Machine Learning and Applied Econometrics

Introduction

Machine Learning and Econometrics

- This introductory lecture is based on
  - Hal R. Varian, Big Data: New Tricks for Econometrics, Journal of Economic Perspectives 28:2 (3-28), Spring 2014.
  - Sendhil Mullainathan and Jann Spiess, Machine Learning: An Applied Econometric Approach, Journal of Economic Perspectives 31:2 (87-106), Spring 2017.
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, Introduction to Statistical Learning with Applications in R, Springer 2013.
  - Darren Cook, Practical Machine Learning with H2O, O'Reilly Media, Inc., 2017.
  - Scott Burger, Introduction to Machine Learning with R: Rigorous Mathematical Analysis, O'Reilly Media, Inc., 2018.

Machine Learning

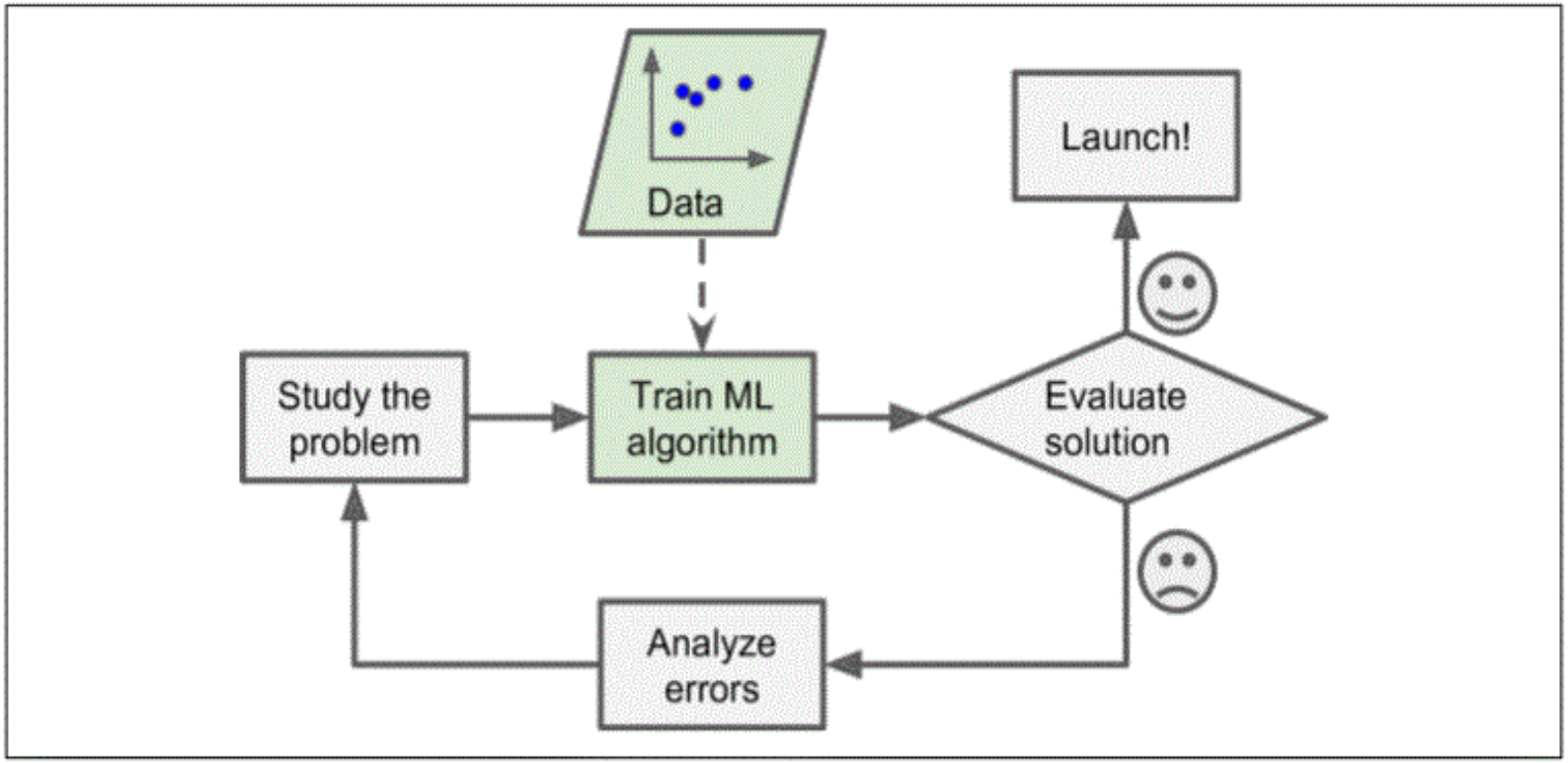

Figure 1-2. Machine Learning approach

Machine Learning

- Model (Problem)
  - Regression
  - Classification
  - Mixed Model
- Learning from Data
  - Training and Testing
  - Validation
- Methods (Algorithms)
  - Supervised vs. Unsupervised Learning
- Prediction

