Introduction to Statistical Learning Notes

Abdullah Pakwashee

July 2020

1 Introduction

Honestly this is just here to save me from having Chapter 2 notes under latex section 1.

2 Chapter 2 Notes

2.1 What is Statistical Learning?

2.1.1 Why estimate f?

Prediction

- Using predictors $X_1, X_2, ...$ we can observe a quantative response Y,

$$Y = f(X) + \epsilon, \tag{1}$$

  such that ϵ is a random error term, independent of predictors with mean zero.
- Accuracy of prediction relies on *reducible error* and *irreducible error*.
- More appropriate statistical learning technique allows the reduction of reducible error.
- *Irreducible error* associated with ϵ cannot be predicted using inputs. Variability associated with ϵ cannot be reduced.
- ϵ can't be eliminated due to: unmeasured variables and immeasurable variation.
- Assume estimate $\hat{f}$ with predictors X, yielding prediction $\hat{Y} = \hat{f}$. The expected

value of the squared difference between the predicted and actual value of Y,

$$\begin{aligned} E(Y - \hat{Y}^2) &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

The error term limits the upper bound for the Y prediction accuracy.

Inference

- We want to understand the way Y is affected by $X_1, X_2...$
- Questions to ask are as follows:
  - *Which predictors are associated with a response?* Need to find the important ones.
  - *What is the relationship between the response and each predictor?* Positive/negative/ nuance?
  - *Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?*
- Is this a problem of *inference*, *prediction* or both?
  - *Inference*: Understanding more about how inputs relate to an output e.g. What increase will I get in sales for increasing TV advertising?
  - *Prediction*: Can I use the use a bunch of inputs to get an accurate output that lets me do stuff? e.g. target key customers.

2.1.2 How Do We Estimate f?

Bit of nomenclature first:

- n always corresponds to different data points
- x_{ij} refers to observation i, with predictors denoted by subscript j.
- y_i represents the response variable for the ith variable.
- We want to train the data to find a function such that $Y \approx \hat{f}(X)$
- Types of statistical learning methods:
  - Parametric Methods:
    1. Make an assumption about functional form e.g. $f(X) = \beta_0 + \beta_1 X_1 + ...$ All you have to do now is find $p + 1$ coefficients of $\beta_0, ..., \beta_p$.

2. Need a way to use the training data to train the model. Need to find the β_p for example from above. Common means is least squares.

Much easier to get a set of parameters than to determine an arbitrary function. Risk is we're not matching the 'true' form of f. Flexible models to fit different functional forms, but the risk is overfitting where the noise in the data is followed too closely.

– Non-parametric methods:

The other option is to not make assumptions about any functional form of f- get a smooth fitting as close to the data as possible. No assumptions about functional form so avoids a potential bad fit. However, since we're not reducing the problem to a few parameters, LOTS of observations are necessary to get an accurate estimate of f.

2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

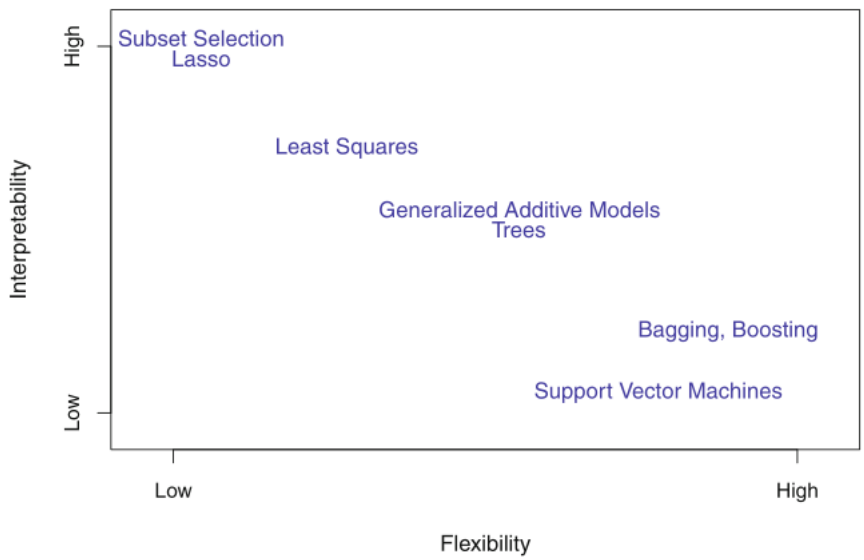

Figure 1: A representation of the trade-off between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.)

- Restrictive model might be better for inference situation. More flexible approach would make it harder to understand how a predictor is related to response.

