Supervised Learning: Model Assessment and Selection

06-20-2023

Supervised Learning

Goal: Uncover associations between a set of predictor variables and a single response (or dependent) variable.

Examples of Statistical Learning Methods/Algorithms:

- GLMS and penalized versions (Lasso, elastic net)
- Smoothing splines, GAMs
- Decision trees and its variants
- Neural nets

Two Main Types:

- Regression models
- Classical models

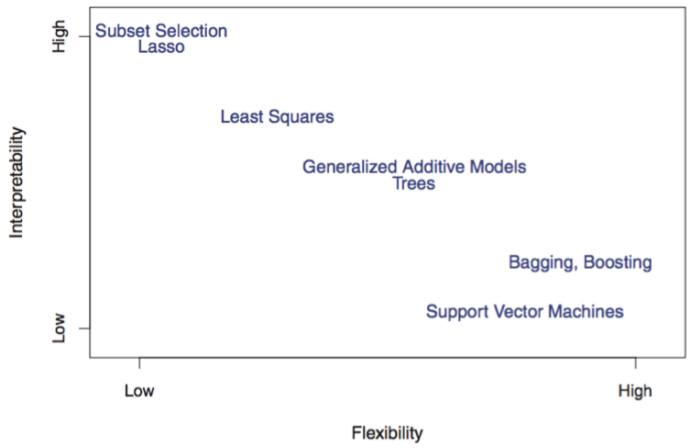
How to Determine which Method to Use?

Depends on your goal: *inference* versus *prediction*

$$\hat{Y} = \hat{f}(X)$$

Any algorithm can be used for prediction, however options are limited for inference. Inference is really about learning the details of $\hat{f}(X)$.

Model Flexibility vs. Interpretability



Generally there is a **trade-off** between a model's flexibility and how interpretable it is (i.e., explainable power).

- **Parametric** models, for which we can write down a mathematical expression for f(X) **before observing the data** and are **inherently less flexible**.
- **Nonparametric** models, in which f(X) is **estimated from the data** (e.g., kernel regression).

Model Assessment vs. Selection

Model Assessment: evaluating how well a learned model performs, via the use of a single-metric

Model Selection: selecting the 'best' model from a suite of learning models.

How Do We Deal with Flexibility?

Goal: have good estimates of f(X) **without overfitting** the data.

Two Common Approaches:

- *Split data into test and training.*

  Training = data used to train models

  Test = data used to test models

- *K-fold cross validation*

  Each observation is placed in "hold-out" aka test data exactly once. Repeat data splitting k times.

  Brief note on Reproducibility: set a seed so random processes/analyses can be reproduced!

Model Assessment Metrics

Loss Function (aka objective or cost function) is a metric that represents **the quality of fit of a model**

For regression, we typically use **mean squared error (MSE)**

Note that MSE is unit-dependent

For classification:

- **Misclassification rate (MCR):** percentage of predictions that are wrong
- **Area under Curve (AUC)**
- interpretation of these metrics can be affected by **class imbalance**

Model Selection