k-NN Algorithm

Aaryan
CO21BTECH11001

Let number of features of dataset = n

Let number of training inputs = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Let test point be x

k-NN is a classification algorithm which works on the principle that similar inputs have similar outputs.

We will assign the test point a label which is most common label amongst its k most similar training inputs.

Denote the set of k nearest neighbors of x as S_x such that $|S_x| = k$

To find S_x , we can make an array of distances of x from every training input.

$$arr[i] = dist\left(x, X^{(i)}\right)$$

Where

$$dist\left(x, X^{(i)}\right) = \left(\sum_{j=1}^{n} \left|x_j - X_j^{(i)}\right|^p\right)^{\frac{1}{p}}$$

Where we generally take p=2 (Euclidean Distance).

Now we take the k smallest values of array arr and store the corresponding training input in S_x

Now the classifier $h(\,)$ is the function returning the most common label in S_x

$$h(\,) = mode(\{y_i: (X^{(i)}, y_i) \in S_x\})$$

Questions –

1. What type of algorithm is k-NN?
   Ans. It is a classification algorithm, the training data is generally labeled and the label of a new test point is to be found.

2. What happens if k is chosen as m?
   Ans. It will then find the m nearest neighbors of x which is the complete array of distances. Therefore, the classifier will return the maximum occurring label as the label of test data.

3. How is the value of k chosen?
   Ans. The value of k generally varies with the problem. We have to make trials and errors for getting the optimal value of k. For starting, we can take the value of k as the odd number nearest to $\sqrt{n}$

4. What happens if we chose p as ∞ ?
   Ans. In that case, the distance formula will return the maximum of $\left|x_j - X_j^{(i)}\right| \forall\, j \in [1, n]$.

5. What are the fields of machine learning where k-NN Algorithm is generally used?
   Ans. k-NN is generally used in search applications where we are looking for similar items.

6. What are disadvantages of using k-NN algorithms?

Ans. 1. Doesn’t work well with large dataset or higher dimensions.

2. Sensitive to outliers and missing values.