Statistical Learning

Statistical Learning: a vast set of tools for <u>understanding data</u>

- **supervised**: building a statistical model for predicting an <u>output</u> based on one or more <u>inputs</u>
- **unsupervised**: learn relationships from structured data

response ~ Y
predictors ~ $(X_1, X_2, X_3, \dots X_n) = X$

$$\boxed{Y = f(X) + \varepsilon}$$

goal is to estimate f

error term

Estimating f

- prediction
- inference

Prediction:

our estimate

- Given X, we predic Y using: $\hat{Y} = \hat{f}(X)$

FIVE STAR

Accuracy of $\hat{Y}$ for prediction of Y depends on

irreducible error

reducible error

reducible error ↳ estimate will always have error b/c Y is also a function of ε

made better by picking better model

cannot be predicted with X (irreducible)

$$E(Y-\hat{Y})^2 = E[f(x)+\varepsilon-\hat{f}(x)]^2$$

$$= \underbrace{[f(x)-\hat{f}(x)]^2}_{\text{Reducible}} + \underbrace{Var(\varepsilon)}_{\text{Irreducible}}$$

we will focus on ways to estimate f & minimize reducible error

Inference: we are interested in understanding how Y is affected as $X_1, \ldots, X_p$ changes

↳ further analyze relationship between X & Y one may ask

- which predictors are associated with the response?
- what is the relationship between response & each predictor
- what model captures X & Y best?

Example:

Advertising Data

- which media contributes more to sales?
- how much would increase in sales associated with TV

How is the probability of purchase affected by the variables?

Modelling for Both Prediction & Inference

Real Estate Setting

inputs

- crimes
- income level
- size of house

...

you can look at these

$\Rightarrow Y$ you can simply predict

picking model for f

simple
- does not capture relationship of X & Y well (underfit)

complex
- does not generalize well (overfits)

How do we estimate f?

we will look at many linear & nonlinear methods

model: $f(x)$

$(x,y)_1, (x,y)_2 \, . \, . \, . \, . \, (x,y)_n$

(training data)

goal: apply a statistical learning method to the training data to estimate f

- parametric
- nonparametric

Parametric Method:

- 2 step approach:

1. make assumption about functional form f
2. fit/train the model

we estimate values of a set of parameters such as $\beta_0, \beta_1, \ldots$

Disadvantages?

- model we choose may not be correct
- may underfit or overfit

follow error term too closely

Non Parametric Methods:

Do not make assumptions about functional form of f

- they seek to make an estimate of f that gets close to the data points as possible without being too rough or wiggly

have a wide range of possibilities to fit for f

Disadvantage?

- very large # of obs required

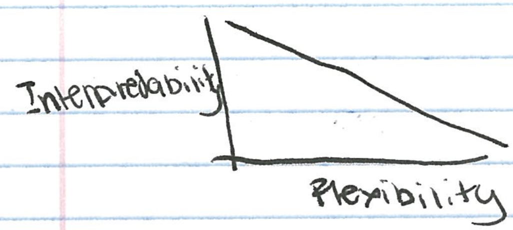
Trade off Between Prediction Accuracy & Model Interpretability

linear regression ⟶ inflexible

thin spline ⟶ wide range of shapes for f

Restrictive models ⟵ more interpretable for inference

hard to interpret when model is overly complex



Supervised vs Unsupervised

only x, no y: we seek to understand relationships between variables & observations
- clustering

Regression vs Classification

qualitative (discrete)

quantitative (continuous)

Assesing Model Accuracy

mse

ROC curve   etc, etc . . .

Bias Variance Trade-Off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\varepsilon)$$

to minimize MSE we need low variance & low bias

mse can never be below $Var(\varepsilon)$

Variance

The amount $\hat{f}$ would change if we used different dataset

- ideal $\hat{f}$ should not change alot between datasets

Bias

error introduced by approximating a real life problem

Good Test set performance require low variance
&
low bias

why Trade off?

b/c $\uparrow V$ & $\downarrow B$ is easy

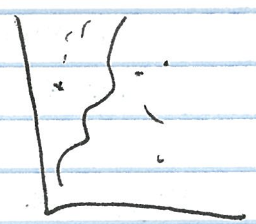
and

$\uparrow B$ $\downarrow V$ is easy

Classification Setting

$$\left.\begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}\right\} \text{qualitative}$$

error rate: $\frac{1}{n} \sum I(y_i \neq \hat{y}_i)$



Training Error
Test Error