Chapter 12

Logistic Regression and Machine Learning

12.1 Introduction to Machine Learning

■scalable statistical algorithms that combine

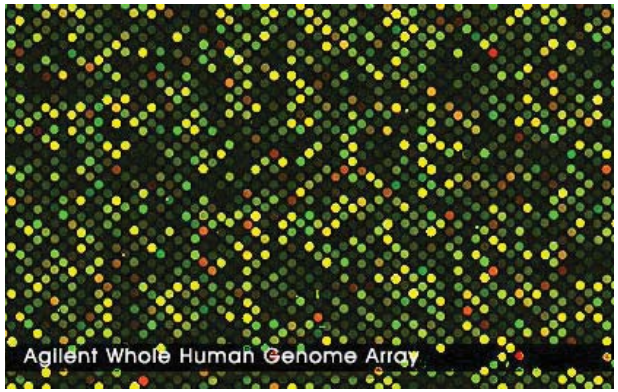
★expertise from statistics on how to extract information from data with

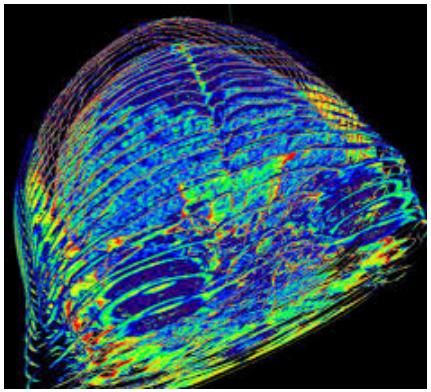
★computational ideas that enable efficient implementation on large data sets.

Data revolution: Enormous datasets are routinely collected

- **Biological Sci.**: Genomics, genetics, neuroscience, medicine

- **Natural Sci.**: Astronomy, earth sciences, meterology.





- **Engineering**: Machine learning, surveil. videos, social media

- **Social Sci**: Economics, finance, marketing, managements.

■Characterize many contemporary scientific and decision problems.

Example 12.1 *Supervised learning — classification*

Labels are provided. Document classification, disease classification, face recognition.

