

hosts over 20,000 datasets with over 50,000 associated machine learning tasks. Working with these datasets can provide a great opportunity to practice your machine learning skills. A disadvantage of competitions is that they already provide a particular metric to optimize, and usually a fixed, preprocessed dataset. Keep in mind that defining the problem and collecting the data are also important aspects of real-world problems, and that representing the problem in the right way might be much more important than squeezing the last percent of accuracy out of a classifier.

## Conclusion

We hope we have convinced you of the usefulness of machine learning in a wide variety of applications, and how easily machine learning can be implemented in practice. Keep digging into the data, and don't lose sight of the larger picture.

## A

- A/B testing, 359
- accuracy, 22, 282
- acknowledgments, xi
- adjusted rand index (ARI), 191
- agglomerative clustering
  - evaluating and comparing, 191
  - example of, 183
  - hierarchical clustering, 184
  - linkage choices, 182
  - principle of, 182
- algorithm chains and pipelines, 305-321
  - building pipelines, 308
  - building pipelines with `make_pipeline`, 313-316
  - grid search preprocessing steps, 317
  - grid-searching for model selection, 319
  - importance of, 305
  - overview of, 320
  - parameter selection with preprocessing, 306
  - pipeline interface, 312
  - using pipelines in grid searches, 309-311
- algorithm parameter, 118
- algorithms (see also models; problem solving)
  - evaluating, 28
  - minimal code to apply to algorithm, 24
  - sample datasets, 30-34
- scaling
  - MinMaxScaler, 102, 135-139, 190, 230, 308, 319
  - Normalizer, 134
  - RobustScaler, 133
  - StandardScaler, 114, 133, 138, 144, 150, 190-195, 314-320
- supervised, classification
  - decision trees, 70-83
  - gradient boosting, 88-91, 119, 124
  - k-nearest neighbors, 35-44
  - kernelized support vector machines, 92-104
  - linear SVMs, 56
  - logistic regression, 56
  - naive Bayes, 68-70
  - neural networks, 104-119
  - random forests, 84-88
- supervised, regression
  - decision trees, 70-83
  - gradient boosting, 88-91
  - k-nearest neighbors, 40
  - Lasso, 53-55
  - linear regression (OLS), 47, 220-229
  - neural networks, 104-119
  - random forests, 84-88
  - Ridge, 49-55, 67, 112, 231, 234, 310, 317-319
- unsupervised, clustering
  - agglomerative clustering, 182-187, 191-195, 203-207
  - DBSCAN, 187-190
  - k-means, 168-181
- unsupervised, manifold learning
  - t-SNE, 163-168
- unsupervised, signal decomposition
  - non-negative matrix factorization, 156-163
  - principal component analysis, 140-155
- alpha parameter in linear models, 50
- Anaconda, 6