Class 2 Review Notes

AI & Machine Learning

Fall 2023 - Laurie Ye

2 Getting Started with Supervised Learning

2.1 Supervised Learning

1. Prediction: a set of explanatory x-variables are used to predict a quantitative y-variable.
2. Classification: the outcome is one of a (generally) small number of categories (e.g., spam or not, fraudulent or not, will click through or not, brand of a purchase, color, etc.).

2.2 Prediction

- Usually, the y-variable is continuous (e.g., income, sales, price, weight, etc.)
- But it also may be a count (# customers, # of clicks).

2.3 Classification

- Nominal variables or attribute variables are other terms used for a categorical variables.
- Sometimes the categories have a natural ordering (e.g., Ph.D. > Master's Degree > Bachelor's Degree).
- This is almost always done by first predicting the probability of each category. These predicted probabilities are then used to determine the particular category given as the "answer."

Note 1. In "real life," most variables are categorical & This is especially true for x-variables.

Note 2. X variables: the x's, independent variables, explanatory variables, regressors, predictors, treatment; Y variable: the y's, dependent variable, outcome variable, response.

2.4 Difference Between Machine Learning & Traditional Statistic

- In ML, the data sets are usually very large. Most data sets in traditional statistics are small to medium sized.
- In ML, we do not care very much about causal explanations. We only care if the procedure works (i.e., makes good predictions or accurate classifications). In traditional statistics, we generally care a lot about causal explanations.

- In ML, we will assess performance via out-of-sample performance.

- In ML, model building is generally done using out-of-sample performance. In traditional statistics, there are clear model-building strategies based on in-sample "goodness-of-fit" measures.

- In ML, there is more use of non-parametric models. In statistics, more emphasis is on parametric models.

2.5 Bias vs. Variance

1. Bias: Is your method on target on average?

$$\text{Bias} = E(\text{Statistic}) - \text{True Value}$$

2. Variance: How consistent is your method.

$$\text{Variance} = E\left[(\text{Statistic} - E[\text{Statistic}])^2\right]$$

3. Mean-Squared Error (MSE): Combines both.

$$\text{MSE} = E\left[(\text{Statistic} - \text{True Value})^2\right]$$

$$\text{MSE} = E\left[(\text{Statistic} - E[\text{Statistic}])^2\right] + [E(\text{Statistic}) - \text{True Value}]^2$$
$$\text{MSE} = \text{Variance} + \text{Bias}^2 \tag{1}$$

2.6 Assessing Performance: Hold-Out Samples

- Hold-out samples help preventing over-fitting.(Use of the validation data set to select the best model(s) usually prevents overfitting.)

- One common approach is to divide data into three groups: training data, validation data, and test data.

- About training data: Fitting a model means estimating the parameters (or coefficients) of the model (i.e., selecting according to some criteria like least squares). So the training data set is used to estimate the parameters of the competing models.

- About validation data: Performance on the validation data set is then used to select among the set of competing models.

- About eliminating models: The idea is that the cross-validation will eliminate models that are over-fitting the training data.

- About test data: Finally, the test data set is used to get a "pure" out-of-sample estimate of performance.

Note 3. Using the validation sample to estimate the performance of the best model will usually give an overly optimistic estimate of performance. Unbiased assessment of the performance of the model requires an independent sample, the test sample.

Note 4. Key idea is a collection of competing models are fit. For example, regressions with different sets of X-variables. Linear least squares, neural-nets, non-parametric regression (CART).

2.7 Reference

Goizueta Business School-Emory University: Professor George S. Easton