2.1.4 Supervised Versus Unsupervised Learning

- Supervised learning: we have predictor measurements x_i (observations) with responses y_i that we wish to fit a model to. This model allows the prediction of the future or understanding how our various predictors affect the response.
- Unsupervised learning concerns having no response; you're effectively blind. One way to use this sort of data is cluster analysis. If I have a bunch of customer data on postcode and family income, we're gonna expect the data to cluster together. We might then find a surprising observation vector which clusters nicely with our other data, giving us new ways to target the customer. p variables leads to $p(p-1)/2$ scatter plots, so automated cluster methods are necessary.

2.1.5 Regression Versus Classification Problems

- Quantitative problems referred to as regression e.g. least squares linear regression.
- Qualitative problems referred to as classification e.g. logistic regression.
- Some SL methods can be used for either.
- Mostly can use methods for either, provided qualitative data is properly *coded* before analysing.

2.2 Assessing Model Accuracy

Let's explore the what's most important to consider when selecting a SL method for a given data set.

2.2.1 Measuring the Quality of Fit

- Most commonly used metric to measure fit in regression is *mean squared error* (MSE) where,
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{2}$$
and obviously we want it to be as small as possible.
- What we really care about is if it can accurately predict on new, unseen test data. Method that gives the lowest *test MSE* as opposed to the lowest training MSE.

- If no test data is available, use the method with the lowest training MSE? No guarantee lowest training MSE = lowest test MSE. Training sets often smaller than test sets.
- Think of Mr. Naylor's explanation of fitting different curves to the same data.

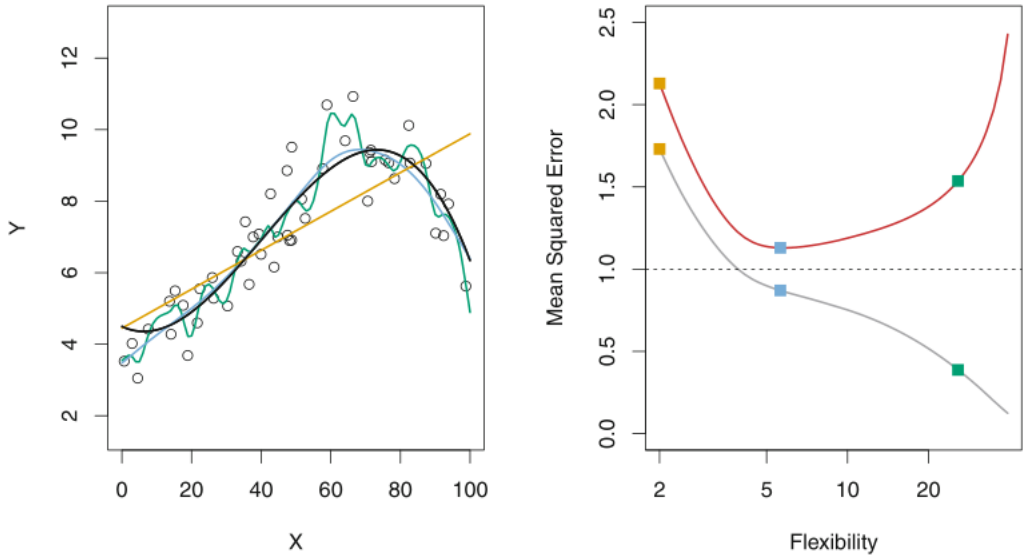

Figure 2: Note: Horizontal dashed line is $Var(\epsilon)$, the irreducible error, the lowest possible test MSE achievable.
Left: Data generated from black curves fit using linear regression (orange), and two smoothing spline fits (blue green).
Right: Grey line (training data) outlines MSE decreasing with increasing D.O.F but red line for test data shows that we minimise ≈ 5.

- *As model flexibility increases, training MSE will decrease, but the test MSE may not.*
- Small training MSE, large test MSE = overfitting. (training $<$ test typically, but specifically overfitting in the case with too many D.O.F- get a less flexible model!)
- Often hard to get test data, want to find that minimum point. Important method is cross-validation.

2.2.2 The Bias-Variance Trade-Off

- Expected test MSE can be decomposed

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{3}$$

- $E(y_0 - \hat{f}(x_0))^2$ represents repeatedly estimating f using lots of training sets, testing each at x_0, then averaging over all possible test values (all your $x_0 s$).
- Hence we want *low variance* and *low bias*. Both positive hence $Var(\epsilon)$ is indeed lower bound.
- Variance - how $\hat{f}$ would change if use trained on different data. Ideally doesn't vary much between training sets. More D.O.F more variance typically (less flexible).
- Bias- bad model will result in bad fit to the data, large bias. More flexible methods have less bias. More flexibility, increased variance, decreased bias. Need to find the optimum point.

2.2.3 The Classification Setting