
H2O.ai

Machine Learning Concepts

1. Out of Sample Data
2. Data Leakage
3. Target Class Imbalance
4. Extreme Values & Outliers
5. Low Frequency Categories
6. Redundant Data
7. Irrelevant Data
8. Missing Data
9. Model Scoring Properties
10. Model Interpretability

Training and Test Data Sets

Training Set vs. Test Set

- Partition the original data (randomly or stratified) into a **training** set and a **test** set. (e.g. 70/30)

Training Error vs. Test Error

- It can be useful to evaluate the training error, but you should not look at training error alone.
- Training error is not an estimate of **generalization error** (on a test set or cross-validated), which is what you should care more about.
- Training error vs test error over time is an useful thing to calculate. It can tell you when you start to overfit your model, so it is a useful metric in supervised machine learning.

Performance Metrics

- Regression: R^2, MSE, RMSE
- Classification: Accuracy, F1, H-measure, Log-loss
- Ranking (Binary Outcome): AUC, Partial AUC

Machine Learning Concepts

1. Out of Sample Data
2. Data Leakage
3. Target Class Imbalance
4. Extreme Values & Outliers
5. Low Frequency Categories
6. Redundant Data
7. Irrelevant Data
8. Missing Data
9. Model Scoring Properties
10. Model Interpretability