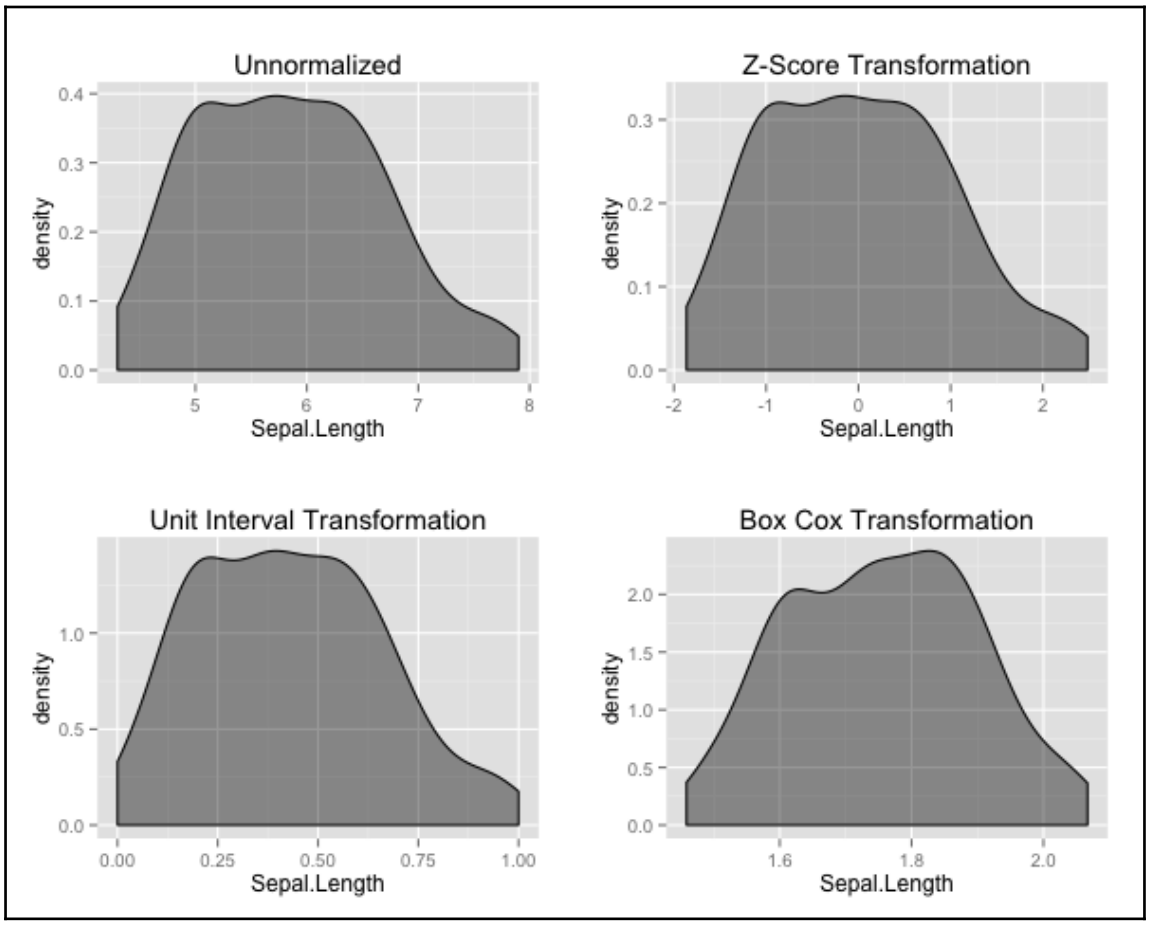
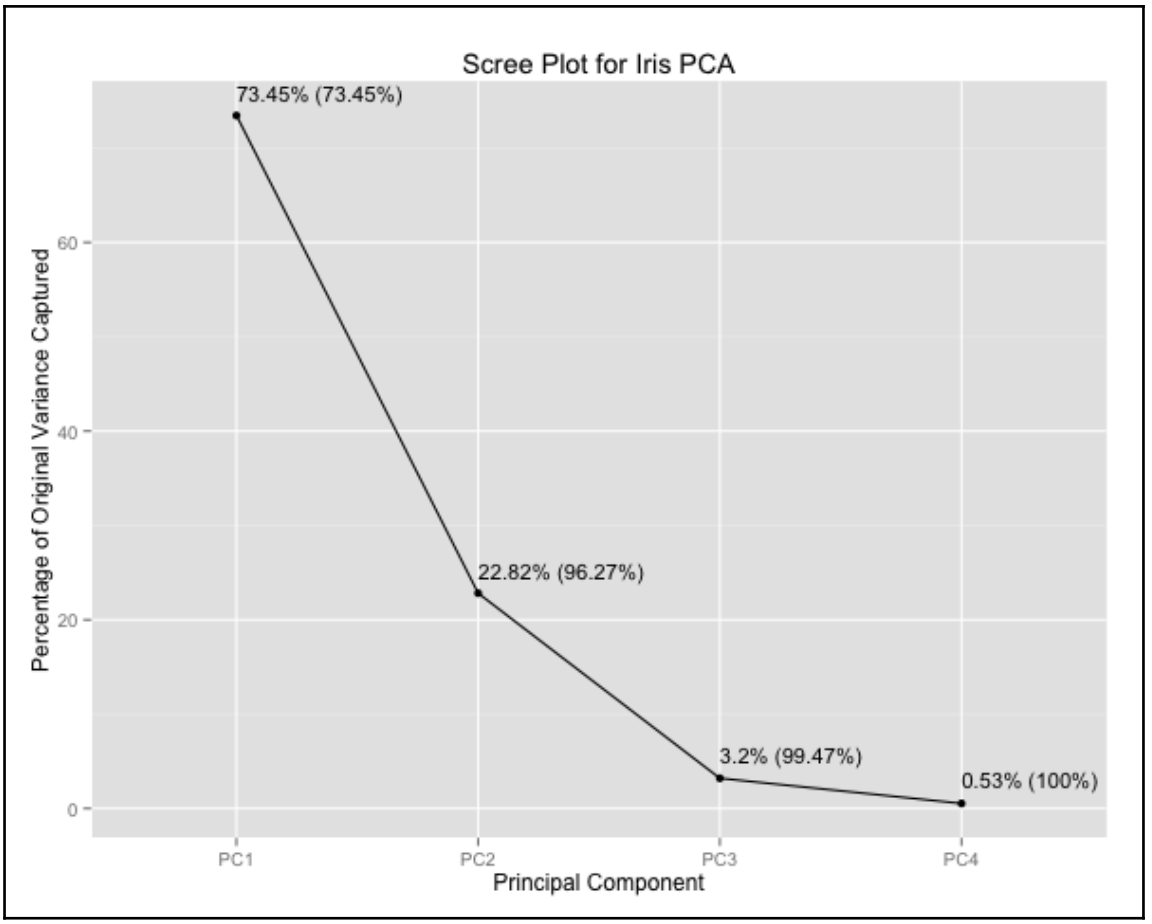
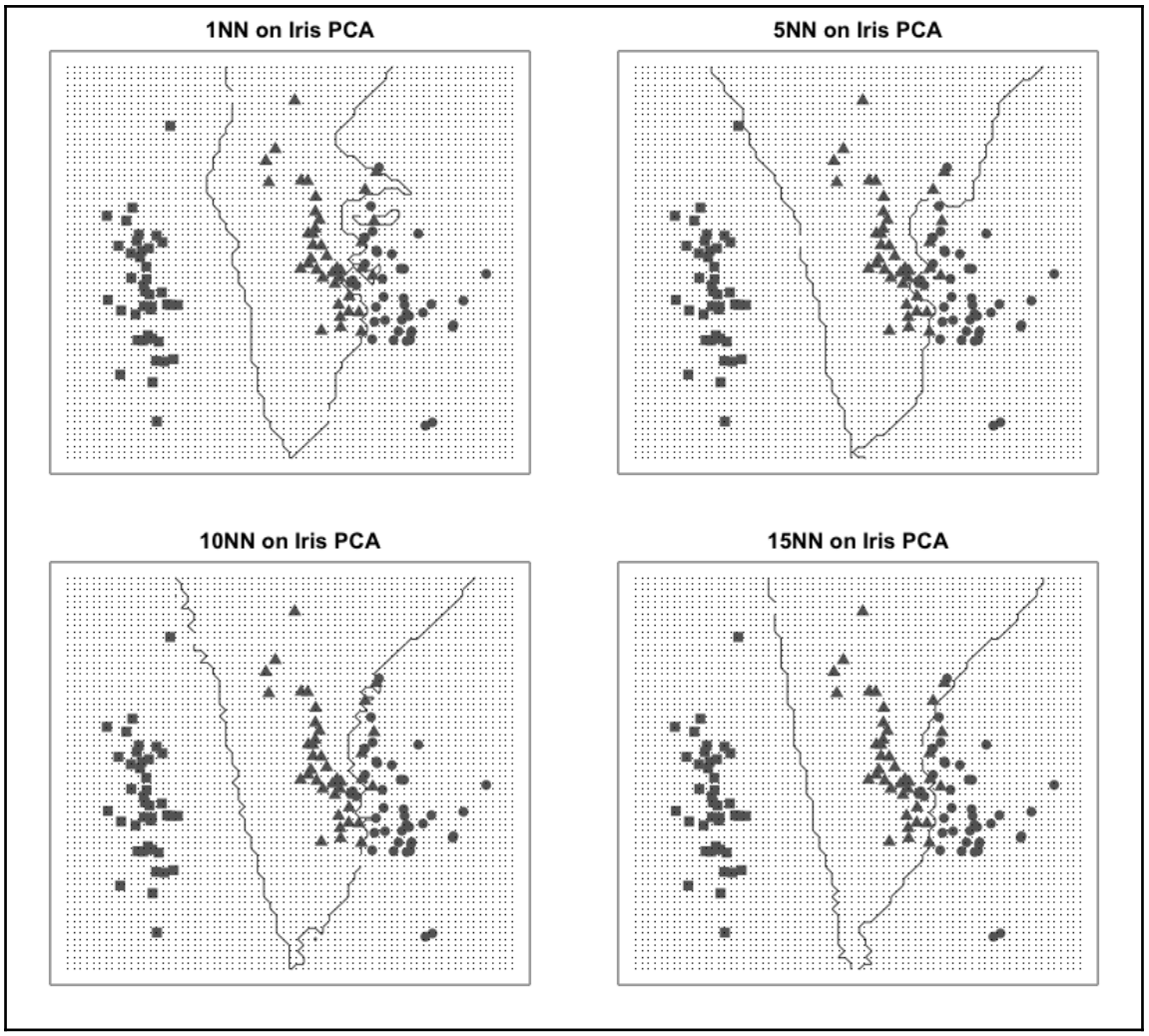
Chapter 1: Gearing Up for Predictive Modeling