Machine Learning

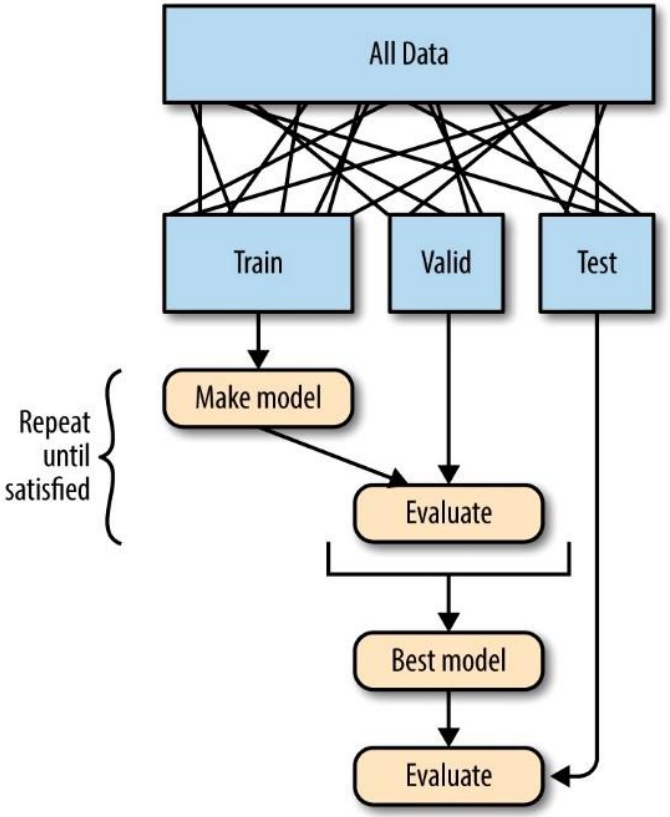

Figure 2-2. Summary of how train, valid, and test data sets are used

Learning from Data

- Data Importing
- Data Manipulation
  - Concatenation (Merging)
  - Transformation
- Data Splitting
  - Training and Testing
  - Validation

Model Evaluation Metrics

- Regression
  - R2 (R Squared)
  - MSE (Mean Squared Error),
  - RMSE
  - MAE
- Classification
  - Gini Index: [0,1]
  - Accuracy Rate
  - Logloss
    - Binomial Classification
    - Multinomial Classification
  - AUC (Area Under the ROC Curve)

Machine Learning and Econometrics

- Machine Learning
  - Statistical learning with machine intelligence on large datasets (i.e., large n and/or p)
  - Focus on nonparametric prediction without over fitting
- Econometrics
  - Causal inference of economic data for decision making based on economic theory
  - Focus on parameter estimation, hypothesis testing, and statistical inference

Econometrics for Machine Learning

- Causal Inference Methods
  - Confounding and Instrumental Variables
  - Regression Discontinuity
  - Difference in Difference
- Causal Effects Estimation
- Program or Policy Evaluation

Machine Learning for Econometrics

- Regression and Classification
  - Linear vs. Nonlinear
  - Parametric vs. Nonparametric
- Model Evaluation
  - Bias and Variance Trade-off
- Cross-Validation*
- Variable Selection
  - Regularization: Ridge, LASSO, Elastic Net, …
- Prediction

