Machine Learning - (One|Simple) Rule - (One Level Decision Tree)

You are here : (https://gerardnico.com/wiki/start)

Table of Contents

1 - About

One Rule is an simple method based on a 1-level decision tree (https://gerardnico.com/wiki/data_mining/decision_tree) described in 1993 by Rob Holte, Alberta, Canada.

Simple rules often outperformed far more complex methods because some datasets are :

- really simple
- so small/noisy/complex that nothing can be learned from them

2 - Articles Related

- Data Mining - (Classifier|Classification Function) (https://gerardnico.com/wiki/data_mining/classification)
- Statistics - (Confidence|likelihood) (Prediction probabilities|Probability classification) (https://gerardnico.com/wiki/data_mining/confidence)
- Data Mining - Decision Tree (DT) Algorithm (https://gerardnico.com/wiki/data_mining/decision_tree)
- Machine Learning - Decision Stump (https://gerardnico.com/wiki/data_mining/decisionstump)
- Machine Learning - Linear (Regression|Model) (https://gerardnico.com/wiki/data_mining/linear_regression)
- Data Mining - Naive Bayes (NB) (https://gerardnico.com/wiki/data_mining/naive_bayes)
- Data Mining - (Decision) Rule (https://gerardnico.com/wiki/data_mining/rule)

3 - Implementation

3.1 - Basic

- One branch for each value
- Each branch assigns most frequent class
- Error rate: proportion of instances that don't belong to the majority class of their corresponding branch
- Choose attribute with smallest error rate

```
For each attribute,
   For each value of the attribute,
   make a rule as follows:
       count how often each class appears
       find the most frequent class
       make the rule assign that class to this attribute-value
   Calculate the error rate of this attribute's rules
Choose the attribute with the smallest error rate
```

Example of output for the weather data set (https://gerardnico.com/wiki/data_mining/weather)

```
outlook:
    if sunny    -> no
    if overcast -> yes
    if rainy    -> yes
```

with this one-level decision tree, 10 instances are correct on 14.

3.2 - Other

Algorithm to choose the best rule

```
For each attribute:
  For each value of that attribute, create a rule:
      1. count how often each class appears
      2. find the most frequent class, c
      3. make a rule "if attribute=value then class=c"
  Calculate the error rate of this rule
Pick the attribute whose rules produce the lowest error rate
```

(Statistics|Probability|Machine Learning|Data Mining|Data and Knowledge Discovery|Pattern Recognition|Data Science|Data Analysis) (https://gerardnico.com/wiki/data_mining/start)
326 pages

The 1 Percent Rule (https://gerardnico.com/wiki/data_mining/1_percent)

A/B (Test|Testing) (https://gerardnico.com/wiki/data_mining/a_b)

(Parameters|Model) (Accuracy|Precision|Fit|Performance) Metrics (https://gerardnico.com/wiki/data_mining/accuracy)

Adjusted R^2 (https://gerardnico.com/wiki/data_mining/adjusted_r_squared)

Akaike information criterion (AIC) (https://gerardnico.com/wiki/data_mining/aic)

Algorithms (https://gerardnico.com/wiki/data_mining/algorithm)

(Anomaly|outlier) Detection (https://gerardnico.com/wiki/data_mining/anomaly_detection)

Analysis of variance (Anova) (https://gerardnico.com/wiki/data_mining/anova)

Apriori algorithm (https://gerardnico.com/wiki/data_mining/apriori)

Association (Rules Function|Model) - Market Basket Analysis (https://gerardnico.com/wiki/data_mining/association)

Sample (Variable|Attribute) (https://gerardnico.com/wiki/data_mining/attribute)

Attribute (Importance|Selection) - Affinity Analysis (https://gerardnico.com/wiki/data_mining/attribute_importance)

Area under the curve (AUC) (https://gerardnico.com/wiki/data_mining/auc)

Automatic Discovery (https://gerardnico.com/wiki/data_mining

4 - One Rule vs Baseline

OneR always outperforms (or, at worst, equals) Baseline (https://gerardnico.com/wiki/data_mining/baseline) when evaluated on the training data. (evaluating on the training data doesn't reflect performance on independent test data.)

