DATA SCIENCE

CLASS TEST

Name: Abhiroop Mukherjee

Enrl. No: 510519109

G-Suite id: 510519109.abhirup@students.iiests.ac.in

---

6) i) K-Nearest Neighbour is Supervised Learning Algorithm which predicts ~~using~~ by calculating "distances" between the datapoints ~~are~~ in the dataset. and use these distances to predict the outcome [label] for an unknown data.

→ As this Algorithm works on the idea of "distance", it doesn't generate any parameter while computing, and hence doesn't "learn" from training set immediately due to this fact, this Algorithm is called

lazy learner Algorithm

→ This Algorithm just stores the dataset, and at the time of classification, it computes the distances ~~and takes distances~~ and uses them to predict label of unknown data.

ii) → There is no definite way to determine the best value of K while training KNN Model.

→ So we generally do trial and error

→ We could use the following way to determine the best value of K

i) Assume a K

ii) "Train" Model [store data]

iii) Test it with a Test Dataset and find it's accuracy

iv) do step (i) to (iii) for a range of K's.

v) select K which gives highest accuracy.

Note: Make sure Test Dataset is not actually a subset of Training Dataset, this doesn't give the real-life accuracy.

→ Most Preferred value of K is 5, [if you want to avoid trail-error]

iii) → During KNN predictions, you may find that the distances computed may rely heavily on some specific attribute of dataset due to it having higher range of values.

→ This is happening because one of the attribute distance has higher magnitude, due to the attribute having higher range.

→ one way to solve this problem is to Scale all the attributes to a specific range to make the distance contribution of data point rely only on data itself and not on the attribute range variation

→ one way to do it is

$$x_{norm} = \frac{x - \mu}{\sigma}$$

where μ → mean of x's
σ → standard deviation of x's

→ This will "normalize" the attributes in range 0 to 1

Eg: Height Range: 1.5 to 1.8 m
Weight Range: 60 to 100 kg
income range: 10k to 200k
→ we normalize them to same range