Chapter 1 Statistical Learning

1. Supervised learning: Building a statistical model for predicting, estimating, an output based on one or more inputs

**Regression Problem:**

Predicting a continuous or quantitative output value

**Classification Problem:**

Predict a non-numerical value (a categorical or qualitative output) eg. Whether a stock will increase or decrease in a given day

**Clustering Problem:**  
Observe input variables, with no corresponding output. Group individuals according to their observed characteristics

1. **PCA (Principle Component Analysis)**
2. **Dimensions**

Statistical Learning

* 1. What Is Statistical Learning

**Inputs: predictors, independent variables, features**

**Output: response, dependent variable**

F represents the systematic information that X provides about Y

**Overall, the errors have approximately mean 0.**

Statistical learning refers to a set of approaches for estimating f.

* + 1. Why Estimate f?

**Prediction**

1. **A set of inputs X are readily available, but the Output Y cannot be easily obtained**
2. **Estimating function is treated as a black box**

**Reducible error**: Potentially improve the accuracy by using more appropriate statistical learning method

**Irreducible error**: because no matter how well we estimate f, we cannot reduce the error introduced by the error term

1. Unmeasured variables
2. Unmeasurable variation

Expected value of the square difference between predicted and actual value of Y

**Inference: Understand the ways that Y is affected by X**

1. Our goal is to estimate f, **but not necessarily make predictions**
2. **F function cannot be treated as a black box**
3. Which predictors are associated with the response
4. What is the relationship between y and x like

Some modeling could be conducted both for prediction and inference

1. Linear model: relatively simple and interpretable inference, but may not yield accurate predictions
2. Non-linear approaches: Provide quite accurate predictions, but at the expense of being less interpretable model for which inference is more challenging
   * 1. How do we estimate f
3. **Training data: use observations to train, teach our method on how to estimate f**

**Parametric Methods**:

1. **Make an assumption about the functional form, or shape of g**
2. After a model has been selected, need a procedure that uses the training data to fit or train the model
3. Disadvantage: chosen form will usually not match the true unknown form of f
4. **Overfitting: model follows the noise too closely**

**Non-parametric Methods:**

1. **Do not make explicit assumptions about the functional form of g**
2. Have the potential to accurately fit a wider range of possible shapes for f
3. Disadvantage: a very large number of observations is required in order to obtain an accurate estimate for f
   * 1. The Trade-Off Between Prediction Accuracy and Model Interpretability
4. **Restrictive models are more interpretable but less flexible//good for inference**
   * 1. Supervised versus Unsupervised Learning

**Supervised learning: for each observation of the predictor measurement, there is an associated response measurement**

**Unsupervised learning: For each observation, we observation a vector of measurements but no associated response y**

1. Cluster analysis

Semi-supervised learning: some predictors have response, but some predictors do not have response

* + 1. Regression vs. Classification Problem

1. **Quantitative variables: take on numerical values**
2. **Qualitative varibales: take on values in one of K different classes, or categories**
3. **Regression Problem: problems with a quantitative response**
4. **Classification Problem: problems with a qualitative response**

Assessing Model Accuracy

* + 1. Measuring the Quality of Fit

1. Mean Square Error (MSE)

A close up of a watch

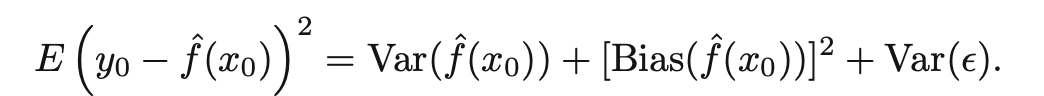
Description automatically generated

1. Whether f(x0) is approximately equal to y0, where (x0,y0) is a previously unseen test observation not used to train the statistical learning method
2. There is no **guarantee that method with the lowest training MSE will also have the lowest test MSE.**
3. Flexibility level corresponding to the model with the minimal test MSE can vary considerably among data sets

2.2.2 The Bias-Variance Trade-Off

The expected test MSE can be decomposed into the sum of three fundamental quantities:

1. **Variance of test function**
2. **Squared bias of the test function**
3. **Variance of the error term**



1. To minimize expected test error, we need **to select a statistical leaning method that simultaneously achieves low variance and low bias**
2. **Variance: the amount f would change if we estimated it using a different training data set. // more flexible statistical methods have higher variance**
3. **Bias: error introduced by approximating a complex problem with a simple model. //more flexible methods result in less bias**
4. **Bias Variance Trade-off:**
5. **As we increase the flexibility, the bias tends to initially decrease faster than the variance increase->MSE declines**
6. **At the inflection point, the impact on the bias starts to significantly increase the variance**