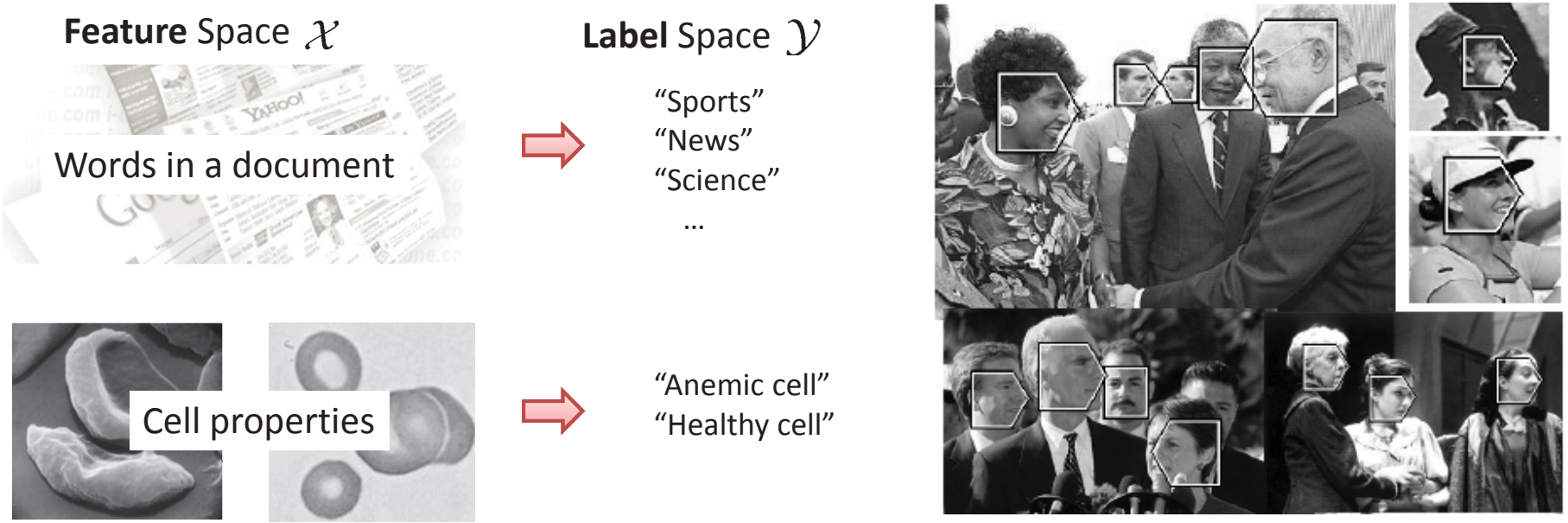


Figure 12.1: Some supervised learning problems

Example 12.2 *Unupervised learning — clustering*

Figure 12.2: Some unsupervised learning problems

No labels provided: Social network and animal phylogenetic tree

Example 12.3 *Gene expression and autism*

Over 60K gene expression profiles (Next Generation Sequencing) are measured among 104 samples: 47 autisms and 57 healthy controls, along with gender, brain region, age, and sites. Of interest is to find the genes that are associated with autism.

12.2 Logistic regression

Modeling binary data: Suppose that latent variable (e.g. severity of disease such as autism and cancers) follows

$$Z = \beta_0^* + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon = \beta_0^* + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon,$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ and $\mathbf{X} = (X_1, \cdots, X_k)^T$. Instead of observing Z, we get $Y = I(Z < c)$ for a threshold c.

Conditional probability: if $\varepsilon \sim G$

$$\begin{aligned} P(Y = 1 | \mathbf{X} = \mathbf{x}^*) &= P(\beta_0^* + \mathbf{x}^{*T}\boldsymbol{\beta} + \varepsilon < c) \\ &= G(c - \beta_0^* - \mathbf{x}^{*T}\boldsymbol{\beta}) \\ &= F(\beta_0 + \mathbf{x}^{*T}\boldsymbol{\beta}) \end{aligned}$$

where $F(x) = G(-x)$ and $\beta_0 = \beta_0^* - c$.

Link function: $F^{-1}(\cdot)$ is called link. Commonly used examples:

★ **logit link**: $F(x) = \frac{\exp(x)}{1+\exp(x)}$, $F^{-1}(p) = \log \frac{p}{1-p}$ —logit function

$$P(Y = 1|\mathbf{X} = \mathbf{x}^*) = \frac{\exp(\beta_0 + \mathbf{x}^{*T}\boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{x}^{*T}\boldsymbol{\beta})},$$

★ **probit link**: $F(x) = \Phi(x)$, normal cdf.

Observed data: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\quad y_i =$ binary.

pmf for Bernoulli: $P(Y = y) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases} = p^y q^{1-y}.$

MLE: Find β_0 and $\boldsymbol{\beta}$ to maximize

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} q_i^{1-y_i}$$

where $p_i = F(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})$. Its log-likelihood is

$$\ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(p_i/q_i) + \log q_i$$

Logistic regression: $p_i = \frac{\exp(\beta_0 + \mathbf{x}_i\beta)}{1+\exp(\beta_0+\mathbf{x}_i\beta)}$ and $q_i = \frac{1}{1+\exp(\beta_0+\mathbf{x}_i\beta)}$ **(logit link)**. Find β_0 and $\boldsymbol{\beta}$ to maximize

$$\ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\beta_0 + \mathbf{x}_i\beta) - \log(1 + \exp(\beta_0 + \mathbf{x}_i\beta)).$$

Solution: β_0 and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_k)'$, by convex optimization.

Predicted probability: $P(Y = 1|\mathbf{x} = \mathbf{x}^*) = F(\widehat{\beta}_0 + \mathbf{x}^*\widehat{\boldsymbol{\beta}})$.

