Cheat Sheet: Machine Learning with KNIME Analytics Platform

Open for Innovation KNIME

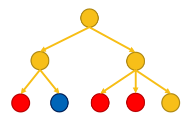
SUPERVISED LEARNING

Supervised Learning: A set of machine learning algorithms to predict the value of a target class or variable. They produce a mapping function (model) from the input features to the target class/variable. To estimate the model parameters during the training phase, labeled example data are needed in the training set. Generalization to unseen data is evaluated on the test set data via scoring metrics.

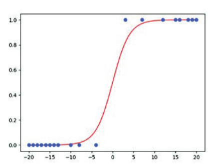
CLASSIFICATION

Classification: A type of supervised learning where the target is a class. The model learns to produce a class score and to assign each vector of input features to the class with the highest score. A cost can be introduced to penalize one of the classes during class assignment.

Decision Tree: Follows the C4.5 decision tree algorithm. These algorithms generate a tree-like structure, creating data subsets, aka tree nodes. At each node, the data are split based on one of the input features, generating two or more branches as output. Further splits are made in subsequent nodes until a node is generated where all or almost all of the data belong to the same class.



Logistic Regression: A statistical algorithm that models the relationship between the input features and the categorical output classes by maximizing a likelihood function. Originally developed for binary problems, it has been extended to problems with more than two classes (multinomial logistic regression).



Naive Bayes Learner

Naive Bayes: Based on Bayes' theorem and assuming statistical independence between input features (thus "naive"), this algorithm estimates the conditional probability of each output class given the vector of input features. The class with the highest conditional probability is assigned to the input data.

SVM Learner

Support Vector Machine (SVM): A supervised algorithm constructing a set of discriminative hyperplanes in high-dimensional space. In addition to linear classification, SVMs can perform non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces, where the two classes are linearly separable.

K Nearest Neighbor

k-Nearest Neighbor (kNN): A non-parametric method that assigns the class of the k most similar points in the training data, based on a pre-defined distance measure. Class attribution can be weighted by the distance to the k-th point and/or by the class probability.

NUMERIC PREDICTION & CLASSIFICATION

RProp MLP Learner

Artificial Neural Networks (ANN, NN): Inspired by biological nervous systems, Artificial Neural Networks are based on architectures of interconnected units called artificial neurons. Artificial neurons' parameters and connections are trained via dedicated algorithms, the most popular being the Back-Propagation algorithm.

Keras Dense Layer

Deep Learning: Deep learning extends the family of ANNs with deeper architectures and additional paradigms, e.g. Recurrent Neural Networks (RNN). The training of such networks, has been enabled by recent advances in hardware performance as well as parallel execution.

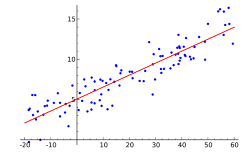
H2O Generalized Linear Model Learner

Generalized Linear Model (GLM): A statistics-based flexible generalization of ordinary linear regression, valid also for non-normal distributions of the target variable. GLM uses the linear combination of the input features to model an arbitrary function of the target variable (the link function) rather than the target variable itself.

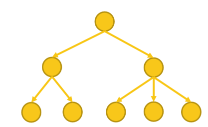
NUMERIC PREDICTION

Numeric Prediction: A type of supervised learning for numeric target variables. The model learns to associate one or more numbers with the vector of input features. Note that numeric prediction models can also be trained to predict class scores and therefore can be used for classification problems too.

Linear/Polynomial Regression: Linear Regression is a statistical algorithm to model a multivariate linear relationship between the numeric target variable and the input features. Polynomial Regression extends this concept to fitting a polynomial function of a pre-defined degree.



Regression Tree: Builds a decision tree to predict numeric values through a recursive, top-down, greedy approach known as recursive binary splitting. At each step, the algorithm splits the subsets represented by each node into two or more new branches using a greedy search for the best split. The average value of the points in a leaf produces the numerical prediction.



TIME SERIES ANALYSIS

Time Series Analysis: A set of numeric prediction methods to analyze/predict time series data. Time series are time ordered sequences of numeric values. In particular, time series forecasting aims at predicting future values based on previously observed values.

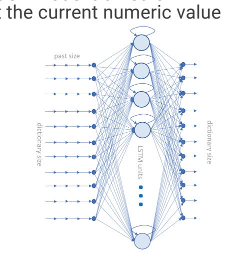
SARIMA Learner

Seasonal Auto-Regressive Integrated Moving Average (SARIMA): A linear Seasonal (S) Auto-Regressive (AR) model is constructed on a specified number p of past (seasonal) values; data are prepared by a degree of differencing d to correct non-stationarity; while a linear combination - named Moving Average (MA) - models the q past (seasonal) residual errors. All SARIMA model parameters are estimated concurrently by various algorithms, mostly following the Box–Jenkins approach.

Lag Column · Linear Regression Learner

ML-based TSA: A numeric prediction model trained on vectors of past values can predict the current numeric value of the time series.