Scree Plot for Iris PCA
73.45% (73.45%)
22.82% (96.27%)
3.2% (99.47%)
0.53% (100%)
Percentage of Original Variance Captured
60
40
20
0
PC1
PC2
PC3
PC4
Principal Component


1NN on Iris PCA
5NN on Iris PCA
10NN on Iris PCA
15NN on Iris PCA

Chapter 2: Tidying Data and Measuring Performance

```
> a<-69
> print(a)
[1] 69
> b<-as.character(a)
> print(b)
[1] "69"
>
```

```
> difftime(as.Date(saledate), as.Date(returndate))
Time difference of -31 days
>
```

$$MSE = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \hat{y}_i = \hat{f}\left(x_i\right)$$

$$ER = \frac{1}{n}\sum_{1}^{n} I\left(y_i \neq \hat{y}_i\right)$$

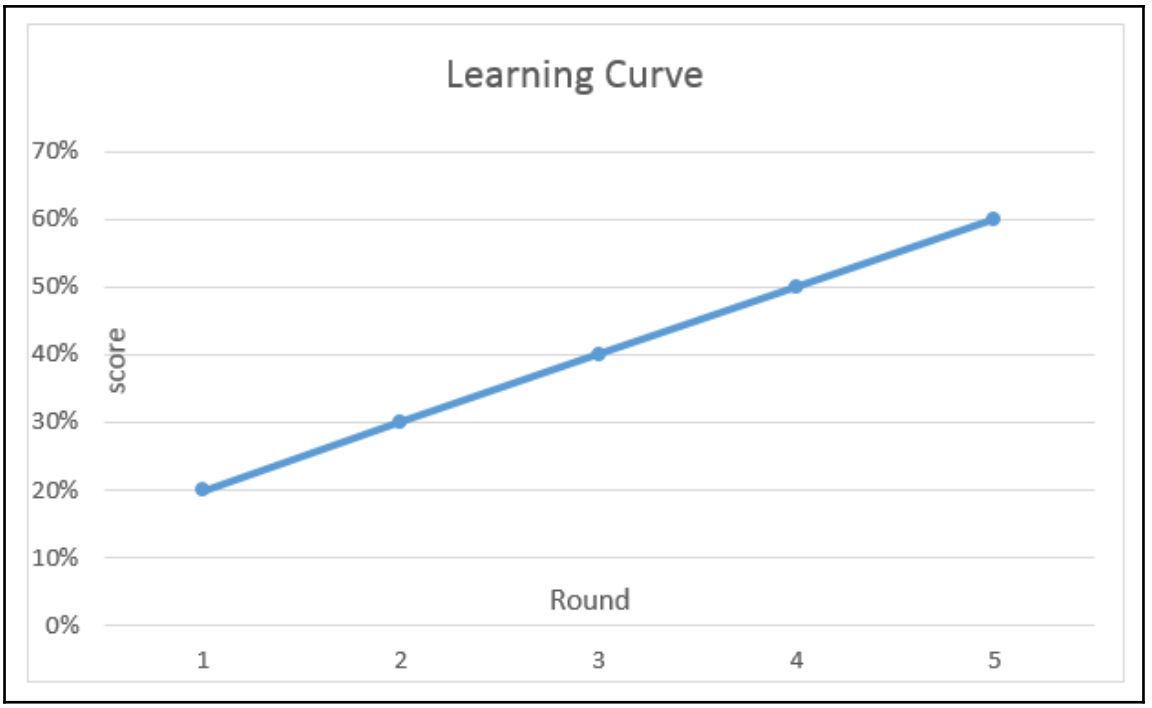
$$Kappa = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

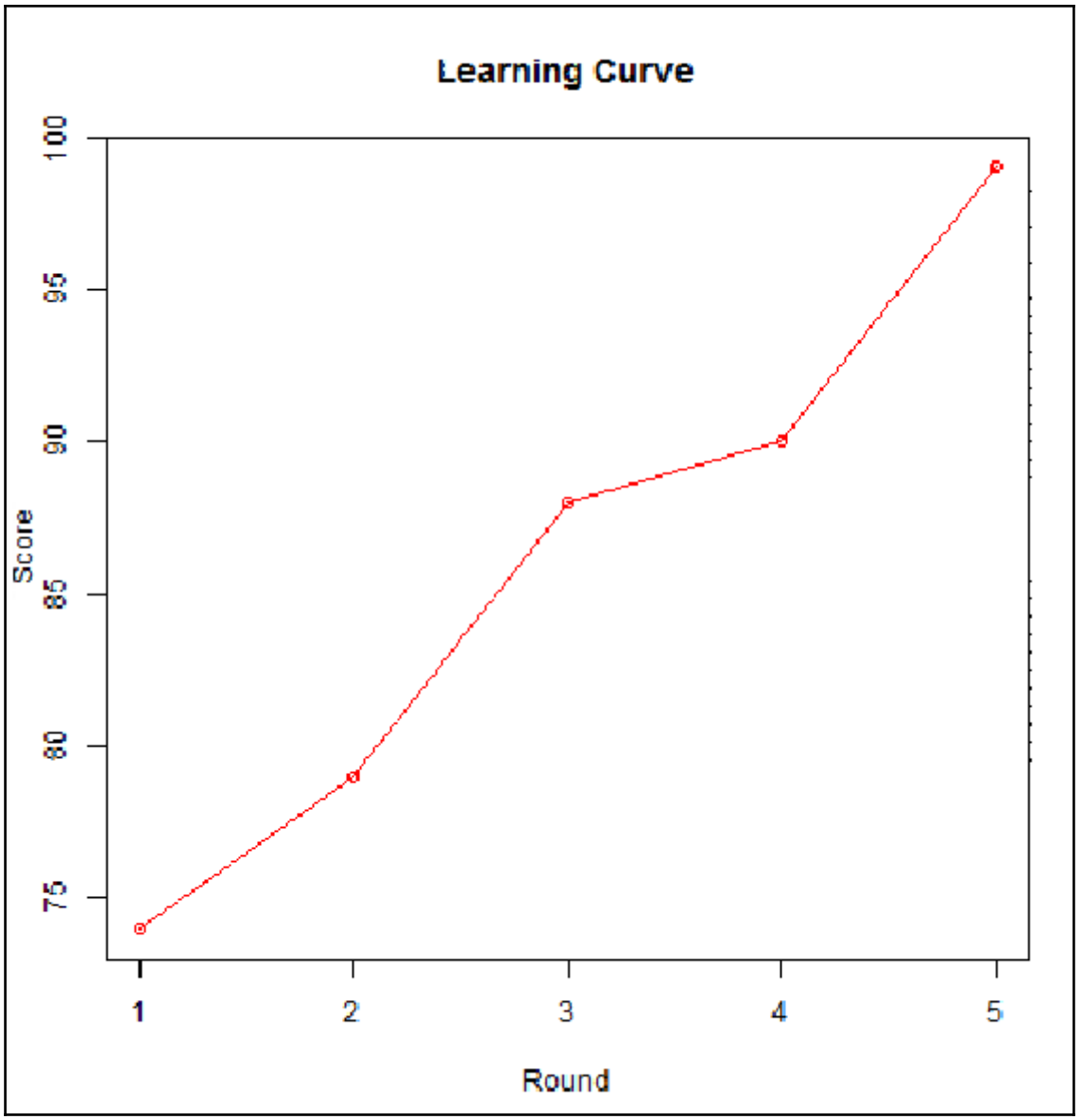
$$Precision = \frac{TruePositives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{TruePositives}{True\ Positives + False\ Negatives}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Characteristic	Validation Percent	Round 1 Percent	Round 2 Percent	Round 3 Percent	Round 4 Percent	Round 5 Percent	Average
GPA>3.0	90%	90%	80%	89%	72%	90%	84%
Alum	37%	36%	37%	30%	35%	37%	35%
Active	79%	77%	78%	79%	79%	79%	78%
Resident	92%	95%	90%	91%	92%	92%	92%
Athlete	69%	69%	69%	69%	69%	61%	67%
InState	90%	90%	90%	90%	90%	90%	90%
International Study	75%	74%	75%	76%	77%	78%	76%
Transfer	5%	5%	6%	4%	5%	5%	5%


