Machine Learning - Lecture k-Nearest-Neighbor classifier

Prof. Dr. Dewan Md. Farid

Professor of Computer Science
United International University

June 13, 2023

k-nearest-neighbor (kNN) classifier

- kNN is a simple classifier, which uses the distance measurement techniques that widely used in pattern recognition.
- kNN finds k instances, $X = \{x_1, x_2, \cdots, x_k\} \in D_{training}$ that are closest to the test instance, x_{test} and assigns the most frequent class label, $c_l \rightarrow x_{test}$ among the X.
- When a classification is to be made for a new instance, x_{new}, its distance to each $A_j \in D_{training}$, must be determined.
- Only the k closest instances, $X \in D_{training}$ are considered further.
- The *closest* is defined in terms of a distance metric, such as Euclidean distance.

kNN classifier (con.)

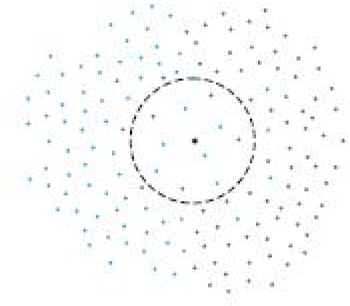

Figure: Nearest neighbour using the 11-NN rule, the point denoted by a "star" is classified to the class.

Euclidean Distance

The Euclidean distance between two points, $x_1 = (x_{11}, x_{12}, \cdots, x_{1n})$ and $x_2 = (x_{21}, x_{22}, \cdots, x_{2n})$, is shown in Eq. 1

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{1}$$

Distance Measures

The distance between the two points in the plane with coordinate (x,y) and (a,b) is given by:

$$EuclideanDistance, (x, y)(a, b) = \sqrt{(x-a)^2 + (y-b)^2} \tag{2}$$

$$ManhattanDistance, (x, y)(a, b) = |x-a| + |y-b| \tag{3}$$

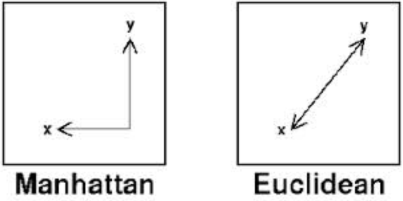

Figure: Euclidean and Manhattan distance.

kNN classifier (con.)

- For, kNN classifier, the unknown instance, $x_{unknown}$ is assigned the most common class, c_l among its k nearest neighbours.
- The k is chosen to be odd for a two class classification and in general not to be a multiple of the number of classes M.
- Usually, kNN achieves good results when the data set is large.
- The value of k should be large for classifying the noisy data.
- Also we can consider the majority voting over the k nearest neighbours to deal with noisy instances.
- Algorithm 1 outlines the k-nearest-neighbor algorithm.

Algorithm 1 k-nearest-neighbor classifier

Input: $D = \{x_1, \cdots, x_i, \cdots, x_n\}$
Output: kNN classifier, kNN.
Method:

1. find $X \in D$ that identify the k nearest neighbours, *regardless* of class label, c_l.
2. out of these instances, $X = \{x_1, x_2, \cdots, x_k\}$, identify the number of instances, k_i, that belong to class c_l, $l = 1, 2, \cdots, M$. Obviously, $\sum_i k_i = k$.
3. assign x_{test} to the class c_l with the maximum number of k_i of instances.

Disadvantage of kNN classifier

- The main disadvantage of the kNN classifier is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification.
- A serious drawback associated with (k)NN technique is the complexity, $(O(kN))^2$, in search of the nearest neighbour(s) among the N available training samples. Although, due to its asymptotic error performance, the kNN rule achieves good results when the data set is large, the performance of the classifier may degrade dramatically when the value of N training instances is relatively small.

An Illustrative Example

Table: Data for Height Classification.

Name	Gender	Height	Output
Kristina	F	1.6 m	Short
Jim	M	2 m	Tall
Maggie	F	1.9 m	Medium
Martha	F	1.88 m	Medium
Stephanie	F	1.7 m	Short
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Short
Dave	M	1.7 m	Short
Worth	M	2.2 m	Tall
Steven	M	2.1 m	Tall
Debbie	F	1.8 m	Medium
Todd	M	1.95 m	Medium
Kim	F	1.9 m	Medium
Amy	F	1.8 m	Medium
Wynette	F	1.75 m	Medium

An Illustrative Example (con.)

- Using the sample data from Table 1 and the **Output** classification as the training set output value, we classify the instance **(Pat, F, 1.6)**.
- Only the height is used for distance calculation so that both the Euclidean and Manhattan distance measures yield the same results; that is, the distance is simply the absolute value of the difference between the values.
- Suppose that $K = 5$ is given. We then have that the K nearest neighbours to the input instance are **(Kristina, F, 1.6), (Kathy, F, 1.6), (Stephanie, F, 1.7), (Dave, M, 1.7)**, and **(Wynette, F, 1.75)**.
- Of these the five item, four are classified as short and one as medium. Thus, the kNN will classify **Pat** as **short**.

*** THANK YOU ***