Long Short Term Memory (LSTM) Units: LSTM units produce a hidden state by processing m x n tensors of input values, where m is the size of the input vector at any time and n the number of past vectors. The hidden state can then be transformed into the current vector of numeric values. LSTM units are suited for time series prediction as values from past vectors can be remembered or forgotten through a series of gates.

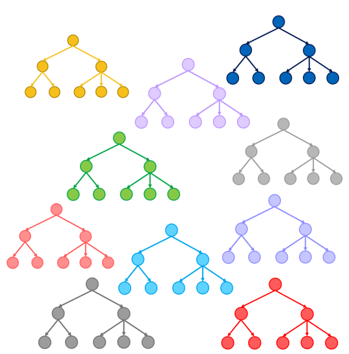


ENSEMBLE LEARNING

Ensemble Learning: A combination of multiple models from supervised learning algorithms to obtain a more stable and accurate overall model. Most commonly used ensemble techniques are Bagging and Boosting.

BAGGING

Bagging: A method for training multiple classification/regression models on different randomly drawn subsets of the training data. The final prediction is based on the predictions provided by all the models, thus reducing the chance of overfitting.



Tree Ensemble of Decision/Regression Trees: Ensemble model of multiple decision/regression trees trained on different subsets of data. Data subsets with less or equal rows and less or equal columns are bootstrapped from the original training set. Final prediction is based on a hard vote (majority rule) or soft-vote (averaging all probabilities or numeric predictions) on all involved trees.

Random Forest of Decision/Regression Trees: Ensemble model of multiple decision/regression trees trained on different subsets of data. Data subsets with the same number of rows are bootstrapped from the original training set. At each node, the split is performed on a subset of sqrt(x) features from the original x input features. Final prediction is based on a hard vote (majority rule) or soft-vote (averaging all probabilities or numerical predictions) on all involved trees.

BOOSTING

Boosting: A method for training a set of classification/regression models iteratively. At each step, a new model is trained on the prediction errors and added to the ensemble to improve the results from the previous model state, leading to higher accuracy after each iteration.

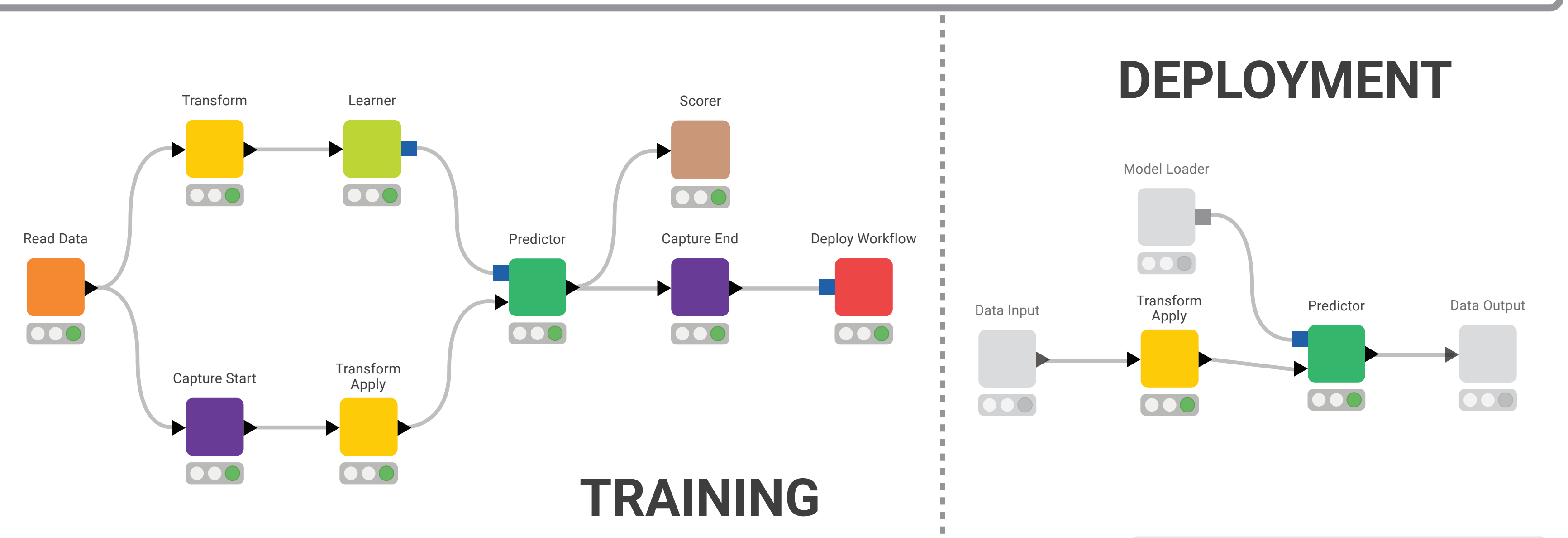
Gradient Boosted Regression Trees: Ensemble model combining multiple sequential simple regression trees into a stronger model. The algorithm builds the model stagewise. At each iteration, a simple regression tree is fitted to predict the residuals of the current model, following the gradient of the loss function. This leads to an increasingly accurate and complex overall model. The same regression trees can also be used for classification.