Learning Curve
70%
60%
50%
40%
30%
20%
10%
0%
score
Round
1
2
3
4
5


Learning Curve
Score
100
95
90
85
80
75
1
2
3
4
5
Round

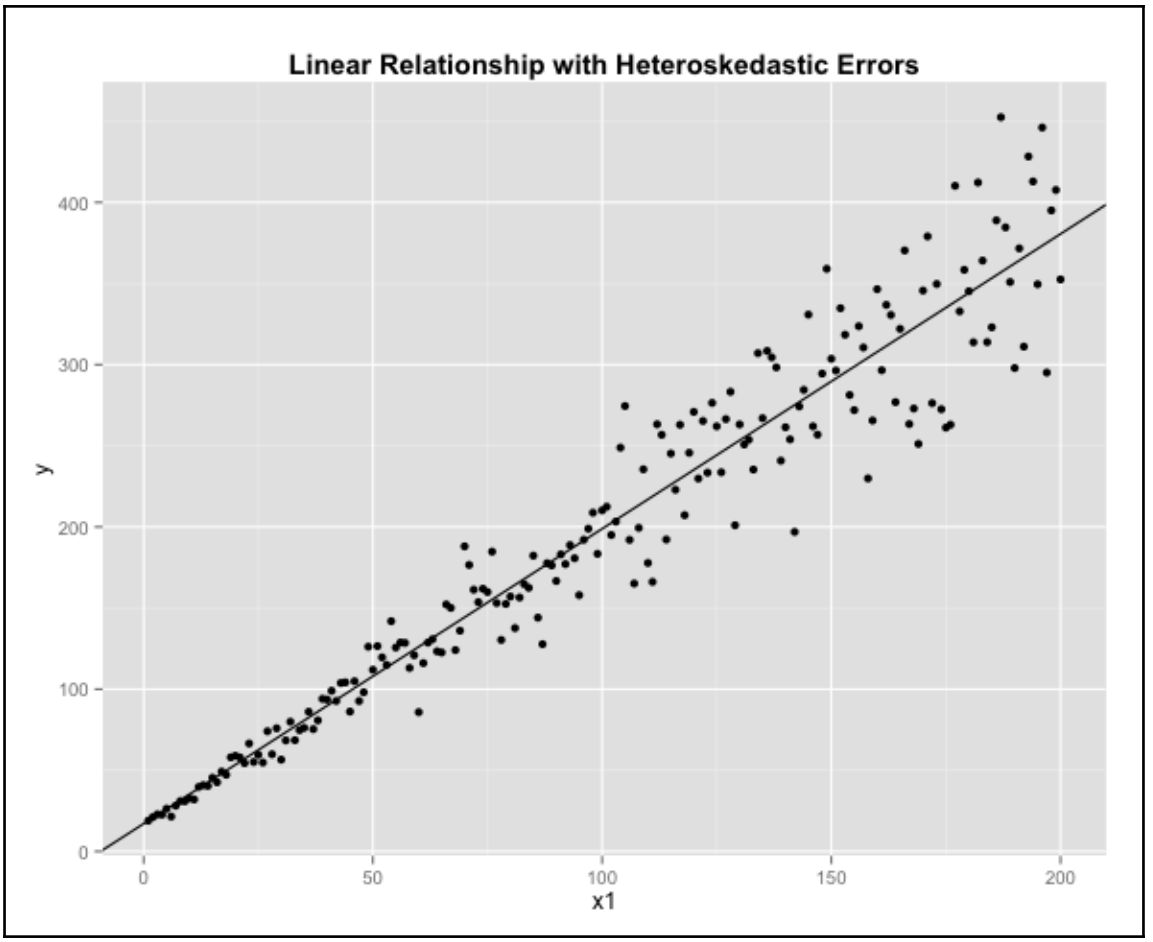
Chapter 3: Linear Regression

$$\hat{y} = \beta_1 x + \beta_0$$

$$\hat{y} = 1.91x_1 + 2.56x_2 - 7.56x_3 + 0.49$$

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

$$y = \beta_2 x_2 + \beta_1 x_1 + \beta_0 + \varepsilon$$



$$y = 0.8\sin\left(z_1\right) - 0.6\ln\left(z_2\right) + 2.3e^{z_3}$$

$$x_1 = \sin(z_1)$$

$$x_2 = \ln(z_2)$$

$$x_3 = e^{z_3}$$

$$y = 0.8x_1 - 0.6x_2 + 2.3x_3$$

$$y_1 = x_1^{\beta_1} + \beta_0$$

$$y_2 = \beta_2 \sin(\beta_1 x_1) + \beta_0$$

$$y = 1.67x_1 - 2.93 + N(0, 2^2)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

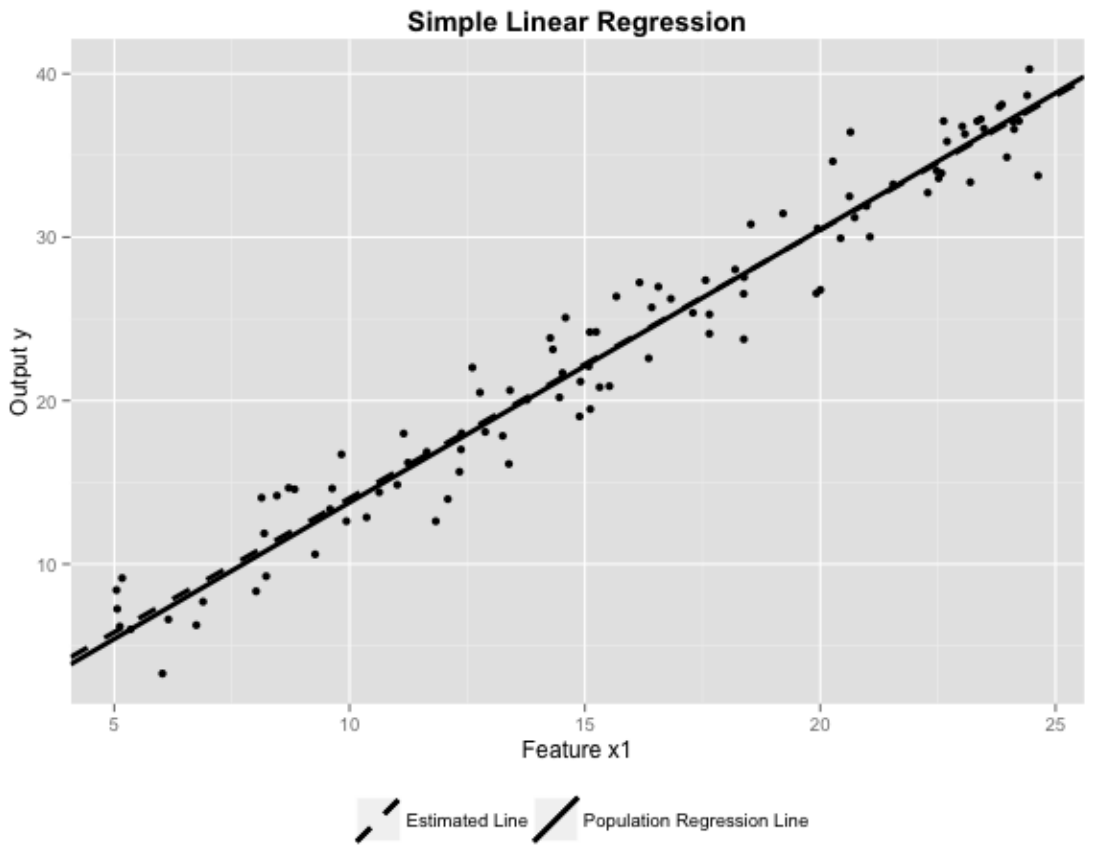
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

