kNN algorithm

**Nikhil Bhargav Raman(1MS15CS078)**

**Sharan Sridhar(1MS15CS112)**

**Subramani R(1MS15CS126)**

**Supreeth S Angadi(1MS15CS129)**

**Abstract:**

The model for kNN is the entire training dataset. When a prediction is required for a unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance.

The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other other types of data such as categorical or binary data, Hamming distance can be used.

In the case of regression problems, the average of the predicted attribute may be returned. In the case of classification, the most prevalent class may be returned.

**Working:**

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given “unseen” observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by

but other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance.

More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

* It runs through the whole dataset computing d between x and each training observation. We’ll call the K points in the training data that are closest to x the set A. Note that K is usually odd to prevent tie situations.
* It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label.

More formally, given a positive integer (usually odd, to prevent tie conditions) K, an unseen observation x and a similarity metric d, kNN classifier performs the following two steps:

* It runs through the whole dataset computing d between x and each training observation. We’ll call the K points in the training data that are closest to x the set A.
* It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label.

Finally, our input x gets assigned to the class with the largest probability.

**Attributes considered:**

* User id.
* Gender(Categorical)
* Age.
* Estimated Salary.
* Purchased (Categorical)

**Inference(s):**

1. Given a tuple containing age and estimated salary, it is estimated whether the person buys a car or not.
2. The comparison of the efficiency of logistic regression for classification and k-Nearest neighbor for the same .
3. The plots of residuals vs. fits, q-q plot, residuals vs. leverage and scale location is thereby used to find the error rate.

