Comparison of Data Mining Algorithms

Classification: Predicting a Categorical Target Variable

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Cases
Decision Trees	Partitions the data into smaller subsets where each subset contains (mostly) responses of one class (either "yes" or "no")	A set of rules to partition a data set based on the values of the different predictors.	No restrictions on variable type for predictors.	The label cannot be numeric. It must be categorical.	Intuitive to explain to nontechnical business users. Normalizing predictors is not necessary.	Tends to overfit the data. Small changes in input data can yield substantially different trees. Selecting the right parameters can be challenging.	Marketing segmentation, fraud detection.
Rule Induction	Models the relationship between input and output by deducing simple IF/THEN rules from a data set.	A set of organized rules that contain an antecedent (inputs) and consequent (output class).	No restrictions. Accepts categorical, numeric, and binary inputs.	Prediction of target variable, which is categorical.	Model can be easily explained to business users. Easy to deploy in almost any tools and applications.	Divides the data set in rectilinear fashion.	Manufacturing, applications where description of model is necessary.
k-Nearest Neighbors	A lazy learner where no model is generalized. Any new unknown data point is compared against similar known data point in the training set.	Entire training data set is the model.	No restrictions. However, the distance calculations work better with numeric data. Data need to be normalized.	Prediction of target variable, which is categorical.	Requires very little time to build the model. Handles missing attributes in the unknown record gracefully. Works with nonlinear relationships.	The deployment runtime and storage requirements will be expensive. Arbitrary selection of value of k. No description of the model.	Image processing, applications where slower response time is acceptable.

Naïve Bayesian	Predicts the output class based on Bayes' theorem by calculating class conditional probability and prior probability.	A lookup table of probabilities and conditional probabilities for each attribute with an output class.	No restrictions. However, the probability calculation works better with categorical attributes	Prediction of probability for all class values, along with the winning class.	Time required to model and deploy is minimum. Great algorithm for benchmarking. Strong statistical foundation.	Training data set needs to be representative sample of population and needs to have complete combinations of input and output. Attributes need to be independent.	Spam detections, text mining.
Artificial Neural Networks	A computational and mathematical model inspired by the biological nervous system. The weights in the network learn to reduce the error between actual and prediction.	A network topology of layers and weights to process input data.	All attributes should be numeric.	Prediction of target (label) variable, which is categorical.	Good at modeling nonlinear relationships. Fast response time in deployment.	No easy way to explain the inner working of the model. Requires preprocessing data. Cannot handle missing attributes.	Image recognition, fraud detection, quick response time applications.

Continued

Classification: Predicting a Categorical Target Variable *Continued*

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Cases
Support Vector Machines	Essentially a boundary detection algorithm that identifies/defines multi-dimensional boundaries separating data points belonging to different classes.	The model is a vector equation that allows us to classify new data points into different regions (classes).	All attributes should be numeric.	Prediction of target (label) variable, which can be categorical or numeric.	Very robust against overfitting. Small changes to input data do not affect boundary and thus do not yield different results. Good at handling nonlinear relationships.	Computational performance during training phase can be slow. This may be compounded by the effort needed to optimize parameter combinations.	Optical character recognition, fraud detection, modeling "black-swan" events.
Ensemble Models	Leverages wisdom of the crowd. Employs a number of independent models to make a prediction and aggregates the final prediction.	A meta-model with individual base models and a aggregator.	Superset of restrictions from the base model used.	Prediction for all class values with a winning class.	Reduces the generalization error.Takes different search space into consideration	Achieving model independence is tricky. Difficult to explain the inner working of the model.	Most of the practical classifiers are ensemble.

Regression: Predicting a Numeric Target Variable

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Case
Linear Regression	The classical predictive model that expresses the relationship between inputs and an output parameter in the form of an equation.	The model consists of coefficients for each input predictor and their statistical significance. A bias (intercept) may be optional.	All attributes should be numeric.	The label may be numeric or binominal.	The workhorse of most predictive modeling techniques. Easy to use and explain to non-technical business users.	Cannot handle missing data. Categorical data are not directly usable, but require transformation into numeric.	Pretty much any scenario that requires predicting a continuous numeric value.
Logistic Regression	Technically, this is a classification method. But structurally it is similar to linear regression.	The model consists of coefficients for each input predictor that relate to the "logit." Transforming the logit into probabilities of occurrence (of each class) completes the model.	All attributes should be numeric.	The label may only be binominal.	One of the most common classification methods. Computationally efficient.	Cannot handle missing data. Not very intuitive when dealing with a large number of predictors.	Marketing scenarios (e.g., will click or not click), any general two-class problem.

Association Analysis: Unsupervised Process for Finding Relationships between Items

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Case
FP Growth and Apriori	Measures the strength of co-occurrence between one item with another.	Finds simple, easy to understand rules like {Milk, Diaper} -> {Beer}	Trans-actions format with items in the columns and transactions in the rows.	List of relevant rules developed from the data set	Unsupervised approach with minimal user inputs. Easy to understand rules.	Requires prepro-cessing if input is of different format.	Recom-mendation engines, cross-selling, and content suggestions.