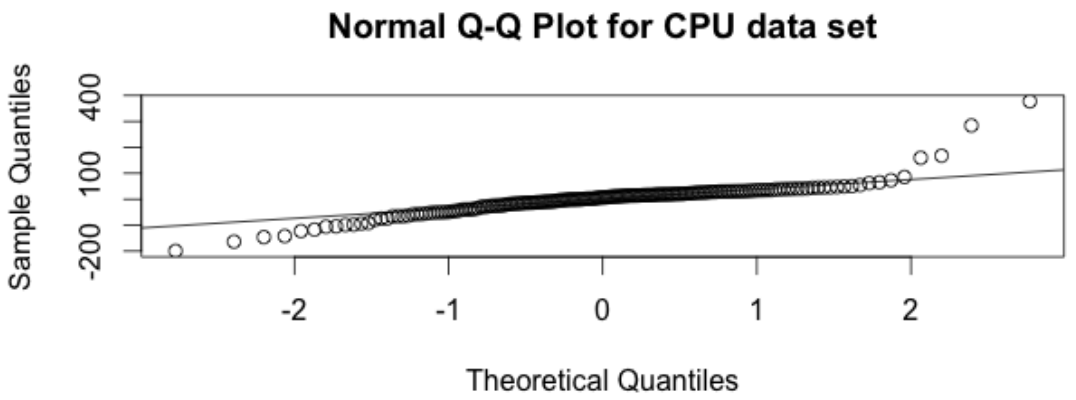
$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

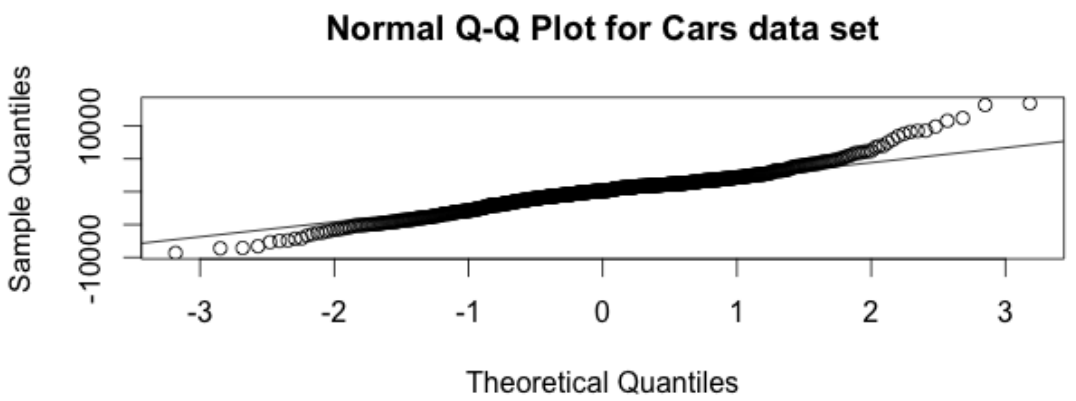
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

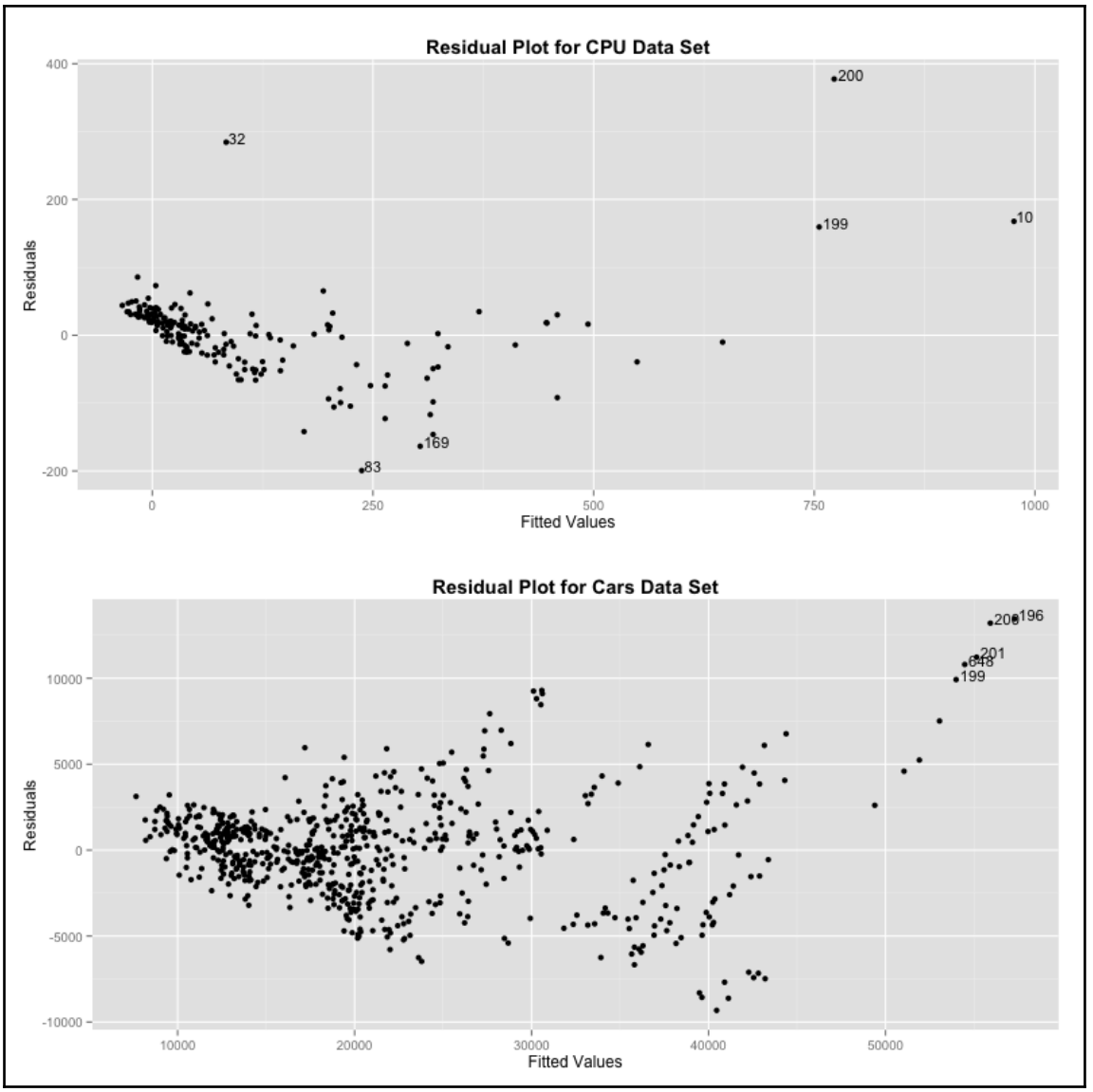


$$y = \beta_k x_k + \beta_{k-1} x_{k-1} + \cdots + \beta_1 x_1 + \beta_0 + \varepsilon$$

$$e_i = y_i - \hat{y}_i$$






Residual Plot for CPU Data Set
Residuals
Fitted Values
400
200
0
-200
0
250
500
750
1000
200
32
199
10
169
83
Residual Plot for Cars Data Set
Residuals
Fitted Values
10000
5000
0
-5000
-10000
10000
20000
30000
40000
50000
196
201
199

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$MSE = \frac{1}{n} \cdot RSS$$

$$RSE = \sqrt{\frac{RSS}{n-k-1}}$$

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

$$TSS(y) = n \cdot Var(y) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot \sqrt{Var(Y)}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