Example 12.4 *Sex classification using heights*

User profile data for 59,946 San Francisco OkCupid users (a free online dating website) from June 2012 are recorded.

```
logistic.model <- glm(is.female ~ height, family=binomial, data=profiles)
summary(logistic.model)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  44.9999     1.1374    39.6   <2e-16 ***
height       -0.6705     0.0169   -39.8   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 8075.1  on 5994  degrees of freedom
Residual deviance: 4460.2  on 5993  degrees of freedom
AIC: 4464
Number of Fisher Scoring iterations: 6
```

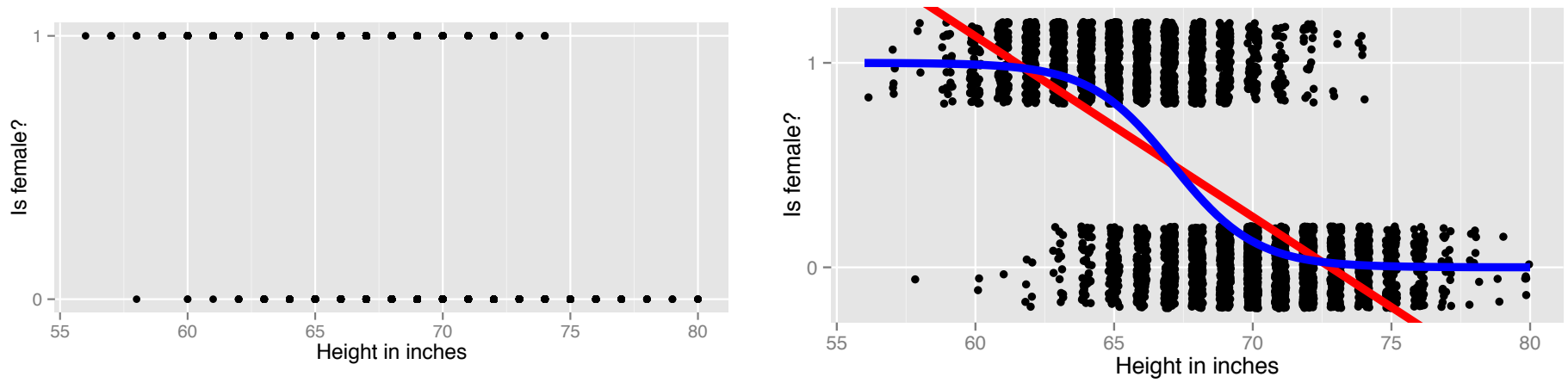


Figure 12.3: Left panel: A random sample of 40 points. Left: Jittered data and fitted probabilty of being female given the height, compared with linear regression

Ex. 12.3 (cont.). We select top 5 differently expressed by using two-sample t-test and fit logistic regression along with other variables.

```
> autism  = read.csv("autism.csv")    #reading the data
> aut.glm = glm(Autism ~ . , family=binomial, data=autism)
>    #fitting the model
> summary(aut.glm)   #summarize the fit

Call:
glm(formula = Autism ~ ., family = binomial, data = autism)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.4105  -0.5834  -0.1647   0.4863   2.5613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.33425    2.56463  -0.520 0.602889
GenderM      0.14585    0.73279   0.199 0.842233
Age         -0.05945    0.02871  -2.071 0.038365 *
SiteM       -3.43602    0.95416  -3.601 0.000317 ***
Reg          1.17445    0.57933   2.027 0.042636 *
Gene1       -0.10237    0.14148  -0.724 0.469332
Gene2        0.43250    0.32752   1.321 0.186658
Gene3        0.78675    0.26275   2.994 0.002751 **
Gene5       -0.66137    0.30426  -2.174 0.029729 *
NA.          0.08676    0.26373   0.329 0.742165
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 143.212  on 103  degrees of freedom
Residual deviance:  74.617  on  94  degrees of freedom
AIC: 94.617
```