Cross Validation

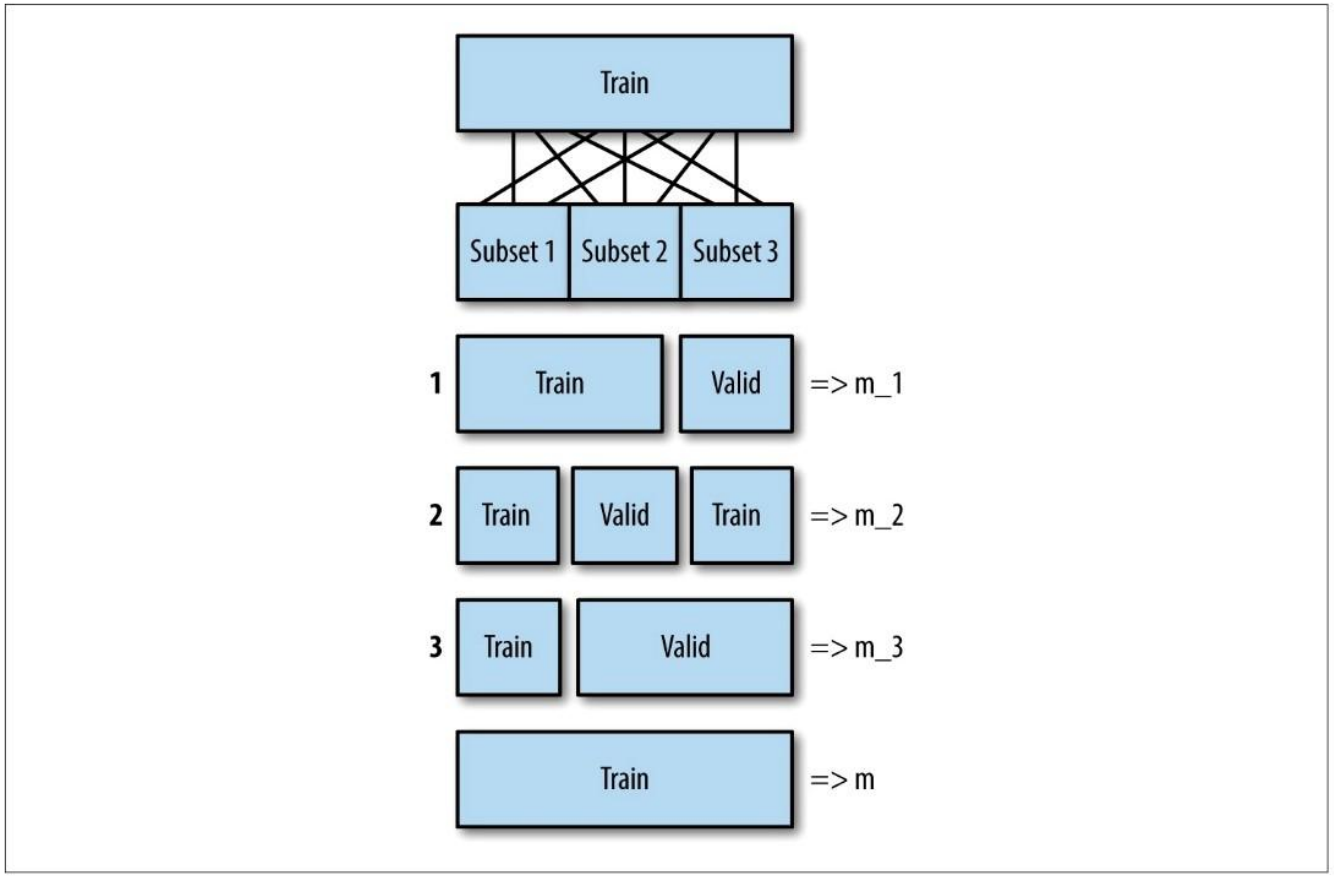

Figure 4-3. The cross-validation process for k=3

Cross Validation

- K-fold Cross Validation

|---------- N ----------|

|-- N_k --|: the number of observations

- The predicted $\hat{Y}_i^k$ is obtained from the estimated model with the k^{th} part's observations removed

Cross Validation

- K-fold Cross Validation (Continued)
  - Randomly divide the sample into K equal-sized parts. Leave out part k, fit the model to the other K-1 parts of combined
  - Obtain the prediction errors for the left-out k^{th} part, and compute CV as

$$CV = \sum_{k=1}^{K} \frac{N_k}{N} MSE_k, \; where \; MSE_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (Y_i - \hat{Y}_i^k)^2$$

    - CV tends to be biased upward

Cross Validation

- K-fold Cross Validation (Continued)
  - If K=N, this is N-fold or leave one out cross validation (LOOCV)
  - Special Case: OLS

$$CV = \sum_{i=1}^{N} \left(\frac{Y_i - \hat{Y}_i}{1 - h_i} \right)^2, \; where \; h = diag[X(X'X)^{-1}X']$$

Machine Learning Methods

- Supervised Learning
  - Regression and Classification
  - Decision Trees and Random Forests
  - Neural Networks, …
- Unsupervised Learning
  - K-Means Clustering
  - Principal Component Analysis, …

Supervised Machine Learning

- Regression-based Methods
  - Generalized Linear Models
    - Linear Regression
    - Logistic Regression
- Deep Learning (Neural Nets)
- Tree-based Ensemble Methods
  - Random Forest (Bagging: Bootstrap Aggregation)
    - Parallel ensemble to reduce variance
  - Gradient Boost Machine (Boosting)
    - Sequential ensemble to reduce bias

Two Examples

- Regression Model: Predicting House Price
  - S. Mullainathan and J. Spiess, Machine Learning: An Applied Econometric Approach, Journal of Economic Perspectives 31:2 (87-106), Spring 2017.
- Classification Model: Credit Card Default
  - I. C. Yeh and C. H. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36:2 (2473-2480), 2009.

Example 1: Predicting House Prices

- Using 10,000 randomly selected owner-occupied units from 2011 metropolitan sample of American Housing Survey, we predict the (log) unit value with 150 features (e.g. number of rooms, the base area, and the census region within the U.S.).[download]
- Based on the separate hold-out set of 41,808 units from the sample, the predictions of different models are evaluated and compared.

Example 2: Credit Card Default

- This example aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default.
- The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Credit Card Default (Continued)

- This example employed a binary variable Y, default payment (Yes = 1, No = 0), as the response variable, with the following 23 variables as explanatory variables:
  - X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
  - X2: Gender (1 = male; 2 = female).
  - X3: Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).
  - X4: Marital status (1 = married; 2 = single; 3 = divorce; 0=others).
  - X5: Age (year).

Credit Card Default (Continued)

- X6 - X11: History of past payment. These variables track the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Machine Learning Using R

H2O Package

- Installation and Use of H2O Package
  - `install.packages("h2o")`
  - `library(h2o)`
  - `h2o.init()`
  - …
  - `h2o.shutdown()`