Abigail Smith

ARS190011

CS 4375.004

Dr. Mazidi

Searching for Similarity, Narrative

In machine learning, kNN and decision trees exist as two similarity models used for classification and regression. Decision trees and kNN provide ways for machine learning applications to classify observations based on partitions (decision trees) and neighbors (kNN) through a provided training set. Both kNN and decision trees provide valuable ways in which to evaluate and classify data.

The kNN algorithm stands as an instance-based learning model used for both classification and regression. At its basis, kNN simply stores all training data in memory and uses said training observations to label observed, or current, observations against. As in the name, k in kNN represents the number of neighbors, or nearby training observations, used to classify said current observation. The value assigned to k has a critical impact on the bias-variance trade off in kNN. The smaller the k, the lower the bias and higher the variance, and as k increases, the model becomes less flexible and the vice-verse for bias-variance occurs. Practices exist to select an optimal, reliable k which include using cross-validation and ensuring k is an odd number. Overall, kNN depends on a predefined k-value to classify new observations against neighboring training observations.

Similar to kNN, decision trees work for both classification and regression by partitioning provided data into distinct regions. Decision trees work by iteratively dividing provided observations into partitions with linear boundaries until the observations in each partition have

uniform distribution. Like kNN, decision trees require a predefined value representing the number of partitions before execution. When performing regression with decision trees, one of the main goals is to minimize the RSS within each region (i.e. aim for uniform distribution between the regions). In contrast, decision trees perform classification by dividing qualitative data into branches where distinct factors have been assigned — meaning a non-binary data set would have one or more of its values assigned to a branch and the remaining to another branch.

Each of the three clustering methods — k-means, hierarchical, and model-based clustering — presents a unique way to perform clustering on a data set. While k-means and hierarchical act as heuristic approaches to clustering that create their own clusters from a provided data set, model-based clustering works by "considering the data as coming from a distribution that is [a] mixture of two or more clusters" (Kassambara).

As an iterative approach, k-mean clustering identifies k number of centroids (or centers) where observations are then grouped depending on their closeness to these centroids. While k-means has a variety of approaches, EM represents one of the most famous approaches where each observation is assigned its group based on the closest centroid (step 'E' in EM) and then the centroids are recomputed (step 'M' in EM). While k-mean is a recursive approach that consistently re-evaluates its centers, an optimal k must be found through execution and experimentation.

In contrast, hierarchical clustering does not depend on a predefined number of clusters. Instead, hierarchical clustering utilizes a defined distance measurement to group instances into clusters organized by hierarchy (thus the name "hierarchical clustering"). These clusterings, like in k-mean clustering, have a variety of approaches; however, generally hierarchical clustering is completed in four steps: place each observation into a unique cluster, find the distance between

each cluster and every other cluster, merge the two nearest clusters, and finally repeat these steps until all individual clusterings combine into one cluster. As a result, hierarchical clusters produce a dendrogram of the clusters.

Lastly, model-based clustering works by treating the provided data as though it were a combination or multiple clusters. In doing so, model-based clustering uses probability to determine whether or not an observation belongs to a cluster. This assignment process makes model-based clustering stand out from k-means and hierarchical clustering as it strays away from a heuristic approach and instead embraces “a soft assignment” in which “each data point has a probability of belonging to each cluster” (Kassambara). Overall, each of these methods -- k-means, hierarchical, and model-based clustering -- provide unique ways to perform clustering in machine learning.

When performing data reduction, PCA (Principal Components Analysis) and LDA (Linear Discriminant Analysis) exist as two unique techniques that reduce the dimensionality of a dataset. PCA works by converting provided data into a new coordinate space and simultaneously lowering the amount of axes. In doing so, the first PC acts as the dimensions of greatest variance while the remaining PCs stand for the decreasing variance. Since PCA works without class, it works like other unsupervised methods. In contrast, LDA works by finding a linear combination of the predictors in order to maximize the separation of the classes all while reducing the in-class standard deviation. Unlike PCA, LDA utilizes class and, as such, will have an advantage over PCA when the class is known. Generally, PCA and LDA exist as two distinct approaches to data reduction, each with their own processes and requirements.

Works Cited

Kassambara, Alboukadel. “Model Based Clustering Essentials.” *DataNovia*, https://www.datanovia.com/en/lessons/model-based-clustering-essentials/.

Mazidi, Karen. Machine Learning Handbook Using R and Python. 2nd ed., 2020.