We now select model by using stepwise procedure `step(aut.glm)`. It selects the model:

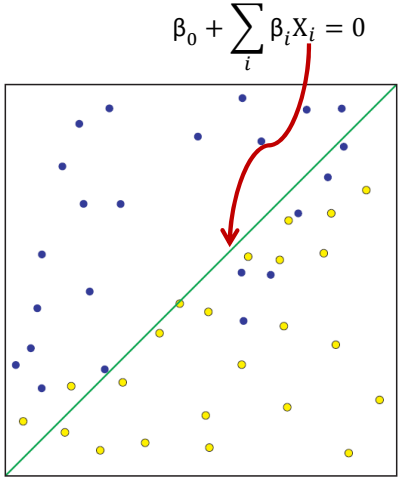
```
> aut.glm1 = glm(Autism ~ Age + Site + Reg + Gene3 + Gene5,
                 family=binomial, data=autism)
> summary(aut.glm1)     #summarize the fit

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.01125    2.12388   0.005 0.995773
Age         -0.06377    0.02804  -2.275 0.022928 *
SiteM       -3.31923    0.85777  -3.870 0.000109 ***
Reg          1.05110    0.52212   2.013 0.044099 *
Gene3        0.89643    0.22623   3.962 7.42e-05 ***
Gene5       -0.51391    0.18172  -2.828 0.004684 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

12.3 Classification

Classification: Given $\mathbf{X} = \mathbf{x}$, classify it as "class 1" if

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = F(\widehat{\beta}_0 + \mathbf{x}^T\widehat{\boldsymbol{\beta}}) > 0.5$$



Example 12.5 *Logistic regression and linear decision boundary*

Classifier is the same as $I(\widehat{\beta}_0 + \mathbf{x}^{*T}\widehat{\boldsymbol{\beta}} > 0)$.

Misclassification rate: Given n^* test data, it is defined by

$$\frac{1}{n^*}\sum_{i=1}^{n^*} I(\widehat{y}_i \neq y_i^*) \qquad \text{—Hamming distance}$$

Ex. 12.3 (cont.). We now use the second model to classify.

```
> logit = predict(aut.glm1)              #fitted log(odd-ratios)
> prob = exp(logit)/(1+exp(logit))     #fitted probability
> classification = (prob > 0.5)        #classification
     ### equivalent to directly using   (logit > 0)
```