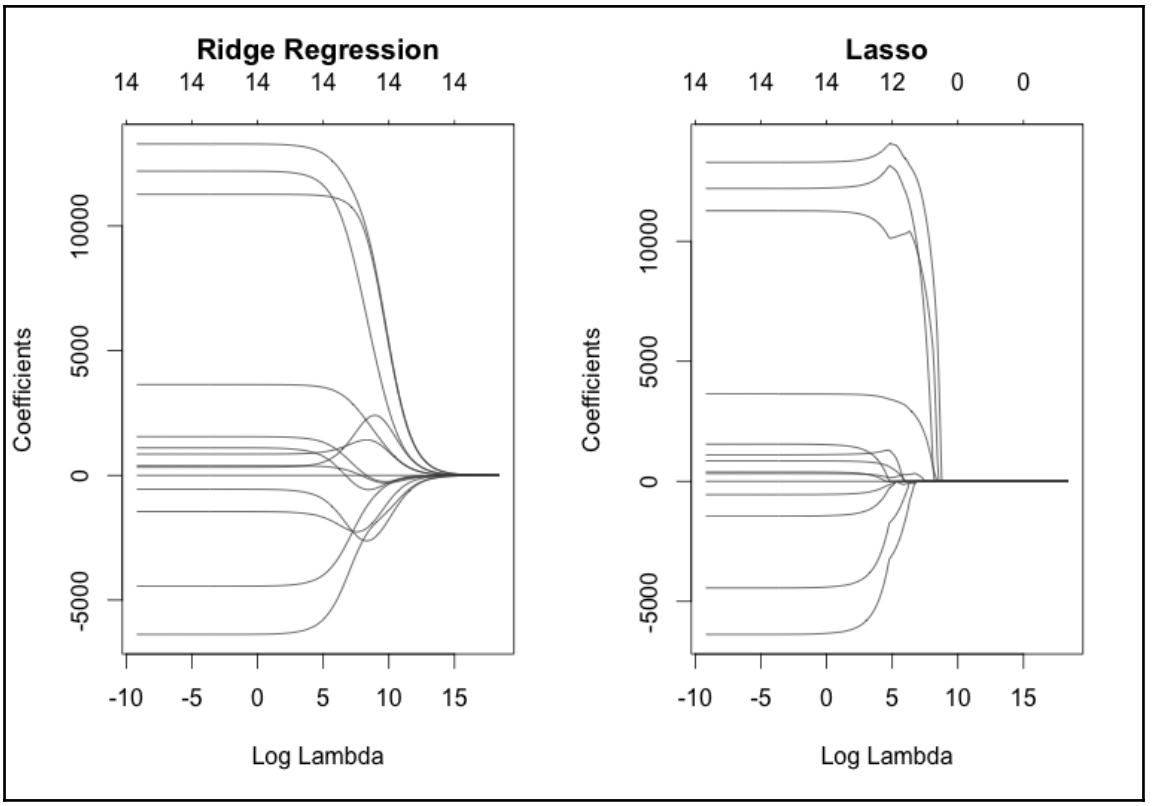
$$R^2_{adjusted} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

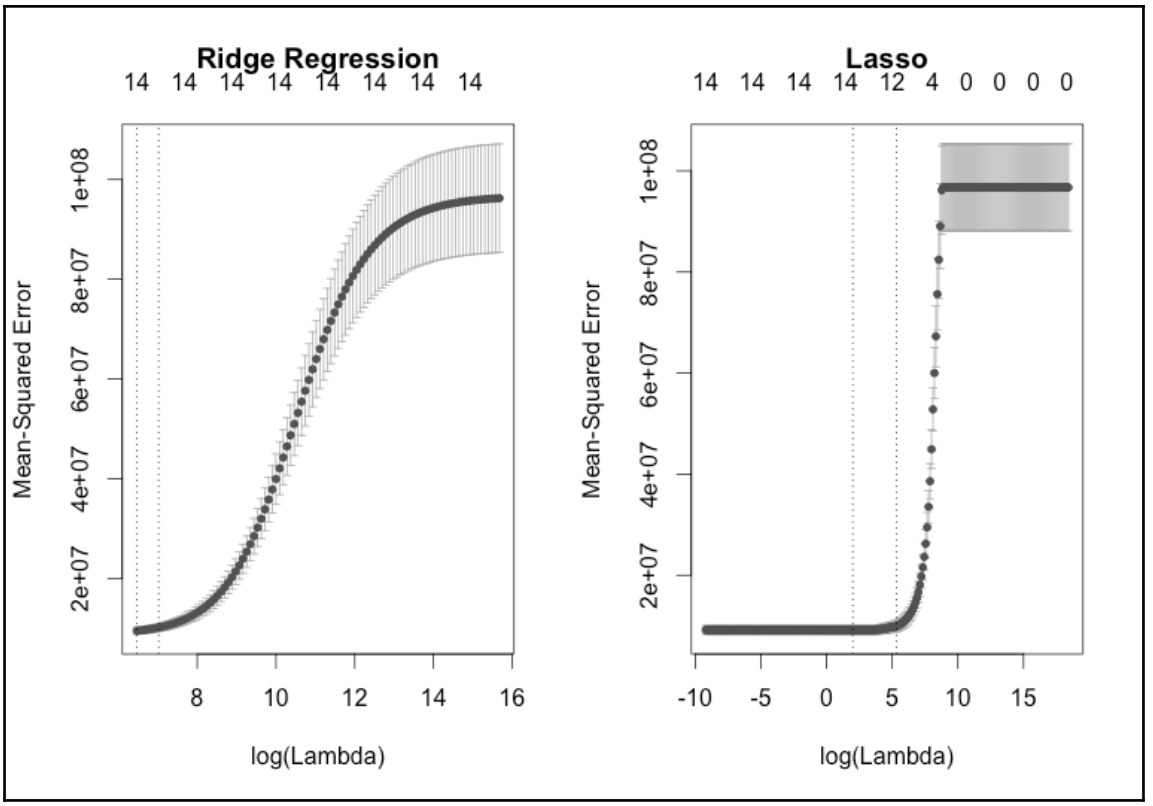
$$1 + (k + (k-1) + \cdots + 1) = 1 + \sum_{i=1}^{k} i = 1 + \frac{k(k+1)}{2}$$

$$RSS + \lambda \sum_{j=1}^{k} \beta_j^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

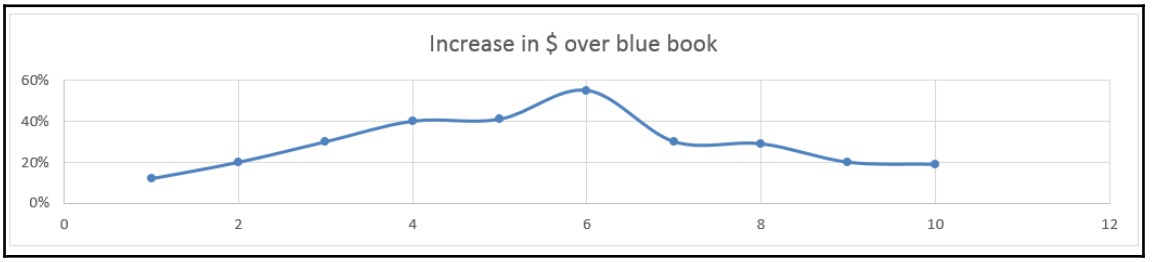
$$RSS + \lambda \sum_{j=1}^{k} \left| \beta_j \right| = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{k} \left| \beta_j \right|$$

