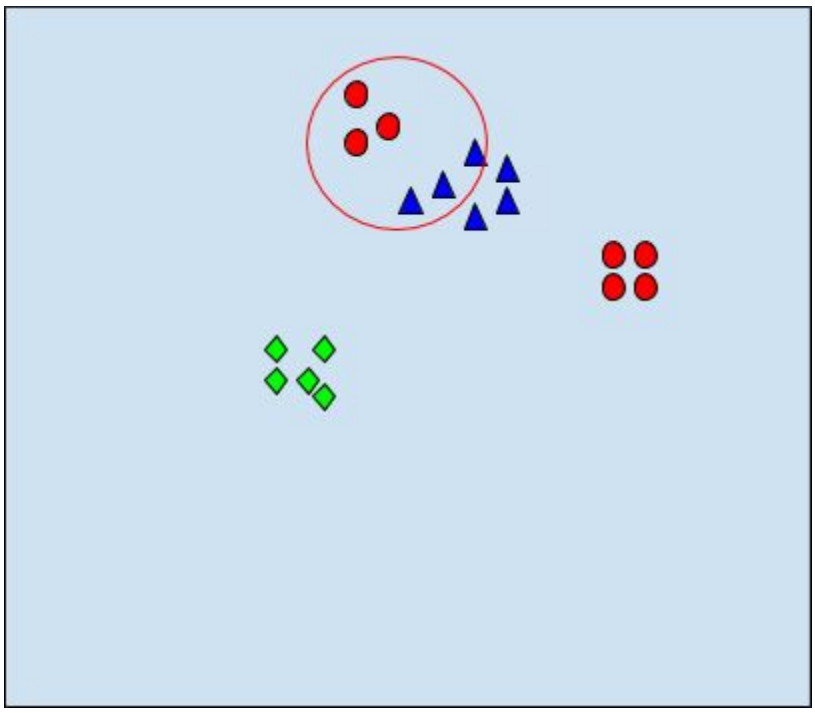
k=3: This is better than k = 1 or 2, since there's a greater possibility that clustering will cause the closest neighbours to be of the correct class. However, since there are 3 classes, it is still **not sufficient** to ensure a higher probability of getting a **tie-free** prediction.

k=4: The accuracies are at their **maximum** here, presumably because by allowing a greater number of closest neighbours we increase our probability of getting the correct class predicted.

k=5: The accuracy is seen to **dip slightly** for Manhattan here, which could be **circumstantial** or the result of each species having **multiple localized clusters** as depicted on the next page.

Another possible scenario is that the test object is at the edge of its class' cluster, thereby resulting in objects of other nearby classes being picked.

Also, the first 2 closest neighbours may be correct and close to the test object, but the remaining 3 may be far away. However, in our current scheme, they get equal weightage. A possible solution is to introduce weights for the neighbours inversely proportional to their distance.



The circled portion shows a possible scenario where we went wrong because we picked up too many neighbours. This issue can be rectified by not blindly using the count of classes of closest neighbours, but instead introducing a weighting factor for each neighbour which is inversely proportional to their distance. This should help if there are incorrect neighbours, which are significantly further away.