```
> mean(autism[,1] != classification)   #compute misclassification rate
[1] 0.1346154
```

Bayes classifer: $f_B(\mathbf{x}) = \text{argmax}_y P(Y = y|\mathbf{X} = x)$.

Likelihood and prior: By Bayes formula,

$$P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\overbrace{P(\mathbf{X} = \mathbf{x}|Y = y)}^{\text{likelihood}}\overbrace{p(Y = y)}^{\text{prior}}}{P(\mathbf{X} = \mathbf{x})}$$

Optimal choice depends only on the numerator.

Risk: For a classifier $f(\mathbf{X})$, its risk is $R(f) = P(f(\mathbf{X}) \neq Y)$

Example 12.6 *Fisher Discriminants:* **Normal** *populations*

Assume $P(Y = 0) = 0.5$ and for population $y = 0$ or 1,

$$P(\mathbf{X} = \mathbf{x}|Y = y) = \frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}_y|}}\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T\mathbf{\Sigma}_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)).$$

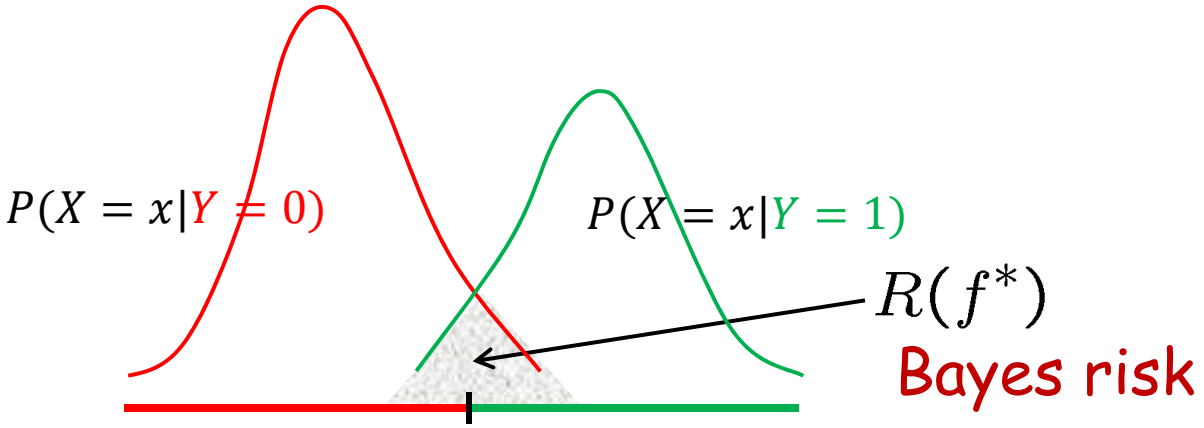


Figure 12.4: Illustration of Bayes classifier and its associated risk when $P(Y = 0) = 0.5$. It compares the likelihood ratio. Green data is classified as green on the left.

Then Bayes rule is the log-likelihood ratio: class 0 if

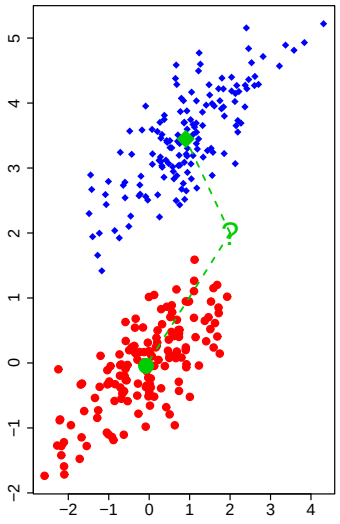
$$(\mathbf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \log |\boldsymbol{\Sigma}_0| \leq (\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \log |\boldsymbol{\Sigma}_1|.$$

This is a **nearest centroid** classifier.

When $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, it becomes