$$\sqrt[p]{\sum_{i=1}^{n} \left| v_i \right|^{\mathrm{p}}}$$

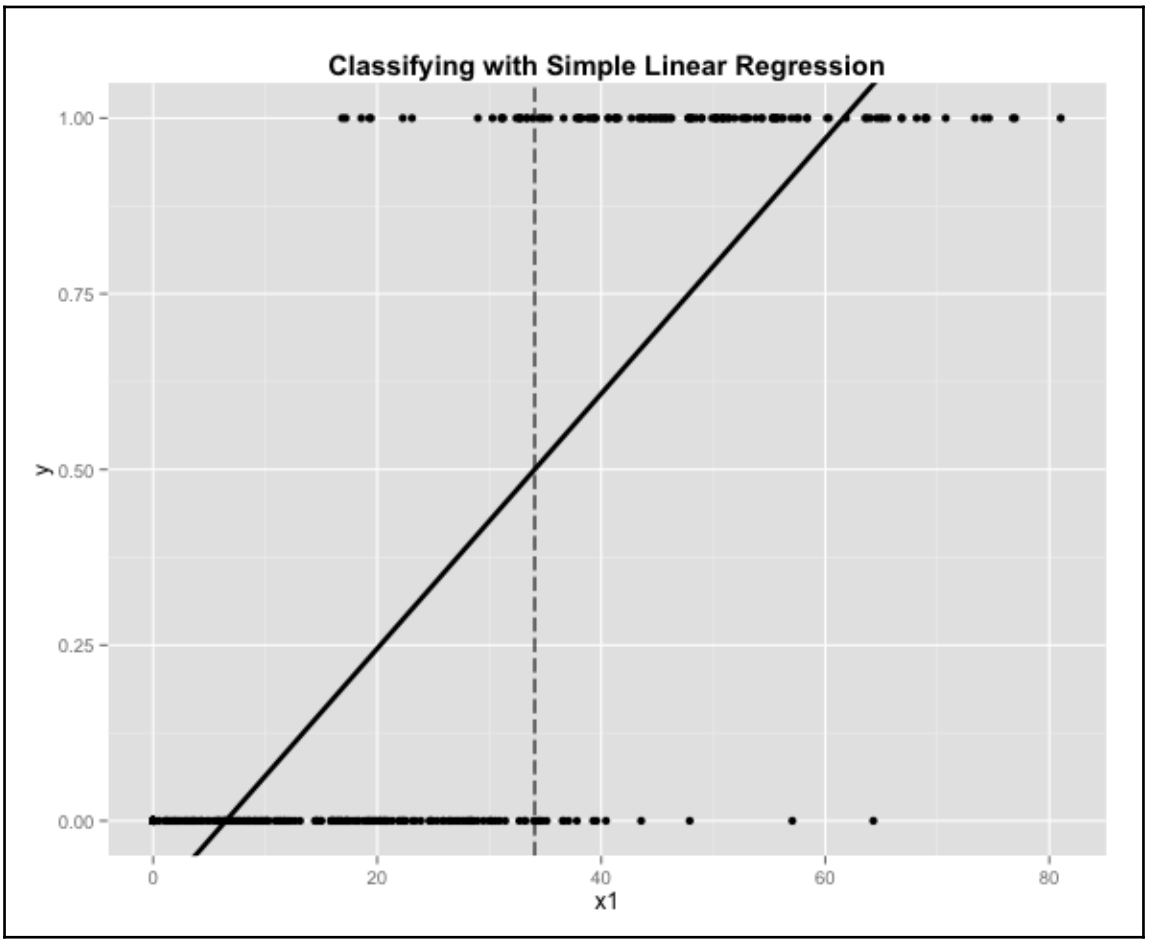

Ridge Regression
14 14 14 14 14 14
Coefficients
10000
5000
0
-5000
-10 -5 0 5 10 15
Log Lambda
Lasso
14 14 14 12 0 0
Coefficients
10000
5000
0
-5000
-10 -5 0 5 10 15
Log Lambda

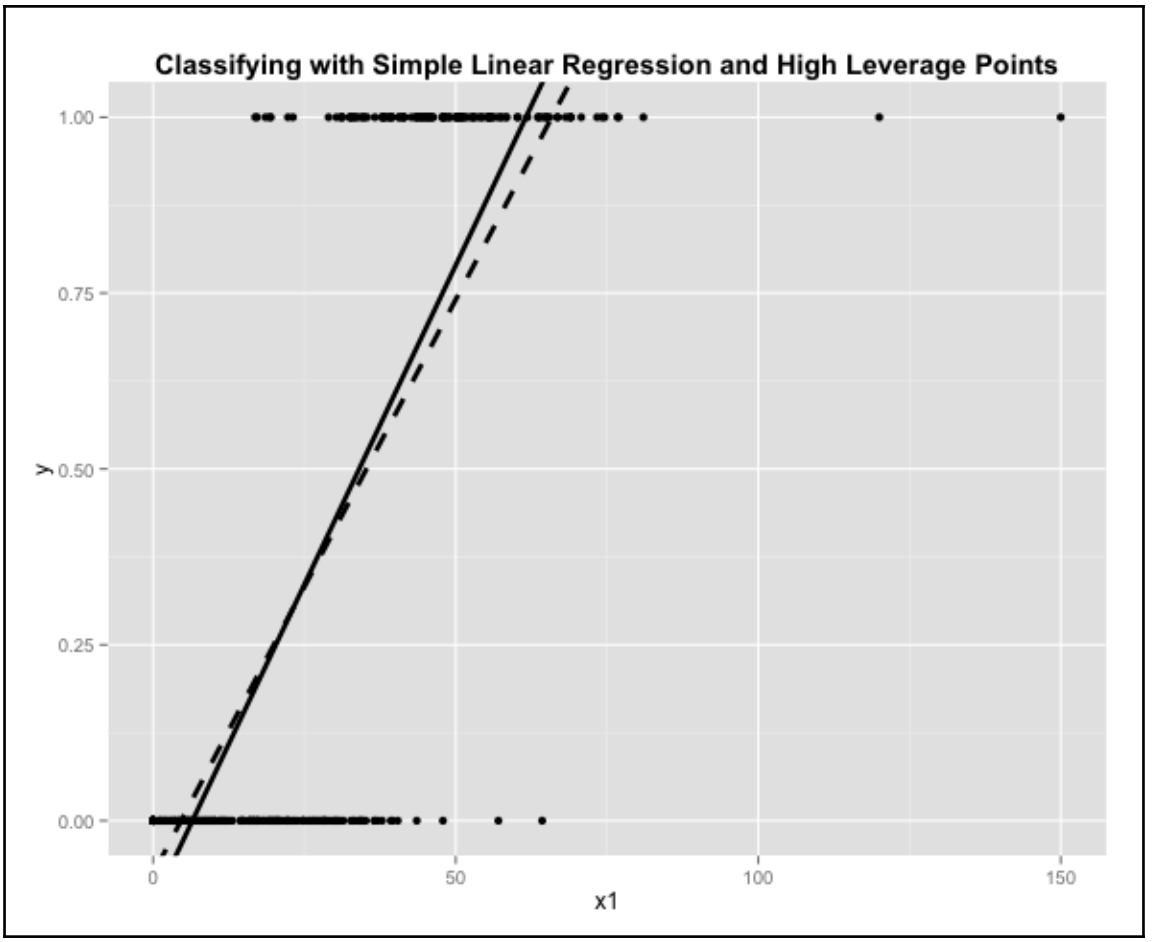


Vehicle Options	1	2	3	4	5	6	7	8	9	10
Increase in $ over blue book	12%	20%	30%	40%	41%	55%	30%	29%	20%	19%

