Data Splitting and Bias-Variance Tradeoff

Aniket Kesari

UC Berkeley

October 1, 2020

Supervised v. Unsupervised Learning

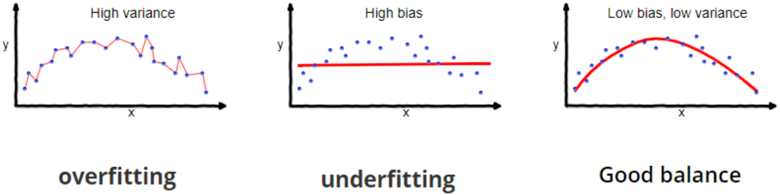
- Machine learning is divided into **supervised** and **unsupervised** learning problems
- Supervised is further divided into **regression** and **classification**
- A core problem in machine learning is that the learning process tends to **overfit** to the training data, and therefore makes poor test and out-of-sample predictions

Definition of Bias-Variance Tradeoff

- **Bias**: The difference between a model's prediction and the actual value of an observation
- **Variance**: The complexity of the model

$$Err(x) = Bias^2 + Variance + \epsilon$$

Overfitting Visual Intuition



overfitting

underfitting

Good balance

Figure 1: Over and Underfitting

Machine Learning

- Develop dynamic models that are data-dependent
- Basic Process:
  - Split data into **training**, **validation**, and **test** sets
  - Train the ML algorithm
  - Use the validation set to make adjustments to the model(s)
  - Test the final model on the test set
  - Operationalize algorithm on new data

Machine Learning

- Supervised Learning
  - Training data contains *labels* for the outcome9s)
  - Machine Learning algorithm infers a function describing the relationship between the inputs and the output
  - Algorithm can be used on a new set of input data to infer the output
  - Examples: Linear Regression, Decision Trees, Support Vector Machines, etc.
- Unsupervised Learning
  - Training data does not contain any labels
  - Algorithm instead trains to uncover underlying patterns in the data
  - Used for clustering, dimensionality reduction, etc.
  - Examples: k-means, Principal Compnents Analysis, Singular Value Decomposition, Expectation-Maximization

Regression and Classification

- Typically two tasks in supervised learning: regression and classification
- Regression
    - Predict a continuous outcome response from the input data
    - Ex. Ordinary Least Squares
- Classification
    - Predict membership in a group
    - Ex. Logistic Regression
- Several ML methods are well suited to both regression and classification problems
- An important first step in any supervised machine learning problem is to identify whether you're dealing with a regression or classification problem, and approach it accordingly

Bias-Variance Tradeoff

- Two goals:
  - Minimize *test bias*: This means using as much data as we can in the training phase, which necessarily means reducing the amount of test data available
  - Minimize *test variance*: But, we also want a decent number of points in the test set, otherwise the estimates will have large variances
- Fewer folds lead to higher test bias, but more folds lead to higher test variance

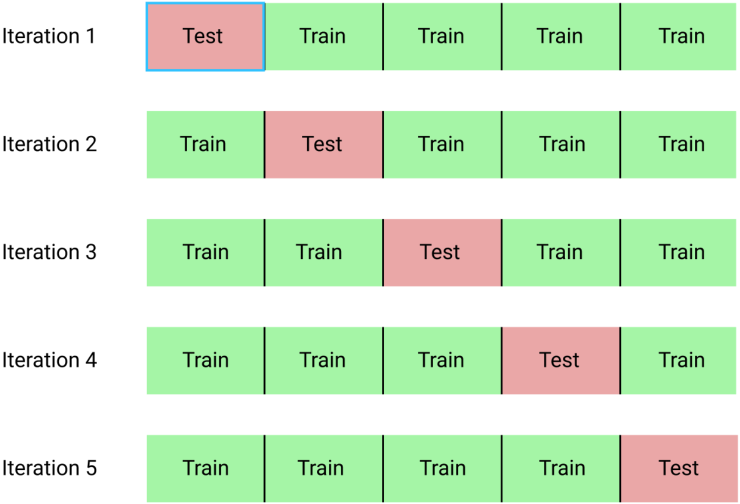
Cross-Validation

- Introduced to statistics from machine learning
- Procedure for k-fold: -Partition the data into a number of folds
    - Train the data on k-1 folds
    - Test on the last fold
    - Rotate so that each fold acts as the test set once
    - Take the mean estiamte

Cross-Validation

- Advantages:
    - Tends to avoid overfitting problems
    - Usable with relatively small datasets (compared to train/test/validation split)
    - Does not make the background assumptions required in information criteria approach
- Disadvantages:
    - Assumes that the out-of-sample data was drawn from the same population as the training data
    - Computationally VERY expensive
- k=5 or 10 is conventionally used, but it is by no means perfectly suited to every context
- In general, this problem lessens the more data you have

K-Fold CV Illustration



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Figure 2: Illustration of k-fold cross validation

Cross-Validatioin Techniques

- K-Fold
  - Divide into k-folds and rotate
- Leave-one-out (LOO)
  - Leave out one observation, train the model on the rest, and calculate on the left out observations

Comparison

- Cross-Validation is generally preferred nowadays because of advances in computing
- However, AIC is asymptotically equivalent to LOOCV if the assumptions are met
- Generally use cross-validation unless it is computationally cost prohibitive

Conclusion

- Machine Learning requires we split our data to evaluate our models
- Data splitting involves substantive choices on the part of the analyst
- Different techniques have different pros and cons, but the main issue comes down to bias-variance tradeoff