Searching For Similarity

Bishal Neupane
Saugat Gyawali
Spencer Gray
Michael Stinnett

10/9/2022

CS 4375.003 - Introduction to Machine Learning - F22

How do kNN and decision trees work for regression?

kNN regression works in regression problems by finding the nearest neighbors' respective values. For example, we want to find the weight based on height and age. For person 11, it is found that the nearest neighbor is person 1 and 5. When k = 2, it selects persons 1 and 5. Similarly, when k = 3, it selects another person's value nearer to it by calculating Euclidean distance. After that, it finds the weight of the 3 known people and assumes the person's weight is the average of three people's weight. This is how kNN regression works. First, it fits the data using the training data. After that, it uses test data to find the average of known data by specifying the value of k on its own. We need to choose k in such a way that there is less error and put the optimal neighbor.

For decision trees, we specify a feature to become the root node. The leaf nodes are the final values. We want to minimize the RSS within each region, where y_r is the mean response for the training region.

We use a top-down, greedy approach is used to partition the data. To start, all predictors are examined to see if they make good splits and for each predictor, the numerical value at which split must be determined. Our first split will divide the data into different regions r1 and r2.

We will consider some predictor X1, which has split1 such that observations in r1 are less than the split value.

Finally, whenever the splitting threshold is reached, splitting process is stopped.

How do kNN and decision trees work for classification?

It is the statistical analysis method which predicts the output based on the prior observation of a data set. Logistic regression focuses on decreasing the loss function on each iteration using the concept of gradient descent and learning rate. It will adjust the value of w. It tries to minimize the loss as long as it can for the given data and output the log odd and this can be later converted to probability.

KNN Classification:
KNN is the machine learning algorithm which can be used for both regression and classification but I am going to focus on classification. It tries to classify different categories based on the distance. It tries to create the group of K data based on the euclidean distance or other distances.

Decision Tree:

This is the recursive, top-down, greedy algorithm used for classification. Decision tree works by classifying the features into two or more branches based on the features. Entropy and Gini index are used as the metric in decision trees.

How do the 3 clustering methods of step 3 work?

The three clustering methods are K-means, Hierarchical, and Model-Based. All of these methods use a form of Euclidean distance to determine how far apart observations are. In multi-dimensional space that is represented in most data sets, distance is not always an easy thing to understand. So that distance measure is pretty much what each algorithm sets out to optimize.

For K-means, the user establishes the number of centroids that will be identified and groups observations by the distance to that centroid. Deciding K centers, or centroids, is not a straightforward task, it takes understanding your data set intimately and lots of trial and error. The best K will minimize the within sum of squares. Even though Distance is what we call it in clustering, for K-means a better word might be variance. Essentially what is desired is to minimize variance between observations in clusters.

Hierarchical breaks away from relying on user defined centroid numbers. By defining 3 metrics called single, complete, and average linkage for distance. Hierarchical places all observation into its own cluster, then merges the closest two clusters. It repeats that until all clusters have merged into one.

Model Based assumes there is a good model to fit our data, applies likelihoods and bayesian principles to identify the best model and the best number of clusters. This mode makes use of Gaussian mixture models that are initialized by hierarchical clustering.

How PCA and LDA work, and why they might be useful techniques for machine learning?

Method of “reducing dimension” count. When creating new models, we can increase our accuracy by also increasing the number of parameters. This gives the model more factors to consider when trying to decide. However, that also means more relationships to calculate, further increasing the learning time of our algorithm.

The models we are working with only take a matter of minutes. But any model slightly more complex can increase that time to hours, days, or years. Hence the need to find a way to reduce our learning time, with less loss in accuracy. This is where PCA and LDA dimension reduction come into play. PCA and LDA are algorithms used to find correlations between various parameters and combine in some way, such as making a new axis using PCA along highest covariance. While LDA finds highest separation amongs different components. Reducing dimension count while maintaining the desired accuracy. Making its performance better than removing the parameter outright.

Reference:

Mazidi, Karen. “Machine Learning Handbook using R and Python”, www.karenmazidi.com. Accessed: October 07, 2022.