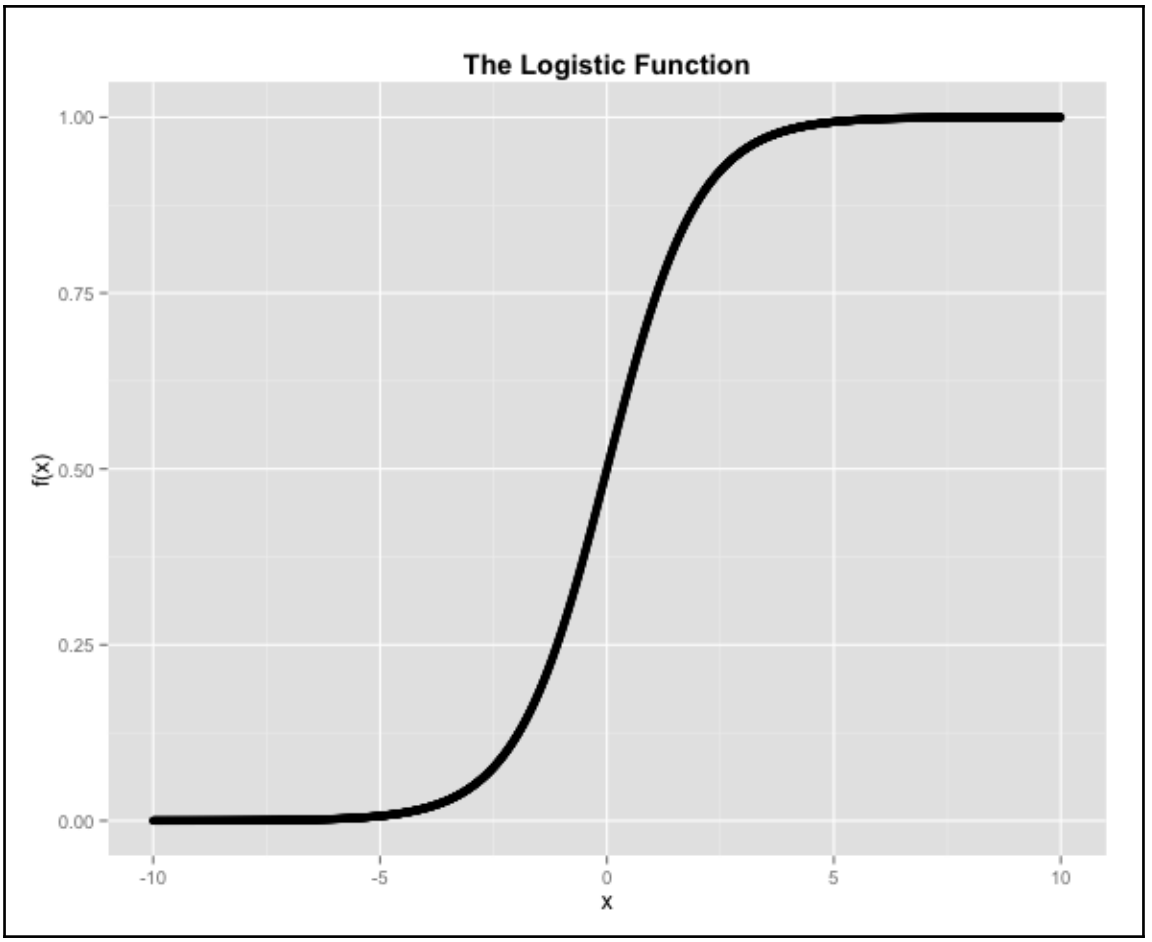


Chapter 4: Generalized Linear Models





$$f(x) = \frac{e^x}{e^x + 1} = \frac{e^{-x}}{e^{-x}} \cdot \frac{e^x}{\left(e^x + 1\right)} = \frac{1}{1 + e^{-x}}$$



$$x = \beta_0 + \beta_1 X_1$$

$$P(Y=1|X)=\frac{e^{\beta_0+\beta_1 X_1}}{e^{\beta_0+\beta_1 X_1}+1}$$

$$\mu_y = P(Y=1|X)=\frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}+1}$$

$$P(Y=1|X)\cdot\left(e^{\beta_0+\beta_1 X}+1\right)=e^{\beta_0+\beta_1 X}$$

$$P(Y=1|X)\cdot e^{\beta_0+\beta_1 X}+P(Y=1|X)=e^{\beta_0+\beta_1 X}$$

$$P(Y=1|X)=e^{\beta_0+\beta_1 X}-P(Y=1|X)\cdot e^{\beta_0+\beta_1 X}$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)}=e^{\beta_0+\beta_1 X}$$

$$ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right)=\beta_0+\beta_1 X$$

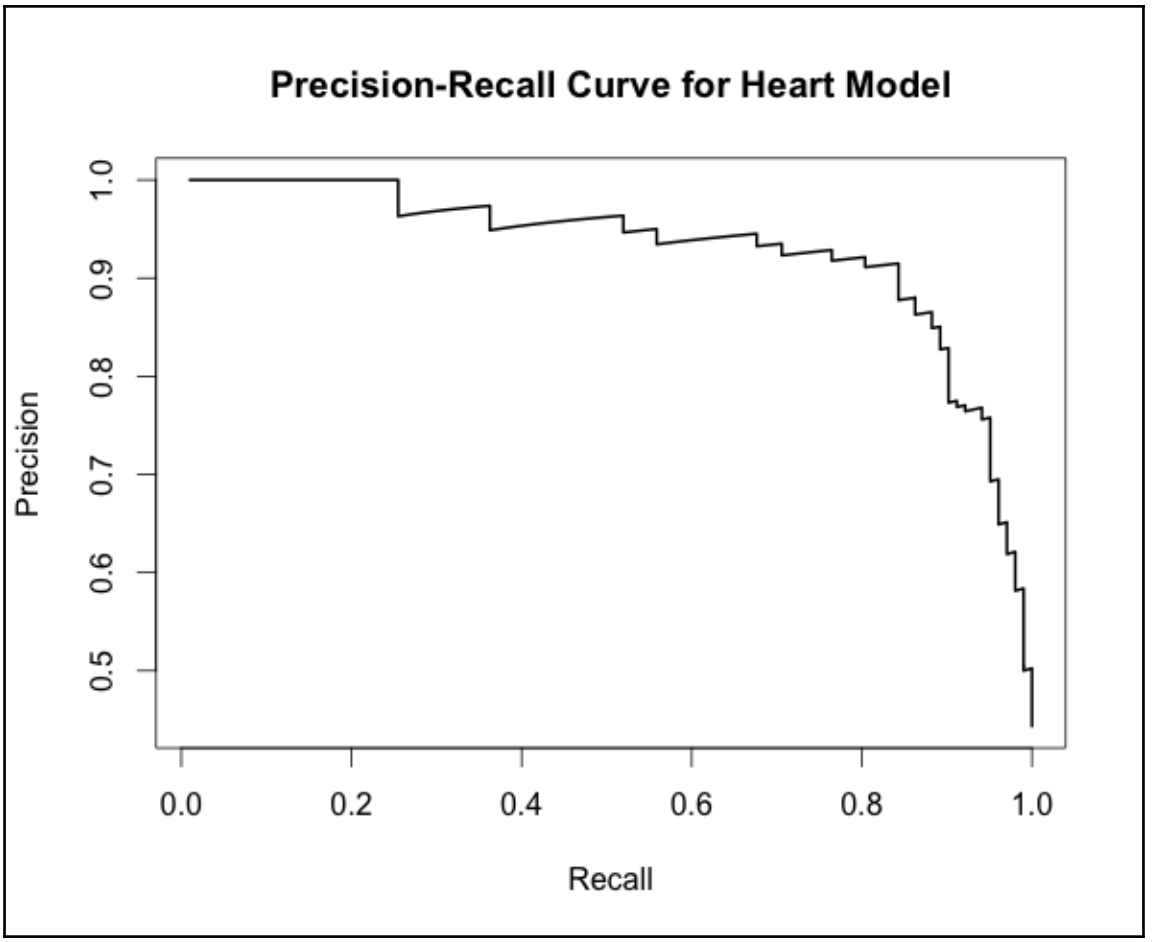
$$e^{\beta_i}$$

$$l\left(\beta_0,\beta_1\cdots\beta_p\right)=\prod_{i:y_i=1}P\left(y_i=1\,|\,x_i\right)\cdot\prod_{j:y_j=0}1-P\left(y_j=1\,|\,x_j\right)$$

$$ll\left(\beta_0,\beta_1\cdots\beta_p\right)=\sum_{i:y_i=1}log\left(P\left(y_i=1\,|\,x_j\right)\right)+\sum_{i:y_i=0}log\left(1-P\left(y_j=1\,|\,x_j\right)\right)$$

$$ll\left(\beta_0,\beta_1\cdots\beta_p\right)=\sum_{i}y_i\cdot log\left(P\left(y_i=1\,|\,x_j\right)\right)+\left(1-y_i\right)\cdot log\left(1-P\left(y_i=1\,|\,x_j\right)\right)$$

$$dr_i=d_i\cdot sign\left(\hat{y}_i-P\left(y_i=1\,|\,x_i\right)\right)$$


Precision-Recall Curve for Heart Model
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Recall

$$ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)=\beta_{10}+\beta_{11}X_1+\beta_{12}X_2$$