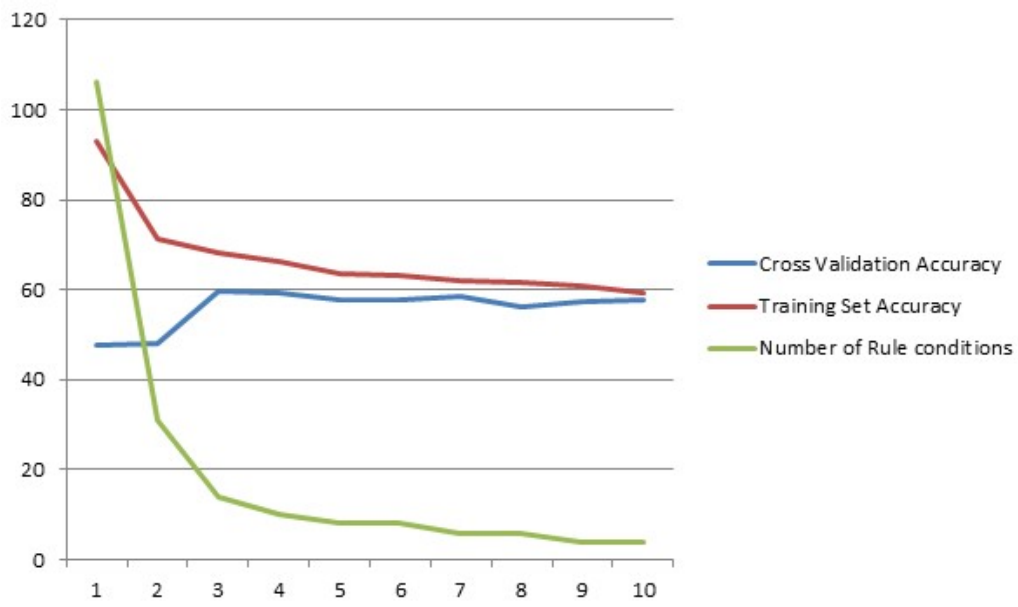
ZeroR (https://gerardnico.com/wiki/data_mining/baseline) sometimes outperforms OneR if the target (https://gerardnico.com/wiki/data_mining/target) distribution is skewed or limited data is available, predicting the majority class can yield better results than basing a rule on a single attribute. This happens with the nominal weather dataset (https://gerardnico.com/wiki/data_mining/weather)

5 - minBucket Size

The “minBucket size” parameter of weka (https://gerardnico.com/wiki/data_mining/weka) limits the complexity of rules in order to avoid overfitting (https://gerardnico.com/wiki/data_mining/overfitting) (Default 6)

With one “minBucket size” the accuracy on the training data set is really high and decreases whereas the “minBucket size parameter” increases.

The cross validation (https://gerardnico.com/wiki/data_mining/cross_validation) evaluation method (10 folders) limits the accuracy effect and make it more stable through the “minBucket size” values.



(https://gerardnico.com/wiki/_detail/data_mining/one_r_graph.jpg?id=data_mining%3Aone_rule)

min Bucket Size Parameter	Eval Method: Cross Valid-ation Accuracy	Eval Method: Training Set Accuracy	Number of conditions generated
1	47.66	92.99	106
2	48.13	71.5	31
3	59.81	68.22	14
4	59.35	66.36	10
5	57.94	63.55	8
6	57.94	63.08	8
7	58.41	62.14	6
8	56.07	61.68	6
9	57.48	60.75	4
10	57.94	59.34	4

/automatic_discovery)

Bootstrap aggregating (bagging) (https://gerardnico.com/wiki/data_mining/bagging)

(Base rate fallacy|Bonferroni's principle) (https://gerardnico.com/wiki/data_mining/base_rate_fallacy)

(Baseline|Naive) classification (Zero R) (https://gerardnico.com/wiki/data_mining/baseline)

Bayes' Theorem (Probability) (https://gerardnico.com/wiki/data_mining/bayes)

Bayesian (https://gerardnico.com/wiki/data_mining/bayesian)

Benford's law (frequency distribution of digits) (https://gerardnico.com/wiki/data_mining/benford)

Best Subset Selection Regression (https://gerardnico.com/wiki/data_mining/best_subset)

Bias (Sampling error) (https://gerardnico.com/wiki/data_mining/bias)

Bias-variance trade-off (between overfitting and underfitting) (https://gerardnico.com/wiki/data_mining/bias_trade-off)

Bayesian Information Criterion (BIC) (https://gerardnico.com/wiki/data_mining/bic)

Data Science - Big Data (https://gerardnico.com/wiki/data_mining/big_data)

R (Big R) (https://gerardnico.com/wiki/data_mining/big_r)

Bimodal Distribution (https://gerardnico.com/wiki/data_mining/bimodal_distribution)

Binary logistic regression (https://gerardnico.com/wiki/data_mining/binary_logistic_regression)

Mathematics - (Combination|Binomial coefficient|n choose k) (https://gerardnico.com/wiki/data_mining/binomial_coefficient)

(Probability|Statistics) -

bkmk=https%3A%2F%2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule)
- (http://www.stumbleupon.com /submit?title=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%29 url=https%3A%2F%2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule)
- (http://www.technorati.com/faves?add=https%3A%2F %2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule)
- (https://favorites.live.com/quickadd.aspx?marklet=1&mkt=en-us&title=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%29& url=https%3A%2F%2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule&top=1)
- (http://myweb2.search.yahoo.com/myresults /bookmarklet?title=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree popup=true&u=https%3A%2F%2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule)
- (https://www.facebook.com/sharer.php?u=https%3A%2F %2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule& t=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%29)
- (http://bookmarks.yahoo.com/toolbar/savebm?opener=tb&u=https%3A%2F %2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule& t=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%29)
- (https://twitter.com /home?status=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%2 %3A%2F%2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule)
- (http://favorites.my.aol.com/ffclient/AddBookmark?url=https%3A%2F %2Fgerardnico.com%2Fwiki%2Fdata_mining%2Fone_rule& title=Machine+Learning+-+%28One%7CSimple%29+Rule+-+%28One+Level+Decision+Tree%29& favelet=true)

data_mining/one_rule.txt · Last modified: 2013/10/11 17:35 by gerardnico

Binomial Distribution (https://gerardnico.com /wiki/data_mining /binomial_distribution)

