Chapter 2

Statistical Learning

2.1 What is statistical learning

Example 1. *The sales of a product and amount spent on TV, radio, newspaper ads was measured across 200 different markets. We want to understand how ads affect sales.*

Code

```
> # Install and load the ISLR package (any CRAN mirror works)
> # Type data() to see data sets
> install.packages("ISLR")
> library(ISLR)
> data()
> # Highlight and copy the data (with header) then run the following code
> Advertising<-read.table("clipboard",header = T)
> dim(Advertising)
[1] 200   4
> Advertising[1:5,]
     TV radio newspaper sales
1 230.1  37.8      69.2  22.1
```

```
2  44.5  39.3      45.1  10.4
3  17.2  45.9      69.3   9.3
4 151.5  41.3      58.5  18.5
5 180.8  10.8      58.4  12.9
> plot(Advertising)
> attach(Advertising)
> plot(radio, sales)
> abline(lm(sales~radio), lwd = 5)
```

- What are the three independent variables / explanatory variable/ inputs X_1, X_2, X_3.

  TV, Radio, Newspaper

- What is the dependent variable/ response variable/ output variable Y?

  Sales

- We in general have a model $Y = f(X_1, ..., X_p) + \epsilon$
  - f is the systematic information that $X = (X_1, ..., X_p)$ provides about Y
  - ϵ is error that is independent of X and has mean 0.
- What are some potential choices for f for the data?

  linear, polynomial, cound use some not all variable etc.

Example 2. *Income, years of education, and "seniority" was simulated for 30 individuals.*

Code

```
> Inc<-read.table("clipboard",header = T)
> dim(Inc)
[1] 30  3
> Inc[1:3,]
  Education Seniority   Income
1  21.58621  113.1034 99.91717
2  18.27586  119.3103 92.57913
3  12.06897  100.6897 34.67873
> attach(Inc)
> plot(Seniority, Income)
> plot(Education, Income)
```

- What are the independent variables / features / inputs?

  Education

  Seniority

- What is the response variable/ dependent variable/ output variable?

  Income

- You might again think that f is linear in $Y = f(X_1, X_2) + \epsilon$ but in fact its not. Here, f was known to be nonlinear (the data was simulated). In practice, do we know f?

2.1.1 Why estimate f

Prediction

- Goal: predict Y for a given value of X via $\hat{Y} = \hat{f}(X)$ where $\hat{f}(X)$ is an estimate of $f(X)$.

- Merit Criteria: Minimize the expected prediction error

$$E(Y - \hat{Y})^2 = E(f(X) - \hat{f}(X))^2 + Var(\epsilon)$$

  - Which term is reducible?

E(f(x) - fhat(x))^2)

  - Which term is irreducible?

Var(e)

- What are f and $\hat{f}$ below? Observe $\hat{Y} \neq Y$.

Code

```
> # Simulation
> set.seed(1)
> x<-rnorm(20)
> y<- x+rnorm(20) ---> f(x) = X
> plot(x,y)
```

```
> f.hat<-lm(y~x)
> abline(0, 1, lwd = 3)
> abline(f.hat, lwd =3, col =2, lty =2)
> round(cbind(y,f = x, f.hat=predict(f.hat), error=resid(f.hat)),2)
       y     f f.hat error
1   0.29 -0.63 -0.46  0.76
2   0.97  0.18  0.18  0.79
3  -0.76 -0.84 -0.63 -0.13
4  -0.39  1.60  1.30 -1.69
5   0.95  0.33  0.29  0.66
6  -0.88 -0.82 -0.62 -0.26
7   0.33  0.49  0.42 -0.09
8  -0.73  0.74  0.62 -1.35
9   0.10  0.58  0.49 -0.39
10  0.11 -0.31 -0.21  0.32
11  2.87  1.51  1.23  1.64
12  0.29  0.39  0.34 -0.05
13 -0.23 -0.62 -0.46  0.23
14 -2.27 -2.21 -1.72 -0.55
15 -0.25  1.12  0.92 -1.18
16 -0.46 -0.04  0.00 -0.46
17 -0.41 -0.02  0.02 -0.43
18  0.88  0.94  0.78  0.10
19  1.92  0.82  0.68  1.24
20  1.36  0.59  0.50  0.85
```

linear model fhat(x) = \beta_not hat + \beta1x) != 0+ 1x

observe/verify that

y = fhat(x) + epsilon hat

Inference

- Main goal: Understand how changes in X affect Y through $f(X)$.
- Merit Criteria: Minimize the expected prediction error *but*, f should be easy to interpret.
- Some specific goals

