K- nearest neighbors

Prachi M Joshi
Machine Learning Demystified

Eager vs Lazy learners

- Eager learners: When given training tuples, they create a generalization model before receiving new tuples to classify.

- Lazy learners are the ones who will simply store the training tuples and wait till a test tuple is given.

K-nearest neighbors

- Method introduced in early 50's, gained popularity later after more computation power was available.
- A lazy learner.
- Widely used in area of pattern recognition.
- The training tuples are stored. Sometimes pre-processed and stored.
- So, all the training tuples are stored in n-dimensional space, where n attributes exist.

K-nearest neighbors approach

- It works based on the minimum distance between the input instance/new unlabeled data to the training samples to determine the k-nearest neighbors.
- After the k-nearest neighbors are gathered we simply take the majority of these k-nearest neighbors to be the prediction.

Predict if a student will complete the assignment or no??

- Training tuples:

Name	Earlier marks	% attendance	Assignment status
Rucha	88	50	Yes
Mayur	87	54	No
Nilesh	90	56	No
Devika	90	55	Yes

- New instance

Dushyant	88	45	???

Calculate Closeness- distance

- For each training tuple calculate the distance between new instance to be classified.

- Generally Euclidean distance is used.

$$\boxed{d(X1, X2) = \sqrt{\sum_{i=1}^{n} (x_1 i - x_2 i)^2}}$$

- X1 and X2 are the tuples to be compared.
- x1i and x2i are the respective parameters of the tuples.

Predict completion of assignment

1. Calculate distance between
   - Dushyant and all the training tuples,
2. Sort the distances and determine the k-minimum distances. (K – the no. of neighbors to compare)
   - Assume that the k- value is selected as 3.

Tuple id	Distance	Result/ Class
Mayur	2	No
Nilesh	2	No
Rucha	5	Yes

3. If out of the 3 nearest neighbors, if 2 have not submitted the assignment, then the prediction is no.
4. So, Dushyant is not likely to submit the assignment.

How to select the k-value?

- A small value of k will means noise will have high influence on the result.
- A large value will make it computationally expensive.
- K is generally selected as odd no. if the no. of classes is 2.
- A simple approach to select k is $\boxed{k = \sqrt{n}}$
- Algorithms that are used for commercial purpose tend to use value of k as 10.

References

- Dunham, Margaret H. *Data mining: Introductory and advanced topics*. Pearson Education India, 2006.
- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.