What is K Nearest Neighbor

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

In data science the data that is provided to create a machine learning model can be of disparate type. They could be numeric, textual, or even pixel position and color. Various algorithms have been devised to deal with these types of data. However, in all of them, the goal is to predict or identify an outcome. In some cases, the outcome is already known. This outcome is known by what is called design of experiment, and the machine learning algorithm “estimates” an outcome of a newly introduces observation into the dataset. In this case, regression algorithms such as linear, or polynomial regression. There is one area that deals with “categorizing” or “classifying” the outcome. In this paper, we will cover K-Nearest Neighbors classifier which is a common method to predict the class of an observation.

Keywords:

KNN- K-Nearest Neighbors

What is K Nearest Neighbor

K-Nearest Neighbors

In linear regression models, we draw a scatter plot of a data set and try to fit the model such that the point in the scatter plot are closest to the line drawn in the plot. In this case, the mean squared error function is used to minimize the distance between observed values versus the predicted ones.

On the other hand, some datasets tend to exhibit a "clustering" behavior where clusters are seen in the graph. For example, consider a dataset that has data on all known viruses. In this case, the data will be clustered for each type of virus. Now, let us say we have data for an unknown virus and want to know how close it is to the known viruses. K-Nearest Neighbors classification algorithm predicts which class the new virus is closest to or even is.

The nearest neighbors' terms sound like the mean squared error functions, but what is K? The K is what is called a hyperparameter or a model tuning knob whose value affects the prediction. The value of K determines how many neighbors must an observation be close to in order to determine which class that observation is closest to. In out example the new data could show the virus is closets to the common cold, flu, or something new!

The distance between observations is calculated using the traditional Euclidean distance $d_{euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. *Manhattan* method replaces the square of distance with the absolute value in the Euclidean method:

$$d_{manhattan} = \sqrt{\sum_{i=1}^{n}|x_i - y_i|}$$

The *Minskowski* method introduces a variable *p* which can be adjusted to yield either Euclidean or Manhattan method of calculating the distance between two observations.

$$d_{minkowski} = \left(\sqrt{\sum_{i=1}^{n} |x_i - y_y|^p} \right)^{1/p}$$

, where with p = 1 yields Manhattan distance and p = 2 yields the Euclidean distance.

As far determining what value of k to use, we must consider that k could be as low as 1(with only one nearest neighbor) to the total number of observations *n*. With k=1 the bias would be low, but variance would be high. These are the two measures that balance the choice. With k= *n,* bias and variance would be inversely impacted.

The sklearn package offers GridSearchCV which allows us to run x-fold cross validation on the KNN classifier with different values of k. Once complete, the function returns the k that produces the best model.

Conclusion

Classification algorithms play an especially important role in machine learning. The applications span from natural language processing to image processing to sentiment analysis and beyond.

References

1. Introduction to K-means Clustering - https://blogs.oracle.com/datascience/introduction-to-k-means-clustering
2. Albon, Chris. Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (p. 196). O'Reilly Media. Kindle Edition.
3. StatQuest: K-nearest neighbors, Clearly Explained https://www.youtube.com/watch?v=HVXime0nQeI