  - Variable selection

  - Nature of association between X_j and Y

  - Interactions between X_j and X_k

- Example: In the advertising data, where X_1, X_2, and X_3 are money spent on advertising media TV, radio, and newspaper ads and Y is sales, describe how you might perform an inference to answer these questions

  - Is sales associated newspaper ads?

  - How strongly is sales associated with newspaper ads?

  - What is the nature of the association?

  - Which media has highest return rate?

– If I increase TV spending by 10K how much will sales increase?

2.1.2 How to estimate f

- We estimate f with training data $(x_1, y_1), ..., (x_n, y_n)$ where $x_i = (x_{i1}, ..., x_{ip})^T$.

- Some assumptions must be made about the form for f.

parametric f

- What is the usual linear model for f?

- What are the parameters?

- Can you think of some estimation methods?

- lm in R gets least squares estimates for a linear model. What are f and $\hat{f}$? below?

Code

```
> model<-lm(Income ~ Seniority + Education)
> model

Call:
lm(formula = Income ~ Seniority + Education)

Coefficients:
(Intercept)    Seniority    Education
   -50.0856       0.1729       5.8956
```

nonparametric f

- A nearest neighbor approach would predict y by

  1. Compare X to $x_1, x_2, ..., x_n$ to determine the "nearing neighbor", i.e. which x_i is closest to X.

  2. If X is the closest to x_i, then take $\hat{Y} = y_i$.

- Example: Suppose that you have (x_i, y_i) given by $(1, 2), (2, 7), (3, 25), (2.8, 19)$ and wish to predict Y for $X = 2.5$. What's $\hat{Y}$?

2.1.3 Prediction accuracy vs. model interpretability

- Consider the nearest neighbor approach vs. the linear model approach for the Income data set. Which should be used to answer the following questions.
  - Which variable is most important: Seniority or Education?
  - Does income level increase as seniority increases at the same rater for all education levels?
  - What should I expect to make if I have 3 years of college and my seniority level is 25?
- In general there is a tradeoff between prediction accuracy and model interpretability. The following methods we'll consider are listed from least flexible (most interpretable) to most flexible (least interpretable)
  - Subset selection and LASSO (least absolute shrinkage and selection operator)

  - Least squares estimation (full linear model)

  - Generalized additive models (GAM) and regression trees

  - Bagging and Boosting

  - Support vector machine (SVM) with nonlinear kernels

2.1.4 Regression vs. classification

- Its typically important to know whether **y** is continuous (quantitative) or categorical (qualitative).

  - If y is continuous then we have a *regression problem*. Otherwise our objective is *classification*.

- Some methods can be applied to categorical or continuous responses and some cannot / should not.

  - Can the nearest neighbor prediction approach work for continuous or categorical y?

- What about the logistic regression (for classification) vs. the usual regression below. Answer the following questions for the simulated data to determine if linear regression is reasonable for categorical data.

Code

```
> # Logistic regression vs. regular linear regression
> set.seed(1)
> x<- seq(-3,3, .1)
> y<-rbinom(length(x), 1, prob=exp(x)/(1+exp(x)))
> plot(x,y)
> curve(exp(x)/(1+exp(x)), lwd =3, add = T)
> logistic.reg<-glm(y~x, family = "binomial")
> linear.reg<-glm(y~x)
> logistic.reg

Call:  glm(formula = y ~ x, family = "binomial")

Coefficients:
(Intercept)            x
    -0.3067       1.1557

Degrees of Freedom: 60 Total (i.e. Null);  59 Residual
Null Deviance:      84.15
Residual Deviance: 51.21        AIC: 55.21
> linear.reg

Call:  glm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.4590       0.1898

Degrees of Freedom: 60 Total (i.e. Null);  59 Residual
Null Deviance:      15.15
Residual Deviance: 8.332        AIC: 57.67
> abline(.4590,.189, lty = 2, lwd =3, col=2)
> curve(exp(-.31 + 1.2*x)/(1+exp( -.31 + 1.2*x)), add = T, lwd=3, lty =3, col=3)
```