Data Mining, Book (https://gerardnico.com /wiki/data_mining/book)

(Boosting|Gradient Boosting|Boosting trees) (https://gerardnico.com /wiki/data_mining/boosting)

Bootstrap Resampling (https://gerardnico.com /wiki/data_mining/bootstrap)

Decision boundary Visualization (https://gerardnico.com /wiki/data_mining/boundary)

(C4.5|J48) algorithm (https://gerardnico.com /wiki/data_mining/c4.5)

(Statistics|Machine Learning|Data Mining) - (Unit|Individual|Case|Subject|Observation|Insta (https://gerardnico.com /wiki/data_mining/case)

(Case-control|retrospective) sampling (https://gerardnico.com /wiki/data_mining /case_control_sampling)

Causation - Causality (Cause and Effect) Relationship (https://gerardnico.com /wiki/data_mining/causality)

Cumulative Distribution Function (CDF) (https://gerardnico.com /wiki/data_mining/cdf)

Centering Continous Predictors (https://gerardnico.com /wiki/data_mining/centering)

Central limit theorem (CLT) (https://gerardnico.com /wiki/data_mining /central_limit_theorem)

Centroid (center of gravity) (https://gerardnico.com /wiki/data_mining/centroid)

Chance (https://gerardnico.com /wiki/data_mining/chance)

Characteristic, Property, Nature (https://gerardnico.com /wiki/data_mining /characteristic)

(Class|Category|Label) Target (https://gerardnico.com /wiki/data_mining/class)

(Classifier|Classification

Function) (https://gerardnico.com/wiki/data_mining/classification)

Clustering (Function|Model) (https://gerardnico.com/wiki/data_mining/cluster)

(Prediction|Recommender System) - Collaborative filtering (https://gerardnico.com/wiki/data_mining/collaborative_filtering)

Competitions (Kaggle and others) (https://gerardnico.com/wiki/data_mining/competition)

Pattern Recognition - Computer Vision (https://gerardnico.com/wiki/data_mining/computer_vision)

Statistics - (Confidence|likelihood) (Prediction probabilities|Probability classification) (https://gerardnico.com/wiki/data_mining/confidence)

Confidence Interval (https://gerardnico.com/wiki/data_mining/confidence_interval)

Confounding (factor|variable) - (Confound|Confounder) (https://gerardnico.com/wiki/data_mining/confounding)

Confusion Matrix (https://gerardnico.com/wiki/data_mining/confusion_matrix)

Continuous Variable (https://gerardnico.com/wiki/data_mining/continuous)

(Scientific) Control (Group) (https://gerardnico.com/wiki/data_mining/control)

Convex (https://gerardnico.com/wiki/data_mining/convex)

Correlation (Coefficient analysis) (https://gerardnico.com/wiki/data_mining/correlation)

Correlation does not imply causation (https://gerardnico.com/wiki/data_mining/correlation_does_not_imply_causation)

Cosine Similarity (Measure of Angle)

(https://gerardnico.com/wiki/data_mining/cosine_similarity)

Covariance (https://gerardnico.com/wiki/data_mining/covariance)

Mallow's Cp (https://gerardnico.com/wiki/data_mining/cp)

Cross Product (of X and Y) (CP|SP) (https://gerardnico.com/wiki/data_mining/cross_product)

(Statistics|Data Mining) - (K-Fold) Cross-validation (rotation estimation) (https://gerardnico.com/wiki/data_mining/cross_validation)

(Periodicity|Periodic phenomena|Cycle) (https://gerardnico.com/wiki/data_mining/cycle)

(Data Mining|Machine Learning) - Data (Analysis|Analyse) (https://gerardnico.com/wiki/data_mining/data_analysis)

(Data|Knowledge) Discovery - Statistical Learning (https://gerardnico.com/wiki/data_mining/data_mining)

Data Point (https://gerardnico.com/wiki/data_mining/data_point)

Data (Preparation | Wrangling | Munging) (https://gerardnico.com/wiki/data_mining/data_preparation)

Data - Processing (Functions, Model) (https://gerardnico.com/wiki/data_mining/data_processing)

Data Product (https://gerardnico.com/wiki/data_mining/data_product)

Data - Science (https://gerardnico.com/wiki/data_mining/data_science)

Data Scientist (https://gerardnico.com/wiki/data_mining/data_scientist)

