
Index of Data Mining Topics

- Data pre-processing
 - Creating new variables, 108–112
 - feature selection, 112–117, 241–251
 - standardization, 62, 124
 - unknown values, 53–63
- Data summarization, 43–52, 239, 240
 - basic statistics, 43, 167–174
 - comparing distributions, 179–183
 - correlation, 56
 - inter-quartile range, 44, 47, 179, 241
 - measures of centrality, 179
 - measures of spread, 179
- Data visualization, 43–52
 - bar plot, 168
 - box plot, 47, 171
 - box-percentile plot, 50
 - candlestick graphs, 110
 - conditioned box plot, 49
 - conditioned histogram, 59
 - conditioned strip plot, 51, 60
 - histogram, 44, 239
 - interacting with plots, 48, 80
 - level plot, 248
 - log scales, 171
 - QQ plot, 45
 - strip plot, 51
- Descriptive models
 - box plot rule, 173, 196–201
 - clustering of variables, 250
 - hierarchical agglomerative clustering, 205, 250
 - LOF, 201–204
 - ORh, 205–208
- Evaluation criteria
 - accuracy, 119, 252
 - AUC, 252
 - Brier score, 252
 - confusion matrix, 120, 266
 - cumulative recall charts, 192
 - error rate, 119
 - errors scatter plot, 79
 - F-measure, 120
 - financial trading criteria, 132, 138–141, 158–162
 - lift charts, 191
 - mean absolute error, 77
 - mean squared error, 78
 - misclassification costs, 252
 - NDTP, 193
 - normalized mean squared error, 78
 - precision, 120, 188, 252
 - precision/recall curves, 188–191
 - recall, 120, 188, 252
- Evaluation methodologies
 - cross validation, 81–91
 - hold out, 194–195
 - leave-one-out cross validation, 253–254, 258–265
 - monte carlo, 141–156
 - significance of differences, 89, 153
 - stratified samples, 194

Modeling tasks

- class imbalance
 - SMOTE, 210
- classification, 117, 185
 - class imbalance, 120, 185, 209–211
- clustering, 184
- outlier detection, 184
- regression, 63–93, 118
- semi-supervised classification, 187
- semi-supervised clustering, 186
- time series forecasting, 121–130
 - growing window, 122
 - regime shift, 121
 - sliding window, 122

Predictive models

- AdaBoost, 217–223
- artificial neural networks, 123–126
- k nearest neighbours, 255–257
- multiple linear regression, 64–71
- multivariate adaptive regression splines, 129–130
- naive bayes, 211–217
- random forests, 88, 255
- regression trees, 71–77
- self-training, 187, 223–229
- support vector machines, 127–129, 254