$$(\mu_1 - \mu_0)^T \boldsymbol{\Sigma}^{-1} (x - (\mu_0 + \mu_1)/2) < 0.$$

called **Fisher linear discriminant**.



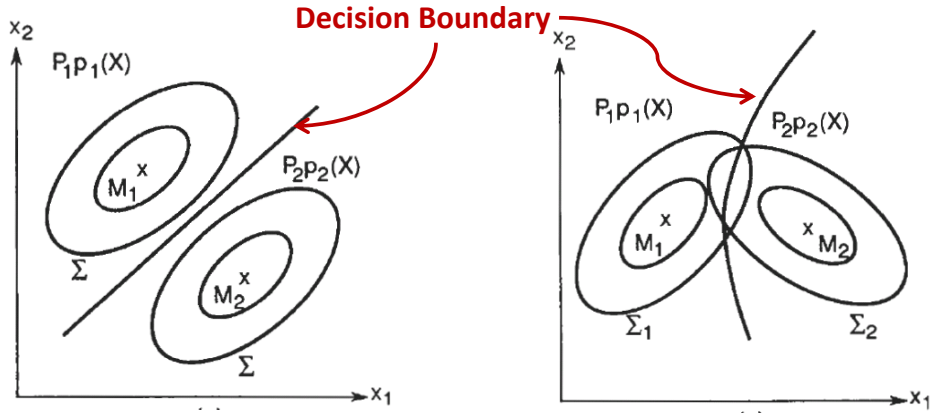


Figure 12.5: Illustration of decision boundary by nearest centroid classier. Case 1 ($\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$): a linear decision boundary. Case 2 ($\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$): Quadratic decision boundary.

12.4 Support Vector Machine

Relabel y as ± 1. Let $f(\mathbf{x})$ be a classifier with decision $\text{sgn}(f(\mathbf{x}))$.

Zero-one Loss: $L(f(\mathbf{x}), y) = I(y * f(\mathbf{x}) \leq 0)$

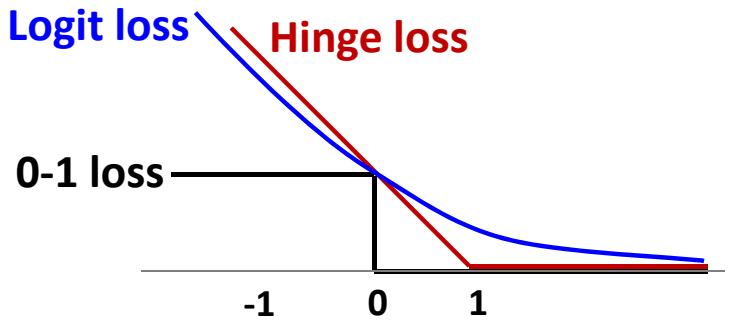
Logit regression: $\log p/q = f(x)$. Then,

$$\begin{cases} -\log p, & \text{if } y = 1 \\ -\log q, & \text{if } y = -1 \end{cases}$$

$$L(f(\mathbf{x}), y) = \log(1 + e^{-y*f(x)}) =$$

Support Vector Machine:

Use hinge loss $L(f, y) = (1 - y * f)_+$



Estimation: Find β_0 and $\boldsymbol{\beta}$ to minimize the emporical loss

$$\sum_i L(f(\mathbf{x}_i), y_i) = \sum_i L(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, y_i)$$

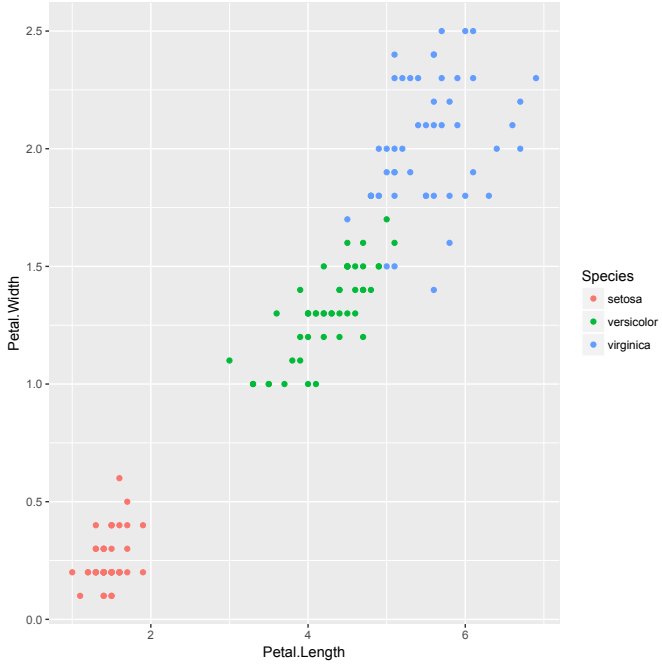
12.5 Clustering

k-means algorithm: Find clusters $\{C_j\}$ and centroids $\{c_j\}$ to min $\sum_{j=1}^k \sum_{i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{c}_j\|^2$

★ Given k cluster centers $\{c_j\}_{j=1}^k$, classify each data into k clusters by the nearest centroid.

★ Given k clusters, update cluster centers by taking their averages.

★ iterate until convergence.