Decision Tree (DT) Algorithm (https://gerardnico.com/wiki/data_mining/decision_tree)

Decision Stump (https://gerardnico.com/wiki/data_mining/decisionstump)

Deep Learning (Network) (https://gerardnico.com/wiki/data_mining/deep_learning)

(Degree|Level) of confidence (https://gerardnico.com/wiki/data_mining/degree_of_confidence)

Degree of freedom (df) (https://gerardnico.com/wiki/data_mining/degree_of_freedom)

(dependent|paired sample) t-test (https://gerardnico.com/wiki/data_mining/dependent_t-test)

Math - Derivative (https://gerardnico.com/wiki/data_mining/derivative)

(Descriptive|Discovery) Analysis (https://gerardnico.com/wiki/data_mining/description)

Deviance (https://gerardnico.com/wiki/data_mining/deviance)

Deviation Score (for one observation) (https://gerardnico.com/wiki/data_mining/deviation_score)

Dimensionality (number of variable, parameter) (P) (https://gerardnico.com/wiki/data_mining/dimension)

(Dimension|Feature) (Reduction) (https://gerardnico.com/wiki/data_mining/dimension_reduction)

(Data|Text) Mining - Word-sense disambiguation (WSD) (https://gerardnico.com/wiki/data_mining/disambiguation)

Discrete Variable (https://gerardnico.com/wiki/data_mining/discrete)

(Discretizing|binning) (bin) (https://gerardnico.com/wiki/data_mining/discretization)

Discriminant analysis (https://gerardnico.com/wiki/data_mining/discriminant_analysis)

Quadratic discriminant analysis (QDA) (https://gerardnico.com

/wiki/data_mining
/discriminant_analysis_quadratic)

(Discriminative|conditional) models (https://gerardnico.com /wiki/data_mining /discriminative_model)

Distance (https://gerardnico.com /wiki/data_mining/distance)

(Probability|Sampling) Distribution (https://gerardnico.com /wiki/data_mining /distribution)

Dummy (Coding|Variable) - One-hot-encoding (OHE) (https://gerardnico.com /wiki/data_mining/dummy)

Effects (between predictor variable) (https://gerardnico.com /wiki/data_mining/effect)

Effect Size (https://gerardnico.com /wiki/data_mining /effect_size)

Elastic Net Model (https://gerardnico.com /wiki/data_mining /elastic_net)

Ensemble Learning (meta set) (https://gerardnico.com /wiki/data_mining/ensemble)

Entropy (Information Gain) (https://gerardnico.com /wiki/data_mining/entropy)

Prediction Error (Training versus Test) (https://gerardnico.com /wiki/data_mining/error)

(Error|misclassification) Rate - false (positives|negatives) (https://gerardnico.com /wiki/data_mining/error_rate)

(Estimation|Approximation) (https://gerardnico.com /wiki/data_mining/estimation)

(Estimator|Point Estimate) - Predicted (Score|Target|Outcome|...) (https://gerardnico.com /wiki/data_mining/estimator)

(Evaluation|Estimation|Validation|Testing) (https://gerardnico.com /wiki/data_mining/evaluation)

(Experimentation|Experimental research) (https://gerardnico.com /wiki/data_mining /experimentation)

Data analysis - Explanatory (https://gerardnico.com /wiki/data_mining

/explanatory)

Data Science - (Data exploration|Exploratory Analysis|Discovery ?) (https://gerardnico.com/wiki/data_mining/exploratory)

Exponential Distribution (https://gerardnico.com/wiki/data_mining/exponential_distribution)

F-distributions (https://gerardnico.com/wiki/data_mining/f-distribution)

(F-Statistic|F-test|F-ratio) (https://gerardnico.com/wiki/data_mining/f_statistic)

Face Recognition (https://gerardnico.com/wiki/data_mining/face_recognition)

(Factor Variable|Qualitative Predictor) (https://gerardnico.com/wiki/data_mining/factor)

Factor Analysis (https://gerardnico.com/wiki/data_mining/factor_analysis)

Factorial Anova (https://gerardnico.com/wiki/data_mining/factorial_anova)

(Feature|Attribute) Extraction Function (https://gerardnico.com/wiki/data_mining/feature_extraction)

Feature Hashing (https://gerardnico.com/wiki/data_mining/feature_hashing)

(Attribute|Feature) (Selection|Importance) (https://gerardnico.com/wiki/data_mining/feature_selection)

Fraud Detection (https://gerardnico.com/wiki/data_mining/fraud)

(Frequency|Rate) (https://gerardnico.com/wiki/data_mining/frequency)

(Frequent itemsets|co-occurring items) (https://gerardnico.com/wiki/data_mining/frequent_itemset)

Frequentist (https://gerardnico.com/wiki/data_mining/frequentist)

Data Model - Fudge factor

(https://gerardnico.com/wiki/data_mining/fudge_factor)

Fuzzy Logic (Partial Truth) (https://gerardnico.com/wiki/data_mining/fuzzy)

