A.1 Statistical Learning

Sources and Objectives

This lesson comes from ISLR Ch. 1 and section 2.1, and is designed to address SRM Syllabus learning objective 1(a).
The ISLR text is available for free in pdf form from the book's website. Additionally, this book has been written to a wide audience, and, although the lessons will not require it, I highly suggest that you consider reading the relevant sections along with the lessons.

Understanding Data

The goal of statistical learning is to understand data. So, what is data?

Data - a collection of recorded observations of certain aspects of a group of subjects, or of the same subject at different times.

Ex. The age, weight, height, and eye color for n individuals.

Ex. For 1 individual, the weight and height at each integer age.

Each measurement may be a *quantitative* variable with some *numerical* value as for age, weight, or height.

Or it may be *qualitative* or *categorical* variable such as for eye color, taking values from a set of *classes* or *levels*.

Notation: Index observations by $i = 1, \ldots, n$, and the measurements for each observation by $j = 1, \ldots, k$.

Supervised Statistical Learning

Suppose Y_i quantitative measurements depend on X_i measurements.

ex. Claim amounts depend on vehicle value.

$Y =$ **Response** or Dependent or Output Variable

$X =$ **Explanatory**, Independent or Input Var., Predictor, Feature

Same $X \longrightarrow$ different Y's $\longrightarrow$ Probabilistic model needed for Y.

$Y = f(X)+\epsilon$, with ϵ a random error term with mean zero.

-Typically, X not random $\longrightarrow \mathrm{E}[Y] = f(X)$

$f(X)$ is the "systematic information that X provides about Y"

Of course, $f(X)$ is usually unknown.

"Supervised Statistical learning"

$\longrightarrow$ methods for estimating f and evaluating the result.

Motivation: Prediction

When the response is not easily measured, an estimated f (called $\hat{f}$) can be used to predict Y from some new measurement X.

$\hat{Y} = \hat{f}(X) =$ **Predicted or fitted value of** Y

Here we don't care what f looks like, only that it make good predictions.

$Y - \hat{Y} =$ **Prediction Error**

$$= (f(X) + \epsilon) - \hat{f}(X) = (f(X) - \hat{f}(X)) + \epsilon$$

$f(X) - \hat{f}(X) =$ **reducible error** – $\hat{f}$ is an imperfect estimate

$\epsilon =$ **irreducible error**

Even if $\hat{f} = f$, it can't predict the variation of Y from unmeasured predictors or unmeasurable random variation

#SquadGoals: Minimize reducible error without *overfitting* into ϵ.

Motivation: Inference/Interpretation

Understanding the way that Y is affected by changes in X.

Here the form of $\hat{f}$ matters more than its predictive power.

Is there an association between Y and X at all?

What is the relationship between Y and X?

Is it reasonable to summarize the relationship using a linear function?

Less complex model (e.g. linear, or fewer features)

$\longrightarrow$ Better mechanistic interpretation, Weaker predictive power.

More complex model (e.g. non-linear, or more features)

$\longrightarrow$ Better predictions, less understanding.

The "best" model for any given setting may depend on the goals as much as on the data. There is no "best" model in any universal sense.

Estimation of Parametric f:

Training Data: Given paired $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, choose $\hat{f}$.

Parametric estimation assumes f takes specific form, e.g. $\beta_0 + \beta_1 X$.

Two primary methods for estimating f :

1. Ordinary Least Squares (OLS): Choose $\hat{\beta}$s by minimizing the sum of squared errors.
2. Maximum Likelihood (MLE): Choose $\hat{\beta}$s that maximize the probability of the response observations.

Characteristics of Parametric Models:

- Easier than estimating an arbitrary function f.
- Modeling error can result from using the wrong form for f.
- More flexible models (more parameters) can help to avoid Modeling error at the risk of overfitting.
- Usually preferred when inference is needed

Estimation of Non-Parametric f:

Training Data: Given paired $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, choose $\hat{f}$.

Non-parametric Methods don't assume a specific shape for f.

- Minimize reducible error subject to smoothness criteria
- Avoids modeling error since can fit any "true" f shape.
- Typically requires more data
- Often preferred when prediction is more important

Unsupervised Statistical Learning

Supervised: We view the organization of X through the lens of its association to Y

Unsupervised Learning refers to case where there is no Y.

- Focus is on the relationships between the Xs.
- One method is *Cluster Analysis* to determine whether the Xs fall into distinct groups.
- For $p = 2$ variables, one scatter plot can be used.
- with $p > 2$ variables, $p(p-1)/2$, 2-dim scatter plots could be produced $\longrightarrow$ need for automated methods if p is large
- *Principal Components Analysis* re-frames the data by directions of greatest variation

Semi-Supervised learning: Some data is missing a response measurement

Exercise 1

For a given set of data, it is estimated that $E[Y] = 0.234 - 0.096X$.
What is the fitted value of the response for new observation $X = 3$?
What is the prediction error corresponding to this new observation if the observed value of the response turns out to be -0.04?

Exercise 1

For a given set of data, it is estimated that $E[Y] = 0.234 - 0.096X$. What is the fitted value of the response for new observation $X = 3$? What is the prediction error corresponding to this new observation if the observed value of the response turns out to be -0.04?

$$\hat{Y} = 0.234 - 0.096 \cdot 3 = -0.054,$$

$$Y - \hat{Y} = -0.04 - (-0.054) = \boxed{0.014}.$$

Exercise 2

It is hypothesized that the height of an individual may be associated to the heights of the individuals parents. If a regression model is to be built to check this hypothesis, which of the following terms could be used to describe the height of the individual, and which the heights of the parents? Is this an example of a supervised or an unsupervised model?

Predictor Dependent Explanatory Feature

Output Response Input

Exercise 2

It is hypothesized that the height of an individual may be associated to the heights of the individuals parents. If a regression model is to be built to check this hypothesis, which of the following terms could be used to describe the height of the individual, and which the heights of the parents? Is this an example of a supervised or an unsupervised model?

Predictor Dependent Explanatory Feature

Output Response Input

Since we think that the height of the child *depends* on the heights of the parents, the height of the child is the Dependent variable, and the heights of the parents are independent variables.

We also say that the child height is the response or the output variable, while the parent heights are explanatory variables, or predictors or features of the model.

This is a supervised model, since there is an identified response variable among the data measurements.