Prediction Fusion

Custom Ensemble Model: Combining different supervised models to form a custom ensemble model. The final prediction can be based on majority vote as well as on the average or other functions of the output results.

XGBoost Tree Ensemble Learner

XGBoost: An optimized distributed library for machine learning models in the gradient boosting framework, designed to be highly efficient, flexible, and portable. It features regularization parameters to penalize complex models, effective handling of sparse data for better performance, parallel computation, and more efficient memory usage.



EVALUATION

Evaluation: Various scoring metrics for assessing model quality - in particular, a model's predictive ability or propensity to error.

Scorer

Confusion Matrix: A representation of a classification task's success through the count of matches and mismatches between the actual and predicted classes, aka true positives, false negatives, false positives, and true negatives. One class is arbitrarily selected as the positive class.

Accuracy Measures: Evaluation metrics for a classification model calculated from the values in the confusion matrix, such as sensitivity and specificity, precision and recall, or overall accuracy.

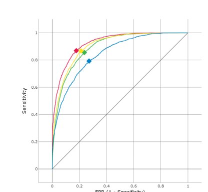
Cross Validation

Cross-Validation: A model validation technique for assessing how the results of a machine learning model will generalize to an independent data set. A model is trained and validated N times on different pairs of training and test set, both extracted from the original data. Some basic statistics on the resulting N error or accuracy measures gives insights on overfitting and generalization.

Numeric Scorer

Numeric Error Measures: Evaluation metrics for numeric prediction models quantifying the error size and direction. Common metrics include RMSE, MAE, or R^2. Most of these metrics depend on the range of the target variable.

ROC Curve: A graphical representation of the performance of a binary classification model with false positive rates on the x-axis and true positive rates on the y axis. Multiple points for the curve are obtained for different classification thresholds. The area under the curve is the metric value.



Resources

- **E-Books:** Learn even more with the KNIME books. From basic concepts in "KNIME Beginner's Luck", to advanced concepts in "KNIME Advanced Luck", through to examples of real-world case studies in "Practicing Data Science". Available for purchase at knime.com/knimepress
- **KNIME Blog:** Engaging topics, challenges, industry news, and knowledge nuggets at knime.com/blog
- **KNIME Hub:** Search, share, and collaborate on KNIME workflows, nodes, and components with the entire KNIME community at hub.knime.com
- **KNIME Forum:** Join our global community and engage in conversations at forum.knime.com
- **KNIME Server:** The enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services. Visit www.knime.com/knime-server for more information.

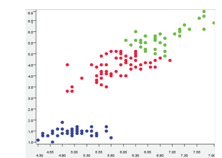
UNSUPERVISED LEARNING

Unsupervised Learning: A set of machine learning algorithms to discover patterns in the data. A labeled dataset is not required, since data are ultimately organized and/or categorized based on similarity or statistical measures.

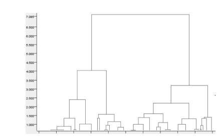
CLUSTERING

Clustering: A branch of unsupervised learning algorithms that groups data together based on similarity measures, without the help of labels, classes, or categories.

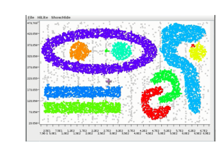
k-Means: The n data points in the dataset are clustered into k clusters based on the shortest distance from the cluster prototypes. The cluster prototype is taken as the average data point in the cluster.



Hierarchical Clustering: Builds a hierarchy of clusters by either collecting the most similar (agglomerative approach) or separating the most dissimilar (divisive a pproach) data points and clusters, according to a selected distance measure. The result is a dendrogram clustering the data together bottom-up (agglomerative) or separating the data in different clusters top-down (divisive).



DBSCAN: A density-based non-parametric clustering algorithm. Data points are classified as core, density-reachable, and outlier points. Core and density-reachable points in high density regions are clustered together, while points with no close neighbors in low-density regions are labeled as outliers.



SOTA Learner

Self-Organizing Tree Algorithm (SOTA): A special Self-Organizing Map (SOM) neural network. Its cell structure is grown using a binary tree topology.

Fuzzy c-Means

Fuzzy c-Means: One of the most widely used fuzzy clustering algorithms. It works similarly to the k-Means algorithm, but it allows for data points to belong to more than one cluster, with different degrees of membership.

RECOMMENDATION ENGINES

Recommendation Engines: A set of algorithms that use known information about user preferences to predict items of interest.

Association Rule Learner

Association Rules: The node reveals regularities in co-occurrences of multiple products in large-scale transaction data recorded at point-of-sale. Based on the a-priori algorithm, the most frequent itemsets in the dataset are used to generate recommendation rules.

Spark Collaborative Filtering Learner (MLlib)

Collaborative Filtering: Based on the Alternating Least Squares (ALS) technique, it produces recommendations (filtering) about the interests of a user by comparing their current preferences with those of multiple users (collaborating).

KNIME Press

Open for Innovation
KNIME

Extend your KNIME knowledge with our collection of books from KNIME Press. For beginner and advanced users, through to those interested in specialty topics such as topic detection, data blending, and classic solutions to common use cases using KNIME Analytics Platform - there's something for everyone. Available for download at www.knime.com/knimepress.