Generalized additive model (GAM) (https://gerardnico.com/wiki/data_mining/gam)

Gaussian processes (modelling probability distributions over functions) (https://gerardnico.com/wiki/data_mining/gaussian_proces)

Generalized Boosted Regression Models (https://gerardnico.com/wiki/data_mining/gbm)

Generative Model (https://gerardnico.com/wiki/data_mining/generative_model)

Getting Started (https://gerardnico.com/wiki/data_mining/getting_started)

Generalized Linear Models (GLM) - Extensions of the Linear Model (https://gerardnico.com/wiki/data_mining/glm)

(Stochastic) Gradient descent (SGD) (https://gerardnico.com/wiki/data_mining/gradient_descendent)

User Group (https://gerardnico.com/wiki/data_mining/group)

Grouping (https://gerardnico.com/wiki/data_mining/grouping)

Head (https://gerardnico.com/wiki/data_mining/head)

Hierarchical Clustering (https://gerardnico.com/wiki/data_mining/hierarchical_clustering)

Hierarchy (https://gerardnico.com/wiki/data_mining/hierarchy)

Data Science - History (https://gerardnico.com/wiki/data_mining/history)

Homoscedasticity (https://gerardnico.com/wiki/data_mining/homoscedasticity)

ID3 Algorithm (https://gerardnico.com/wiki/data_mining/id3)

Intrusion detection systems (IDS) (https://gerardnico.com/wiki/data_mining/ids)

Independent t-test (https://gerardnico.com/wiki/data_mining/independent_t-test)

Statistical - Inference (https://gerardnico.com/wiki/data_mining/inference)

Information Gain (https://gerardnico.com/wiki/data_mining/information_gain)

Information Retrieval (https://gerardnico.com/wiki/data_mining/information_retrieval)

(Interaction|Synergy) effect (https://gerardnico.com/wiki/data_mining/interaction)

Intercept - Regression (coefficient|constant) B_0 (https://gerardnico.com/wiki/data_mining/intercept)

Model Interpretation (https://gerardnico.com/wiki/data_mining/interpretation)

(Interval|Delta) (Measurement) (https://gerardnico.com/wiki/data_mining/interval)

Java API for data mining (JDM) (https://gerardnico.com/wiki/data_mining/jdm)

K-Means Clustering algorithm (https://gerardnico.com/wiki/data_mining/k-means)

Kernel (https://gerardnico.com/wiki/data_mining/kernel)

Keep it simple (https://gerardnico.com/wiki/data_mining/kiss)

K-Nearest Neighbors (KNN) algorithm - Instance based learning (https://gerardnico.com/wiki/data_mining/knn)

Knots (Cut points) (https://gerardnico.com/wiki/data_mining/knot)

Kurtosis (Distribution Tail extremity) (https://gerardnico.com/wiki/data_mining/kurtosis)

Statistical Learning - Lasso (https://gerardnico.com/wiki/data_mining/lasso)

Standard Least Squares Fit (https://gerardnico.com/wiki/data_mining/least_square)

Leptokurtic distribution (https://gerardnico.com/wiki/data_mining/leptokurtic_distribution)

(Level|Label) (https://gerardnico.com/wiki/data_mining/level)

(Lying|Lie) (https://gerardnico.com/wiki/data_mining/lie)

(Life cycle|Project|Data Pipeline) (https://gerardnico.com/wiki/data_mining/lifecycle)

Lift Chart (https://gerardnico.com/wiki/data_mining/lift_chart)

Statistical Learning - Simple Linear Discriminant Analysis (LDA) (https://gerardnico.com/wiki/data_mining/linear_discriminant_analysis)

Fisher (Multiple Linear Discriminant Analysis|multi-variant Gaussian) (https://gerardnico.com/wiki/data_mining/linear_discriminant_analysis_multiple)

Linear (Regression|Model) (https://gerardnico.com/wiki/data_mining/linear_regression)

(Linear spline|Piecewise linear function) (https://gerardnico.com/wiki/data_mining/linear_spline)

Little r - (Pearson product-moment Correlation coefficient) (https://gerardnico.com/wiki/data_mining/little_r)

Global vs Local (https://gerardnico.com/wiki/data_mining/local)

LOcal (Weighted) regrESSion (LOESS|LOWESS) (https://gerardnico.com/wiki/data_mining/local_regression)

Log-likelihood function (cross-entropy) (https://gerardnico.com/wiki/data_mining/log_likelihood)

Logistic regression (Classification Algorithm) (https://gerardnico.com/wiki/data_mining/logistic_regression)

(Logit|Logistic) (Function|Transformation) (https://gerardnico.com/wiki/data_mining/logit)

Loss functions (Incorrect predictions penalty) (https://gerardnico.com/wiki/data_mining/loss_function)

Data Science - (Kalman Filtering|Linear quadratic estimation (LQE)) (https://gerardnico.com/wiki/data_mining/lqe)

Machine Learning (https://gerardnico.com/wiki/data_mining/machine_learning)