```
> library(datasets)       #get the data set "iris"
> iris[1:3,]              #first 3 cases of the data
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
> irisCluster <- kmeans(iris[, 3:4], 3, nstart = 20)
   #20 random initial choices of centroids, using variables 3 and 4
```

```
> irisCluster             # show the resulds
K-means clustering with 3 clusters of sizes 50, 52, 48

Cluster means:
  Petal.Length Petal.Width
1     1.462000    0.246000
2     4.269231    1.342308
3     5.595833    2.037500

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [57] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
[113] 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1]  2.02200 13.05769 16.29167
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
> table(irisCluster$cluster, iris$Species)

    setosa versicolor virginica
  1     50          0         0
  2      0         48         4
  3      0          2        46
```

Hierarchical Clustering:

★ Initially, each object is assigned to its own cluster.

★ at each stage joining the two most similar clusters, continuing until there is just a single cluster.

★ at each stage distances between clusters are recomputed by the LanceWilliams dissimilarity update formula.

```
> clusters <- hclust(dist(iris[, 3:4]))
> plot(clusters,col="blue")
```

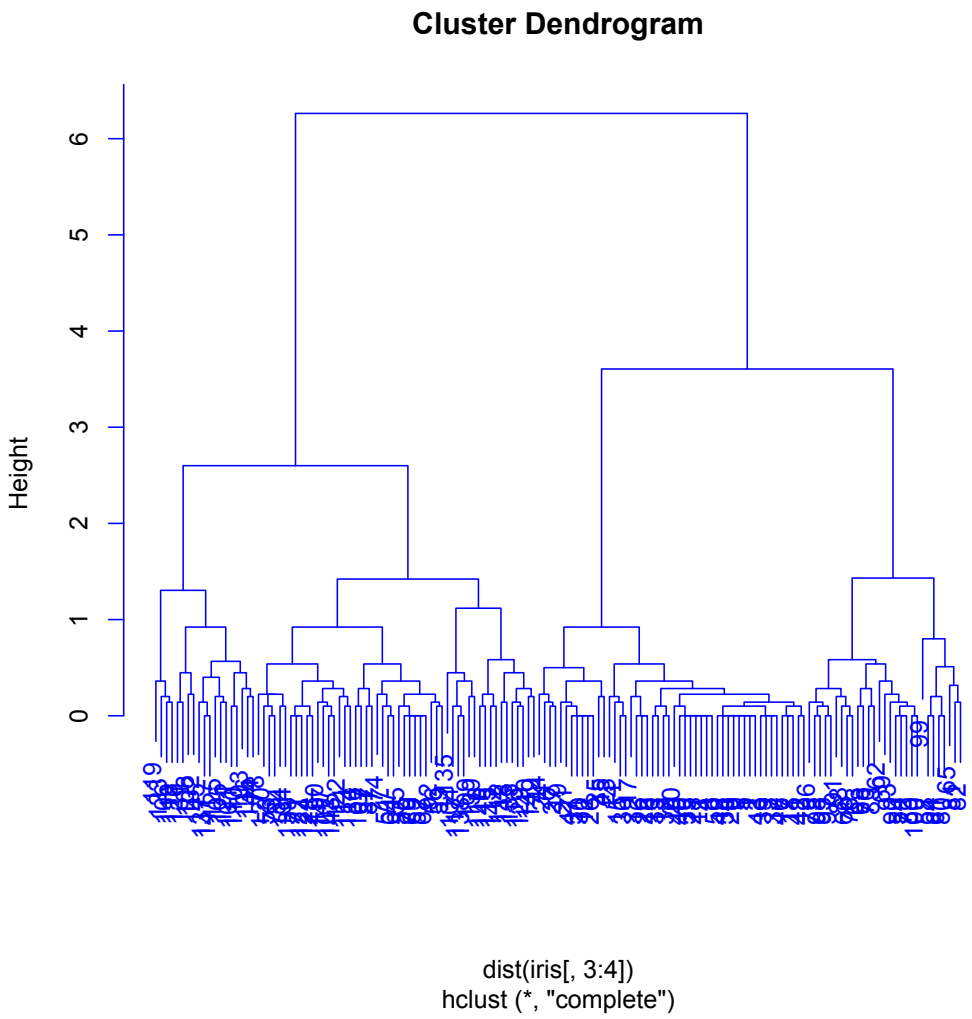


Figure 12.6: Hierarchical clustering by using iris data.