2.2.3 The Classification Setting

A picture containing object, clock

Description automatically generated

**Training error rate: the proportion of mistakes that are made if we apply our estimate to f to training observations**

A close up of a logo

Description automatically generated

**Test Error Rate: Predicted class label that results from applying the classifier to the test observation with predictor x0.**

**Bayes Classifier**

A picture containing object, clock

Description automatically generated

Assign a test observation with predictor vector to the class which the conditional probability is largest

**Bayes Decision Boundary: Bayes prediction determine by the bayes decision boundary**

Bayes Error Rate:

A close up of a logo

Description automatically generated

Similar to Irreducible Error

**K-Nearest Neighbors**

Given a positive integer K and a test observation x, the KNN classifier first identifies K point in the training data that are closest to x, represented by N. It then estimates the conditional probability for class j as the fraction of points in N whose values equal j:

A picture containing object, clock

Description automatically generated

The choice of K has a drastic effect on the KNN classifier obtained:

1. **When K=1, the decision boundary is overly flexible-> low bias but a very high variance**
2. **When k grows, the decision boundary becomes less flexible-> high bias but very low variance**
   1. Lab: Introduction to R
      1. Basic Commands:

**C()**: concatenate, create vectors

**<-/=**: assingnment

**?funcname**: open a new help file window

**Length()**: return length

**Ls():** return a list of all objects that have saved so far

**Rm():** delete any that we don’t want

**Remove(list=ls()):** remove all objects

**matrix(data=vector object, nrow=int):**

**Default**: creates function by successively filling in columns

**Byrow=**True: populate matrix in order of the rows

**Sqrt():** returns the square root of each element of a vector or matrix

**Rnorm(int n, mean=x,sd=y):** generates a vector of random variables with sample size n

Deafault: mean=0, sd=1

**Set.seed(int x):** allow the rnorm to produce certain results

**Cor():** compute the correlation between two functions

**Mean(vector/matrix):** return the mean

**Var(vector/matrix):** return the variance

**Sqrt(var(functiom))/sd():** return the standard deviation

* + 1. Graphics

**Ploty(x, y):** return a scatterplot of x versus y

**Xlab=/ylab=:** passing labels to x and y axis

**Main=:** pass title to the plot

**Pdf(titlename):** create a pdf with name titlename

**Seq(a,b, length=x)/a:b :** return a sequence of x numbers equally spaced between a and b

**Contour(vector1:xValues, vector2: yValues, vector3: zValues):** returns a contour plot that represents three-dimensional data

**Image()**: works the same as contour, returns a heatmap

**Persp(, theta=, phi=):** used to produce a three-dimensionlal plot, theta and phi used to control the angles

* + 1. Indexing Data

**Matrixname[nrow=x, ncol=y]:** accesing the element in the second row and third column

The first argument always refers to rows

The second argument always refers to columns

**Dim():** outputs the number of rows followed by the number of columns of a given matrix

* + 1. Loading Data

**Read.table():** importing a dataset into R

1. **Header=True**: first line of the file contains the variable names
2. **Na.string=:** any occurrence of particular character should be treated as null

**Write.table():** export data

**Fix():** View data in a spreadsheet like window

**Read.csv(file path)**: importing a csv file

**Header=**True;

**Na.strings=;**

**Na.omit(dataset):** omit the missing rows

**Dim():** returns the total number of observations

**Names():** check the variable names

Additional Graphical and Numerical Summaries

**Plot():** produce scatterplot of the quantitative variables

Plot specific variables and variable names:

Dataset$variableName OR

**Attach(dataset);**

**Plot(variable1, variable2)**

If the variable on the x-axis is categorical, then boxplots will automatically be produced

**Col=:** color

**Varwidth=:** width

**Horizontal=**: horizontal axis

**As.factor():** converts quantitative variables into qualitative variables

**Hist():** plot a histogram

**Pairs(**): creates a scatterplot matrix, i.e. scatterplot for every pair of variables for any given data set

**Identify():** the x-axis variable, the y-axis variable

**Summary():** produces a numerical summary of each variable in a particular data set.

**Q()**: shut down R

**Savehistory():** save all the commands in the most recent session

**Loadhistory():** load the saved history