Main Effect (https://gerardnico.com/wiki/data_mining/main)

Probability mass function (PMF) (https://gerardnico.com/wiki/data_mining/mass)

Maximum (https://gerardnico.com/wiki/data_mining/maximum)

Maximum Entropy Algorithm (https://gerardnico.com/wiki/data_mining/maximum_entropy)

Maximum likelihood (https://gerardnico.com/wiki/data_mining/maximum_likelihood)

Measure (https://gerardnico.com/wiki/data_mining/measure)

(Scales of measurement|Type of variables) (https://gerardnico.com/wiki/data_mining/measurement)

(Missing Value|Not Available) (https://gerardnico.com/wiki/data_mining/missing)

(Function|Model) (https://gerardnico.com/wiki/data_mining/model)

Model Selection (https://gerardnico.com/wiki/data_mining/model_selection)

Model Size (d) (https://gerardnico.com/wiki/data_mining/model_size)

Moderator Variable (Z) - Moderation (https://gerardnico.com/wiki/data_mining/moderator)

Monte Carlo (method|experiment) (stochastic process simulations) (https://gerardnico.com/wiki/data_mining/monte_carlo)
(Average|Mean) Squared (MS) prediction error (MSE) (https://gerardnico.com/wiki/data_mining/mse)
Multi-variant logistic regression (https://gerardnico.com/wiki/data_mining/multi-variant_logistic_regression)
Multi-class (classification|problem) (https://gerardnico.com/wiki/data_mining/multi_class)
(Multiclass Logistic|multinomial) Regression (https://gerardnico.com/wiki/data_mining/multiclass_logistic_regression)
Multidimensional scaling (similarity of individual cases in a dataset) (https://gerardnico.com/wiki/data_mining/multidimensional_scaling)
Multiple Linear Regression (https://gerardnico.com/wiki/data_mining/multiple_regression)
Naive Bayes (NB) (https://gerardnico.com/wiki/data_mining/naive_bayes)
(Probabilistic?) Neural Network (PNN) (https://gerardnico.com/wiki/data_mining/neural_network)
Null Hypothesis Significance Testing (NHST) (https://gerardnico.com/wiki/data_mining/nhst)
(No Predictor|Mean|Null) Model (https://gerardnico.com/wiki/data_mining/no_model)
Noise (Unwanted variation) (https://gerardnico.com/wiki/data_mining/noise)
(Discrete | Nominal | Category | Reference | Taxonomy | Class | Enumerated | Factor | Qualitative | Constant) Data (https://gerardnico.com/wiki/data_mining/nominal)
Non-linear (effect|function|model) (https://gerardnico.com

/wiki/data_mining/non-linear)

Non-Negative Matrix Factorization (NMF) Algorithm (https://gerardnico.com /wiki/data_mining /non-negative_matrix_factorization)

Multi-response linear regression (Linear Decision trees) (https://gerardnico.com /wiki/data_mining /non_linear)

(Normal|Gaussian) Distribution - Bell Curve (https://gerardnico.com /wiki/data_mining /normal_distribution)

Orthogonal Partitioning Clustering (O-Cluster or OC) algorithm (https://gerardnico.com /wiki/data_mining/o-cluster)

Odds (Ratio) (https://gerardnico.com /wiki/data_mining/odds)

(One|Simple) Rule - (One Level Decision Tree) (https://gerardnico.com /wiki/data_mining/one_rule)

Outliers Cases (https://gerardnico.com /wiki/data_mining/outlier)

(Overfitting|Overtraining|Robust|Generalization) (Underfitting) (https://gerardnico.com /wiki/data_mining/overfitting)

Data Science - Over-generalization (https://gerardnico.com /wiki/data_mining /overgeneralization)

P-value (https://gerardnico.com /wiki/data_mining/p-value)

(Mathematics|Statistics) - Statistical Parameter (https://gerardnico.com /wiki/data_mining/parameter)

(Non) Parametrics (method|statistics) (https://gerardnico.com /wiki/data_mining /parametric)

Pareto (https://gerardnico.com /wiki/data_mining/pareto)

Pattern (https://gerardnico.com /wiki/data_mining/pattern)

Principal Component (Analysis|Regression) (PCA) (https://gerardnico.com /wiki/data_mining/pca)

(Probability) Density

Function (PDF) (https://gerardnico.com/wiki/data_mining/pdf)

Mathematics - Permutation (Ordered Combination) (https://gerardnico.com/wiki/data_mining/permutation)

Piecewise polynomials (https://gerardnico.com/wiki/data_mining/piecewise_polynomial)

Partial least squares (PLS) (https://gerardnico.com/wiki/data_mining/pls)

Predictive Model Markup Language (PMML) (https://gerardnico.com/wiki/data_mining/pmml)

Poisson (Process|distribution) (https://gerardnico.com/wiki/data_mining/poisson_distribution)

(Global) Polynomial Regression (Degree) (https://gerardnico.com/wiki/data_mining/polynomial)