– Because Y is 0 or 1, we aim to estimate $f(X) = \Pr(Y = 1|X)$. Identify $f(X)$ and the two estimated models.

– Which model assumes y is continuous?

– Which model assumes y categorical?

– Which model always gives reasonable estimates for $f(X)$?

2.2 Assessing model accuracy

2.2.1 Mean squared error

In statistical learning we have *training data* that is used to estimate f and *test data* that is used to estimate how well the estimated model works.

Training MSE

Definition 1. *The mean squared error (MSE) on the training data is* $MSE = \frac{1}{n}\sum_i (y_i - \hat{f}(x_i))^2$

Code

```
> # MSE
> set.seed(1)
> x<-1:3
> y<-rnorm(3,mean =x)
> plot(x,y)
> my.lm<-lm(y~x)
> my.lm

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -0.2170       0.8954

> abline(my.lm)
> xsq<-x*x
> my.qm<-lm(y~x+xsq)
> my.qm

Call:
lm(formula = y ~ x + xsq)

Coefficients:
(Intercept)            x          xsq
    -3.2659       4.5542      -0.9147

> curve(-3.2695 + 4.5542*x - .9147*x^2, add = T, col =2, lty=2)
```

- What is the true f above?

- Two different $\hat{f}$'s are estimated. Write them out.

- Which model has the smallest training MSE?

- In general for n pairs (y_i, x_i) we can always find an $(n-1)$th degree polynomial that has 0 training MSE. In regression model building, this results in an $R^2 = 1$. Why don't we just always use this model then?

Test MSE

Definition 2. *The MSE on the test data is* $Ave(y_0 - \hat{f}(x_0))^2$ *where* (y_0, x_0) *represents test data and "Ave" is some average. Note that training data is used to get* $\hat{f}$*.*

- Which $\hat{f}$ will have smaller test MSE in the above example?

- Sketch a plot with flexibility on the X axis, MSE on the Y axis and draw curves representing the anticipated test MSE and training MSE vs. flexibility.

- When an estimated model $\hat{f}$ has small training MSE and large test MSE then we are *overfitting* the data. Identify overfitting on the sketch above.

Cross Validation

- In practice we'll need to use some of the training data to estimate f and other test data to estimate the test MSE. This is called *cross validation.*

- In classical statistics, we use mathematical arguments to derive optimal estimators for parameters in f. As f gets more flexible, however, the math gets more daunting.

  - Example: The Gauss-Markov Theorem provides the best linear unbiased estimator of parameters in $f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.

  - Example: In generalize linear models, we relax linearity. What technique do we use to derive optimal estimators here?

2.2.2 Bias vs. variance

Definition 3. *The expected test MSE can be decomposed*

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

- *where* $Bias(\hat{f}(x_0)) = E[\hat{f}(x_0)] - E[y_0]$ *and*
- $Var(\hat{f}(x_0)) = E[\hat{f}(x_0) - E(\hat{f}(x_0))]^2$

Illustration

Model 1: Suppose $y_1, ..., y_n$ are iid $N(\mu, 1)$ or $y_i = \mu + \epsilon_i$ for $\epsilon_i \sim N(0, 1)$. Assume μ is unknown.

- Consider predicting y_0 with $\bar{y}$
  - What is $Var(\bar{y})$ and what happens as $n \to \infty$
  - What is $Bias(\bar{y})$?
  - What is $Var(\epsilon)$?

- What if we had overfit and predicted y_0 with $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$ where parameters are estimated via *least squares* (still assume $\beta_0 = \mu$ and $\beta_1 = 0$ so that the data are generated the same way.)

