Ricardo Aler Mur

SELF-ASSESSMENT K-Nearest Neighbour, WITH QUESTIONS AND ANSWERS

1) Why classifiers based on nearest neighbor(s) perform poorly in the presence of irrelevant attributes?

Answer: because they are based on the idea of proximity, which is measured by a distance. For example, if we use the Euclidean distance, this distance is calculated as the sum of several squared differences, as many as attributes. If many of the attributes are irrelevant, many of the components of the distance will also be arbitrary and also the classification resulting from it.

2) What does KNN become when K = number of training instances?

Answer: it becomes a trivial classifier that classifies all instances as a constant class, the majority class. More in detail, if K == number of training instances, all data instances are part of the neighborhood data to classify, and that is so regardless of the instance to be classified. Therefore this classifier will always respond with the same class, the majority class present in the training data.

3) How can we reduce the influence of noisy data in KNN classification?

Answer: in several ways. We can use K> 1. We can also remove noisy instances (by algorithms such as the Wilson editing rule).