Population (https://gerardnico.com/wiki/data_mining/population)

Population Parameter (https://gerardnico.com/wiki/data_mining/population_parameter)

Post-hoc test (https://gerardnico.com/wiki/data_mining/post-hoc)

Power of a test (https://gerardnico.com/wiki/data_mining/power)

(Prediction|Guess) (https://gerardnico.com/wiki/data_mining/prediction)

Predictive Model Markup Language (PMML) (https://gerardnico.com/wiki/data_mining/predictive_model_markup_language)

(Machine|Statistical) Learning - (Predictor|Feature|Regressor|Characteristic) - (Independent|Explanatory) Variable (X) (https://gerardnico.com/wiki/data_mining/predictor)

Privacy (Anonymization) (https://gerardnico.com/wiki/data_mining/privacy)

Probability (https://gerardnico.com/wiki/data_mining/probability)

Probit Regression

(probability on binary problem) (https://gerardnico.com/wiki/data_mining/probit)

Problem (https://gerardnico.com/wiki/data_mining/problem)

Process control (SPC) (https://gerardnico.com/wiki/data_mining/process)

Pruning (a decision tree, decision rules) (https://gerardnico.com/wiki/data_mining/pruning)

R-squared (R^2|Coefficient of determination) for Model Accuracy (https://gerardnico.com/wiki/data_mining/r_squared)

Random forest (https://gerardnico.com/wiki/data_mining/random_forest)

Random Variable (https://gerardnico.com/wiki/data_mining/random_variable)

Range (https://gerardnico.com/wiki/data_mining/range)

Rare Event (https://gerardnico.com/wiki/data_mining/rare)

(Fraction|Ratio|Percentage|Share) (Variable|Measurement) (https://gerardnico.com/wiki/data_mining/ratio)

Raw score (https://gerardnico.com/wiki/data_mining/raw_score)

Regression (https://gerardnico.com/wiki/data_mining/regression)

(Regression Coefficient|Weight|Slope) (B) (https://gerardnico.com/wiki/data_mining/regression_coefficient)

Assumptions underlying correlation and regression analysis (Never trust summary statistics alone) (https://gerardnico.com/wiki/data_mining/regression_correlation_assumption)

(Machine learning|Inverse problems) - Regularization (https://gerardnico.com/wiki/data_mining/regularization)

Reinforcement learning (https://gerardnico.com/wiki/data_mining

/reinforcement)

Sampling - Sampling (With|without) replacement (WR|WOR) (https://gerardnico.com/wiki/data_mining/replacement)

ReSampling Validation (https://gerardnico.com/wiki/data_mining/resampling)

Research (https://gerardnico.com/wiki/data_mining/research)

(Residual|Error Term|Prediction error|Deviation) (e|ϵ) (https://gerardnico.com/wiki/data_mining/residual)

Resistant (https://gerardnico.com/wiki/data_mining/resistant)

Result Considerations (https://gerardnico.com/wiki/data_mining/result)

Ridge regression (https://gerardnico.com/wiki/data_mining/ridge_regression)

Root mean squared (Error|Deviation) (RMSE|RMSD) (https://gerardnico.com/wiki/data_mining/rmse)

ROC Plot and Area under the curve (AUC) (https://gerardnico.com/wiki/data_mining/roc)

Rote Classifier (https://gerardnico.com/wiki/data_mining/rote)

Residual sum of Squares (RSS) = Squared loss ? (https://gerardnico.com/wiki/data_mining/rss)

(Decision) Rule (https://gerardnico.com/wiki/data_mining/rule)

(Data Set|Sample) (https://gerardnico.com/wiki/data_mining/sample)

Sample size (N) (https://gerardnico.com/wiki/data_mining/sample_size)

Sampling (https://gerardnico.com/wiki/data_mining/sampling)

Sampling Distribution (https://gerardnico.com/wiki/data_mining/sampling_distribution)

Sampling Error

(https://gerardnico.com/wiki/data_mining/sampling_error)

Scale (https://gerardnico.com/wiki/data_mining/scale)

Python scikit-learn (https://gerardnico.com/wiki/data_mining/scikit-learn)

Scoring (Applying) (https://gerardnico.com/wiki/data_mining/scoring)

(Random) Seed (https://gerardnico.com/wiki/data_mining/seed)

(Shrinkage|Regularization) of Regression Coefficients (https://gerardnico.com/wiki/data_mining/shrinkage)

Signal (Wanted Variation) (https://gerardnico.com/wiki/data_mining/signal)

Significance level (https://gerardnico.com/wiki/data_mining/significance_level)

(Significance | Significant) Effect (https://gerardnico.com/wiki/data_mining/significant)

Similarity (https://gerardnico.com/wiki/data_mining/similarity)

Simple Effect (https://gerardnico.com/wiki/data_mining/simple)

(Univariate|Simple) Logistic regression (https://gerardnico.com/wiki/data_mining/simple_logistic_regression)