  - We know from regression that is $Var(\hat{f}(x_0)) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\hat{\beta}_0) + x_0^2 Var(\hat{\beta}_1) \to 0$ as $n \to \infty$.

  - Note that least squares estimators are unbiased. What is the bias then of $\hat{f}(x_0)$?

  - What is $Var(\epsilon)$?

  - Which estimator has a smaller MSE for fixed n?

  - When will the discrepancy in MSE's across models be largest (small or large n?)

Model 2: Suppose $y_1, ..., y_n$ are independent $N(x_i, 1)$.

- What if we predict y_0 with $\bar{y}$ but in fact $y_0 \sim N(x_0, 1)$. Note that $\bar{y} \sim N(\bar{x}, 1/n)$.
  - Note that $Var(\bar{y}) \to 0$
  - Note that $Var(\bar{y}) < Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) \to 0$
  - What is $Bias(\bar{y})$ vs. $Bias(\hat{\beta}_0 + \hat{\beta}_1 x_0)$?
  - What is $Var(\epsilon)$?

Summary of Bias vs. Variance

- The reducible portion of the MSE is decomposed into two measures. What are they?

- Consider three models ranging from least flexible to most flexible f_1, f_2, f_3. Plot the MSE on the y axis and the sample size on the x axis. Identify the asymptotes.

2.2.3 Classification

The main ideas in our regression examples carry over to classification, but now y_i represents a class label (population A vs. population B, $y_i \in \{0, 1\}$, etc.)

Classification Error

Definition 4. *Let (x_i, y_i) be training data where y_i is a class label and let $\hat{y}_i$ be the predicted value for y_i. The training error rate is $MSE = \frac{1}{n}\sum_i I(y_i \neq \hat{y}_i)$.*

Example 3. *This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.*

Code

```
> # Iris classification illustration
> data(iris)
> iris2<-iris[51:150,]
> attach(iris2)
> plot(Petal.Length,Sepal.Width, pch = c(rep(1,50), rep(2,50)),
+ col = c(rep(1,50),rep(2,50)))
```

- Here, we could try to classify a species as Versicolor (circle) or Virginica (triangle) using inputs Sepal Width and Petal Length. Draw a line in the plot that best separates the species, and estimate / eyeball the training classification error.

Definition 5. *The test error rate for test data* (x_0, y_0) *is* $Ave(I(y_0 \neq \hat{y}_0))$.

- We can consider much more complex classifiers than the linear classifier. Should we choose a classifier that minimizes the training error rate or test error rate?

Bayes Classifier

Definition 6. *The Bayes classifier predicts* y *to be in the* j*th class if*

$$\Pr(Y = j|X = x_0)$$

is larger than $\Pr(Y = k|X = x_0)$ *for all other* k *classes.*

- We can estimate the above probabilities under a multivariate normal model.
  - The line in the iris plot above actually resents values of x_0 where

    $$\widehat{\Pr}(Y = versicolor|X = x_0) = \widehat{\Pr}(Y = virginica|X = x_0) = 1/2$$

    where the probabilities are estimated under a normal model.

Code

```
> library(MASS)
> lda.model<-lda(Species ~ Petal.Length + Sepal.Width, data = iris)
> predict(lda.model)$posterior[1:5,]
  setosa   versicolor    virginica
1      1 8.482780e-16 9.379239e-28
2      1 1.605552e-12 4.441676e-24
3      1 1.104325e-14 9.975726e-27
4      1 2.522903e-12 1.232883e-23
5      1 1.875517e-16 1.726203e-28
```

- The nearest neighbor classifier is a nonparametric version of the above classifier. It estimates the probability as the proportion of the k nearest neighbors that have class j, i.e.

$$\widehat{\Pr}(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}(x_0)} I(y_i = j)$$

  - Draw a circle around one of the points that contains the 3 nearest neighbors and get the estimate of $\Pr(Y = j|X = x_0)$.

  - What is the training error rate if $k = 1$?

  - Would small or large k tend to result in overfitting?

  - Consider a test observation on the plot, and suppose $y_0 =$ versicolor. What happens to the test error rate as k increases?