Clustering: An Unsupervised Process for Finding Meaningful Groups in Data

Algorithm	Description	Model	Input	Output	Pros	Cons	Use case
k-means	Data set is divided into k clusters by finding k centroids.	Algorithm find k centriods and all the data points are assigned to the nearest centriods, which form a cluster.	No restrictions. However, the distance calculations work better with numeric data. Data should be normalized.	Data set is appended by One of k cluster labels.	Simple to implement. Can be used for dimension reduction.	Specification of k is arbitrary and may not find natural clusters. Sensitive to outliers.	Customer segmentation, anomaly detection, applications where globular clustering is natural.
DBSCAN	Identifies clusters as a high-density area surrounded by low-density areas.	List of clusters and assigned data points. Default Cluster 0 contains noise points.	No restrictions. However, the distance calculations work better with numeric data. Data should be normalized.	Cluster labels based on identified clusters.	Finds the natural clusters of any shape. No need to mention number of clusters.	Specification of density parameters. A bridge between two clusters can merge the cluster. Can not cluster varying density data set.	Applications where clusters are nonglobular shapes and when the prior number of natural groupings is not known.
Self-Organizing Maps	A visual clustering technique with roots from neural networks and prototype clustering.	A two-dimensional lattice where similar data points are arranged next to each other.	No restrictions. However, the distance calculations work better with numeric data. Data should be normalized.	No explicit clusters identified. Similar data points occupy either the same cell or are placed next to each other in the neighborhood.	A visual way to explain the clusters. Reduces multidimensional data to two dimensions.	Number of centriods (topology) is specified by the user. Does not find natural clusters in the data.	Diverse applications including visual data exploration, content suggestions, and dimension reduction.

Anomaly Detection: Supervised and Unsupervised Techniques to Find Outliers in the Data

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Case
Distance Based	Outlier identified based on distance if kth nearest neighbor.	All data points are assigned a distance score based on nearest neighbor.	Accepts both numeric and categorical attributes. Normalization is required since distance is calculated.	Every data point has a distance score. The higher the distance, the more likely the data point is an outlier.	Easy to implement. Works well with numeric attributes.	Specification of k is arbitrary.	Fraud detection, preprocessing technique.
Density Based	Outlier is identified based on data points in low-density regions.	All data points as assigned a density score based on the neighborhood.	Accepts both numeric and categorical attributes. Normalization is required since density is calculated.	Every data point has a density score. The lower the density, the more likely the data point is an outlier.	Easy to implement. Works well with numeric attributes.	Specification of distance parameter by the user. Inability to identify varying density regions.	Fraud detection, preprocessing technique.
Local outlier factor	Outlier is identified based on calculation of relative density in the neighborhood.	All data points as assigned a relative density score based on the neighborhood.	Accepts both numeric and categorical attributes. Normalization is required since density is calculated.	Every data point has a density score. The lower the relative density, the more likely the data point is an outlier	Can handle the varying density scenario.	Specification of distance parameter by the user.	Fraud detection, preprocessing technique.

Feature Selection: Selection of Most Important Attributes

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Case
PCA (Filter Based)	PCA is in reality a dimension reduction method. It combines the most important attributes into a fewer number of transformed attributes.	N/A	Numerical attributes	Numerical attributes (reduced set). Does not really require a label.	Efficient way to extract predictors that are uncorrelated to each other. Helps to apply Pareto principle in identifying attributes with highest variance.	Very sensitive to scaling effects, i.e., requires normalization of attribute values before application. Focus on variance sometimes results in selecting noisy attributes.	Most numeric-valued data sets that require dimension reduction.
Info Gain (Filter Based)	Selecting attributes based on relevance to the target or label.	Similar to decision tree model.	No restrictions on variable type for predictors.	Data sets require a label. Can only be applied on data sets with nominal label.	Same as decision trees.	Same as decision trees.	Applications for feature selection where target variable is categorical or numeric.
Chi-Square (Filter Based)	Selecting attributes based on relevance to the target or label.	Uses the chi-square test of independence to relate predictors to label.	Categorical (polynominal) attributes	Data sets require a label. Can only be applied on data sets with a nominal label.	Very robust. A fast and efficient scheme to identify which categorical variables to select for a predictive model.	Sometimes difficult to interpret.	Applications for feature selection where all variables are categorical.

Continued

Feature Selection: Selection of Most Important Attributes *Continued*

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Case
Forward Selection (Wrapper Based)	Selecting attributes based on relevance to the target or label.	Works in conjunction with modeling methods such as regression.	All attributes should be numeric.	The label may be numeric or binominal	Multicollinearity problems can be avoided. Speeds up the training phase of the modeling process	Once a variable is added to the set, it is never removed in subsequent iterations even if its influence on the target diminishes.	Data sets with a large number of input variables where feature selection is required.
Backward Elimination (Wrapper Based)	Selecting attributes based on relevance to the target or label.	Works in conjunction with modeling methods such as regression.	All attributes should be numeric.	The label may be numeric or binominal.	Multicollinearity problems can be avoided. Speeds up the training phase of the modeling process.	Need to begin with a full model, which can sometimes be computationally intensive.	Data sets with few input variables where feature selection is required.