(Univariate|Simple) Linear Regression (https://gerardnico.com/wiki/data_mining/simple_regression)

Skew (-ed Distribution|Variable) (https://gerardnico.com/wiki/data_mining/skewed_distribution)

Data Mining / (Software|Tool|Programming Language) (https://gerardnico.com/wiki/data_mining/software)

(Spread | Variability) of a sample (https://gerardnico.com/wiki/data_mining/spread)

Stacking (https://gerardnico.com/wiki/data_mining/stacking)

Standard Deviation (SD) (https://gerardnico.com/wiki/data_mining/standard_deviation)

Standard Error (SE) (https://gerardnico.com/wiki/data_mining/standard_error)

(Normalize|Standardize) (https://gerardnico.com/wiki/data_mining/standardize)

Statistic (https://gerardnico.com/wiki/data_mining/statistic)

Statistics (https://gerardnico.com/wiki/data_mining/statistics)

Forward and Backward Stepwise (Selection|Regression) (https://gerardnico.com/wiki/data_mining/stepwise_regression)

(Stochastic|random) process (https://gerardnico.com/wiki/data_mining/stochastic_process)

(Data|Data Set) (Summary|Description) (https://gerardnico.com/wiki/data_mining/summary)

(Supervised|Directed) Learning ("Training") (Problem) (https://gerardnico.com/wiki/data_mining/supervised)

Support Vector Machines (SVM) algorithm (https://gerardnico.com/wiki/data_mining/support_vector_machine)

Singular Value Decomposition (SVD) (https://gerardnico.com/wiki/data_mining/svd)

(Student's) t-test (Mean Comparison) (https://gerardnico.com/wiki/data_mining/t-test)

(t-value|t-statistic) (https://gerardnico.com/wiki/data_mining/t-value)

T-distributions (https://gerardnico.com/wiki/data_mining/t_distribution)

Tail (https://gerardnico.com/wiki/data_mining/tail)

(Machine|Statistical) Learning - (Target|Learned|Outcome|Dependent|Response) (Attribute|Variable) (Y|DV) (https://gerardnico.com

/wiki/data_mining/target)

Hypothesis (Tests|Testing) (https://gerardnico.com /wiki/data_mining/test)

(Test|Expected|Generalization) Error (https://gerardnico.com /wiki/data_mining/test_error)

Test Set (https://gerardnico.com /wiki/data_mining/test_set)

(Threshold|Cut-off) of binary classification (https://gerardnico.com /wiki/data_mining/threshold)

Titanic Data Set (https://gerardnico.com /wiki/data_mining/titanic)

(Training|Building|Learning|Fitting) (https://gerardnico.com /wiki/data_mining/training)

Training Error (https://gerardnico.com /wiki/data_mining /training_error)

Training (Data|Set) (https://gerardnico.com /wiki/data_mining /training_set)

Nested (Transactional|Historical) Data (https://gerardnico.com /wiki/data_mining /transactional_data)

Transform (https://gerardnico.com /wiki/data_mining/transform)

Treatments (Combination of factor level) (https://gerardnico.com /wiki/data_mining/treatment)

True score (Classical test theory) (https://gerardnico.com /wiki/data_mining /true_score)

(True Function|Truth) (https://gerardnico.com /wiki/data_mining/truth)

(Total) Sum of the square (TSS|SS) (https://gerardnico.com /wiki/data_mining/tss)

Tuning Parameter (https://gerardnico.com /wiki/data_mining /tuning_parameter)

(two class|binary) classification problem (https://gerardnico.com /wiki/data_mining/two_class)

Statistical Learning - Two-fold validation (https://gerardnico.com

/wiki/data_mining
/two_fold_validation)

Data - Uncertainty
(https://gerardnico.com
/wiki/data_mining
/uncertainty)

Uniform Distribution
(platykurtic)
(https://gerardnico.com
/wiki/data_mining
/uniform_distribution)

Unsupervised Learning
("Mining")
(https://gerardnico.com
/wiki/data_mining
/unsupervised)

Resampling through
Random Percentage Split
(https://gerardnico.com
/wiki/data_mining
/validation_set)

Validity (Valid Measures)
(https://gerardnico.com
/wiki/data_mining/validity)

(Variance|Dispersion|Mean
Square) (MS|σ)
(https://gerardnico.com
/wiki/data_mining/variance)

Probability and Vizualization
(https://gerardnico.com
/wiki/data_mining/viz)

Statistics vs (Machine
Learning|Data Mining)
(https://gerardnico.com
/wiki/data_mining/vs)

Random Walk
(https://gerardnico.com
/wiki/data_mining/walk)

(Golf|Weather) Data Set
(https://gerardnico.com
/wiki/data_mining/weather)

Weka
(https://gerardnico.com
/wiki/data_mining/weka)

Z Scale
(https://gerardnico.com
/wiki/data_mining/z_scale)

Z Score (Zero Mean) or
Standard Score
(https://gerardnico.com
/wiki/data_mining/z_score)

Back to top