Umar Ali-Salaam

Cory Pekkala

Carolline Osei

Benjamin Frenkel

Searching for Similarity

a) how kNN and decision trees work for classification and regression

i) KNN regression is a non-parametric approach that approximates the relationship between independent variables and continuous outcomes by averaging data in the same neighborhood [1]. KNN works by calculating the distances between a query and all of the instances in the data, then picking the number of examples (K) closest to the query and voting for the most common label in the case of classification. The main distinction is that KNN regression uses a local average to estimate the value of the output variable. By computing the local probability, KNN classification attempts to forecast the class to which the output variable belongs.

ii) A classification tree splits a dataset based on similarities. Assume x and y are the two factors that determine whether or not a choice is taken. If the training data indicates that the majority of the decisions are based on x, the data is separated and the x variable is elevated to the top node of the tree. This division makes a particular fraction of the data 'pure.' Impurity measurements are used to quantify data homogeneity in classification trees. Each independent variable in a regression tree is used to fit a regression model to the target variable. For each independent variable, the data is then separated into multiple points. A regression model is fit to the target variable using each of the independent variables in a regression tree. The data is then divided into many points for each independent variable.

b) how the 3 clustering methods of step 3 work

i) For K-means Clustering, first you pick a random number of data points to be your starting centroids or “k”. We chose to do the elbow method to determine which “k” for k-means to choose, but there are others After you got your centroids initialized, you look at the other data points on the scatter plot and calculate the distance between the data point and the nearest centroid; assign that data point to its nearest centroid. Each time you assign data points to a centroid, the more you have to move the centroid around, by averaging all of the centroid’s assigned data points, and then that is your new centroid. You just repeat adding data points and moving centroids around until you get through all of the data.

ii) For Hierarchical Clustering, first you need to make sure your data set size isn’t too big, sampling of a large data set may have to be cut down to a sample. After that is sorted you need to compute a proximity matrix of the data points, using some sort of distance measurement tool. Then data points are assigned to their clusters, and are merged based on a measurement for similarity between the clusters. Then you update the matrix. Repeat the merging and updating of the clusters and the matrix, until you get one giant cluster.

iii) For Model-Based Clustering, first you need to assign k clusters randomly. Then by the expectation maximization algorithm or you can just use the mclust package in R to classify observations and continuously update the clusters on a scatter plot graph. Just like K-means Clustering you update the cluster means and variances, until you run out of data points.

c) how PCA and LDA work, and why they might be useful techniques for machine learning

i) PCA is used to help visualize multidimensional data, simplify complex decisions, and reduce dimensionality of data. One way to calculate PCA is to standardize the mean and variances of the features in the data. Then you obtain the matrix covariation of your data, and calculate the eigendecomposition of the matrix. Sort each eigenvalue from highest to

lowest, and then select the top N eigenvectors to become the principal components.

ii) LDA is used to reduce dimensionality for supervised classification problems in machine learning. LDA first estimates the probability that each data point should be in its respective classification. The class with the highest probability is then selected to be used for predictions.

iii) Both PCA and LDA are great for helping to reduce the dimensionality of data, and simplifying features to make it more easily understandable. They make the difficulty of classifying data much easier.

[Sources]

[1] Harrison, Onel. “Machine Learning Basics with the K-Nearest Neighbors Algorithm.” Medium, Towards Data Science, 14 July 2019, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.