k-Nearest Neighbours
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

18 Jul 2022

1. Introduction

k-nearest neighbours is a supervised machine learning algorithm that is used for both classification and regression although it is mostly used as a classification algorithm. It is a very simple and intuitive algorithm.

The basis of this algorithm is the assumption that points closer to each other are similar and points far from each other are not. When used for classification, the label assigned to a point is the label that a majority of its k nearest neighbours have. When used for regression, the label assigned to a point is the average of the labels of its k nearest neighbours.

2. Mathematical Formulation

Suppose we are given m labelled samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)$ where $x_i \in \mathbb{R}^n \ \forall i \in [m]$
Let $\mathcal{N}_k(\mathbf{x}_i)$ denote the set of the k nearest neighbours of $\mathbf{x}_i$ and $d(\mathbf{x}_i, \mathbf{x}_j)$ be a metric that denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$

I) If $y_i \in \{1, 2, \dots, c\}$ (classification into c categories)
Let $\text{count}(\mathcal{N}_k(\mathbf{x}_i), p)$ denote the number of points in $\mathcal{N}_k(\mathbf{x}_i)$ that have the label p. Then,

$$y_i = p \text{ if } \text{count}(\mathcal{N}_k(\mathbf{x}_i), p) > \text{count}(\mathcal{N}_k(\mathbf{x}_i), j) \ \forall j \neq p \tag{2.1}$$

II) If $y_i \in \mathbb{R}$ (regression)
Then,

$$y_i = \frac{1}{k} \sum_{j : \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} y_i \tag{2.2}$$

3. Weighted Classifier

One of the biggest drawbacks of the *k*-nearest neighbours algorithm is that the classification is skewed when the distribution of categories in the input data is uneven. If there is a category that has a much larger number of data points than other categories, then a random query point is very likely to get assigned to that category irrespective of the distance, if k is large enough.

In order to prevent this, weights can be assigned to neighbours based on their distance from the query point so that closer points have more contribution to the determination of the label of the query point. Without weights, a very close point would have had the same contribution as that of a very far away point provided they were both among the k nearest neighbours. With weights, even if there is a category with a disproportionately large number of points, their contribution will be less if they are far away from the query point.

4. Choosing the Optimal k

The performance of the model is crucially dependent on the value of k. Hence, it is important to choose k properly. Unfortunately, there is no mathematically rigorous algorithm or statistical technique to find the optimal k, which is why we usually have to resort to trial-and-error.

Some people choose k to be equal to $\sqrt{m}$ or $\frac{\sqrt{m}}{2}$ where m is the number of input samples. These are reasonably good estimates. Alternately, one can choose k by cross-validation, i.e., iterating over some values of k and choosing the one that gives the least error. The error is usually measured with the help of a confusion matrix.

5. Questions

i) Give an example of a weight that can be used for a weighted classifier.
ii) What happens when k is too low?
iii) What happens when k is too high?
iv) Let the points $(0,0),(1,0),(0,1),(-1,0),(0,-1)$ have the categorical labels $1,2,1,3,2$ respectively. What will be the label of the point $(1,1)$ with $k = 3$?
v) Repeat the above but with real labels $0.4,-1,2.1,1.6,0$

6. Answers

i) The inverse of the distance between two points can be used as a weight.

$$w = \frac{1}{d} \tag{6.1}$$

ii) When k is too low, noise and irrelevant features have a huge impact on the classification because the label of a query point can get decided by just the noisy points.
iii) When k is too high, we are including points that are very far away from the query point too and giving them the same weightage as those close by, which can end up skewing the results.
iv) The three closest points to $(1,1)$ are $(0,0),(1,0),(0,1)$ which have the labels $1,2,1$. Since label 1 is the most common among these, $(1,1)$ will be assigned the label 1
v) The three closest points to $(1,1)$ are $(0,0),(1,0),(0,1)$ which have the labels $0.4,-1,2.1$. The label assigned to $(1,1)$ will be the average of these, i.e.,

$$\frac{0.4-1+2.1}{3} = \frac{1.5}{3} = 0.5 \tag{6.2}$$