Cluster Analysis

Cluster Analysis is a type of unsupervised Analysis which is used to classify the cases based on their similarity using their features. As the result, the objects in the same group (called a cluster) are more similar to each other than to those in other clusters.

Dissimilarity

In clustering Analysis, as explained above, similar cases are clustered in same group. Similarity of cases is measures based on the distance between each two cases which is the measure of dissimilarity.

Multidimensional Scaling

Multidimensional Scaling is a tool for visualizing the level of similarity of individual cases of a dataset. MDS is used to translate information about the pairwise distances among a set of n objects or individuals into a configuration of n points mapped into an abstract Cartesian space.

Rejection sampling

Rejection sampling is a method of generating samples from distribution *f* using another distribution *g*, which is used to define the envelop function h. The samples then are generated in 2 dimensions (i.e. (x,u)) from *g* and a uniform distribution between (0, y) where y is h(x). The samples with u > f(x) are rejected and the remaining sample would have the f distribution.

Ensemble

Ensemble is a method that combines several decision trees to produce better predictive performance than utilizing a single decision tree. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner.

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Factor Analysis is a method of Data Reduction.

Varimax Rotation

The sub-space found with principal component analysis is expressed as a basis with many non-zero weights which makes it hard to interpret. Varimax rotation maximizes the sum of the variances of the squared loadings.

Missing Completely At Random (MCAR)

A variable is missing completely at random if the probability of missingness is the same for all units, for example, if each survey respondent decides whether to answer the “earnings” question by rolling a die and refusing to answer if a “6” shows up. If data are missing completely at random, then throwing out cases with missing data does not bias your inferences.

Missing Not At Random (MNAR)

A missing value that is neither MAR nor MNAR is missing not at random. In particular, this type of missing is because of the value of the missing values itself, or because of a hidden variable.

Missing At Random (MAR)

A general assumption for missing at random is that the probability of missingness in a variable for a case depends only on the information provided by other variables. For instance, if sex, race, education, and age are recorded for all the people in the survey, then “earnings” is missing at random if the probability of nonresponse to this question depends only on these other recorded variables. It would therefore be reasonable to model this type of missingness using a logistic regression, where the outcome variable equals 1 for observed cases and 0 for missings.

Random forest

Random Forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees.

Bagging

(Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.

Cross Validation

Cross validation is a technique for model selection. A proper model is the one that performs well on a dataset that was not seen before. In other words, if a model performs well on the training dataset, this is not a guarantee to be a good model. It would therefore be an appropriate approach to use multiple training datasets to end up to the model that works well with various datasets. We therefore: 1) Create multiple training datasets, which all are a subset of the larger dataset. 2) Keep some data aside to check the performance of the model. These two steps together is what is called Cross Validation.