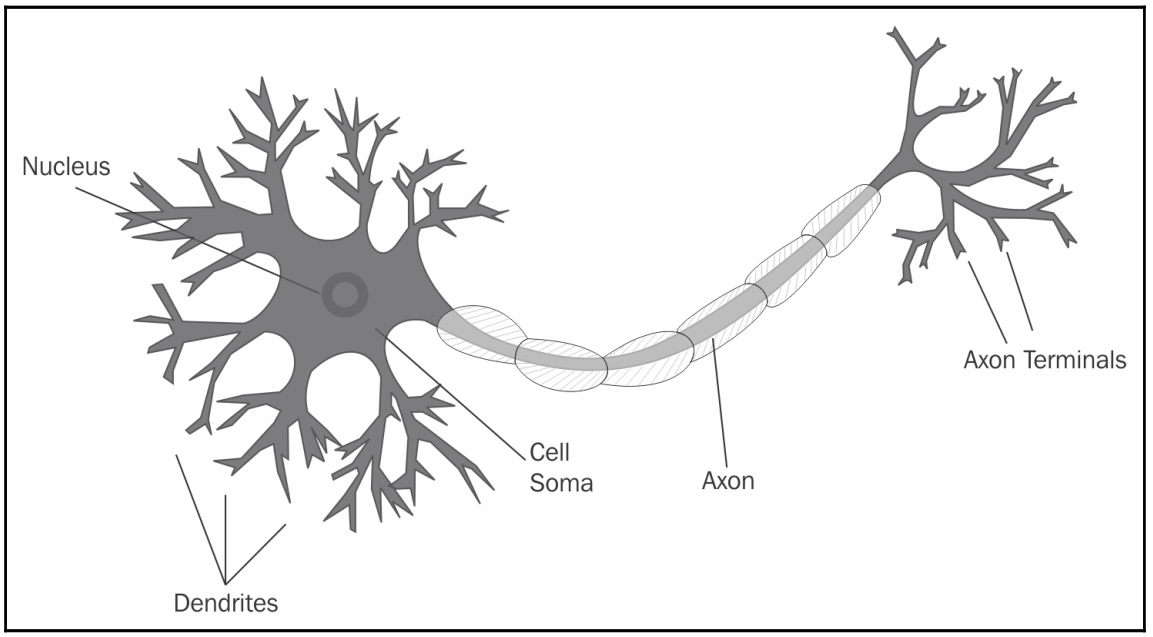
$$ln\left(\frac{P(Y=2|X)}{P(Y=0|X)}\right)=\beta_{20}+\beta_{21}X_1+\beta_{22}X_2$$

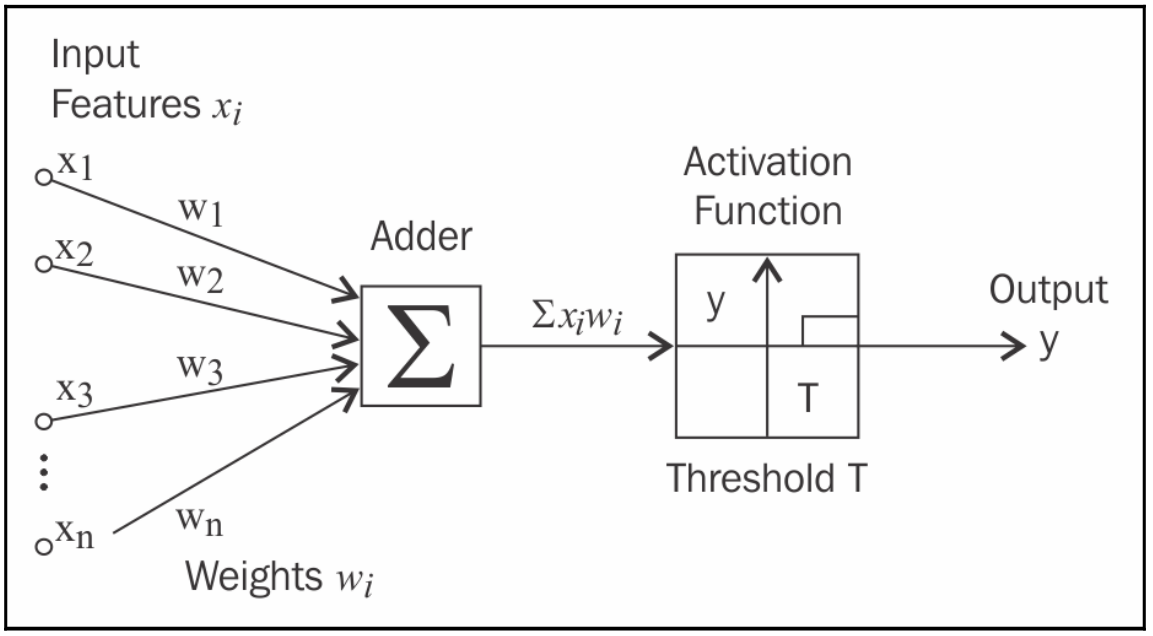
$$P(Y=k|X)=\begin{cases}\dfrac{1}{1+\sum_{j=1}^{K-1}e^{f_j(x)}}, & k=0\\ \dfrac{e^{f_j(x)}}{1+\sum_{j=1}^{K-1}e^{f_j(x)}}, & k>0\end{cases}$$

$$f_j(x)=\beta_{j0}+\beta_{j1}X_1+\cdots+\beta_{jp}X_p$$

$$\ln\left(\frac{P(Y\le k|X)}{P(Y>k|X)}\right)=\beta_k+\beta_1X_1+\cdots+\beta_pX_p,\ \forall k\in\{0,\ldots K-1\}$$

Chapter 5: Neural Networks





$$y = g\left(w_0 + \sum_{i=1}^{p} w_i x_i \right)$$

$$g(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

$$y = g\left(\sum_{i=0}^{p} w_i x_i \right)$$

$$\frac{1}{2n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2$$

$$\hat{y}_i$$

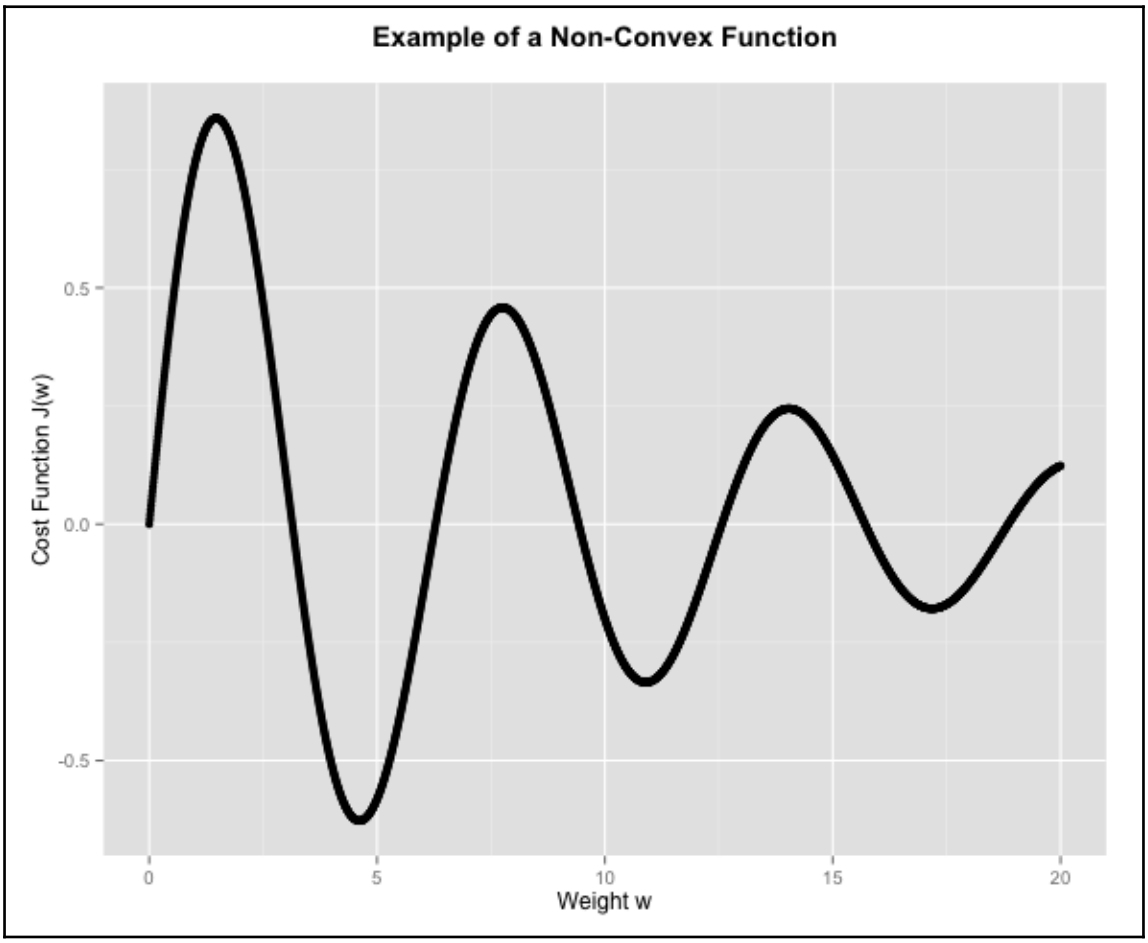
$$J\left(\vec{w}\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{p} w_j x_j\right) - y_i\right)^2$$

$$\vec{w}$$

$$\frac{\partial J\left(\vec{w}\right)}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{p} w_j x_j\right) - y_i\right) x_{ik}$$

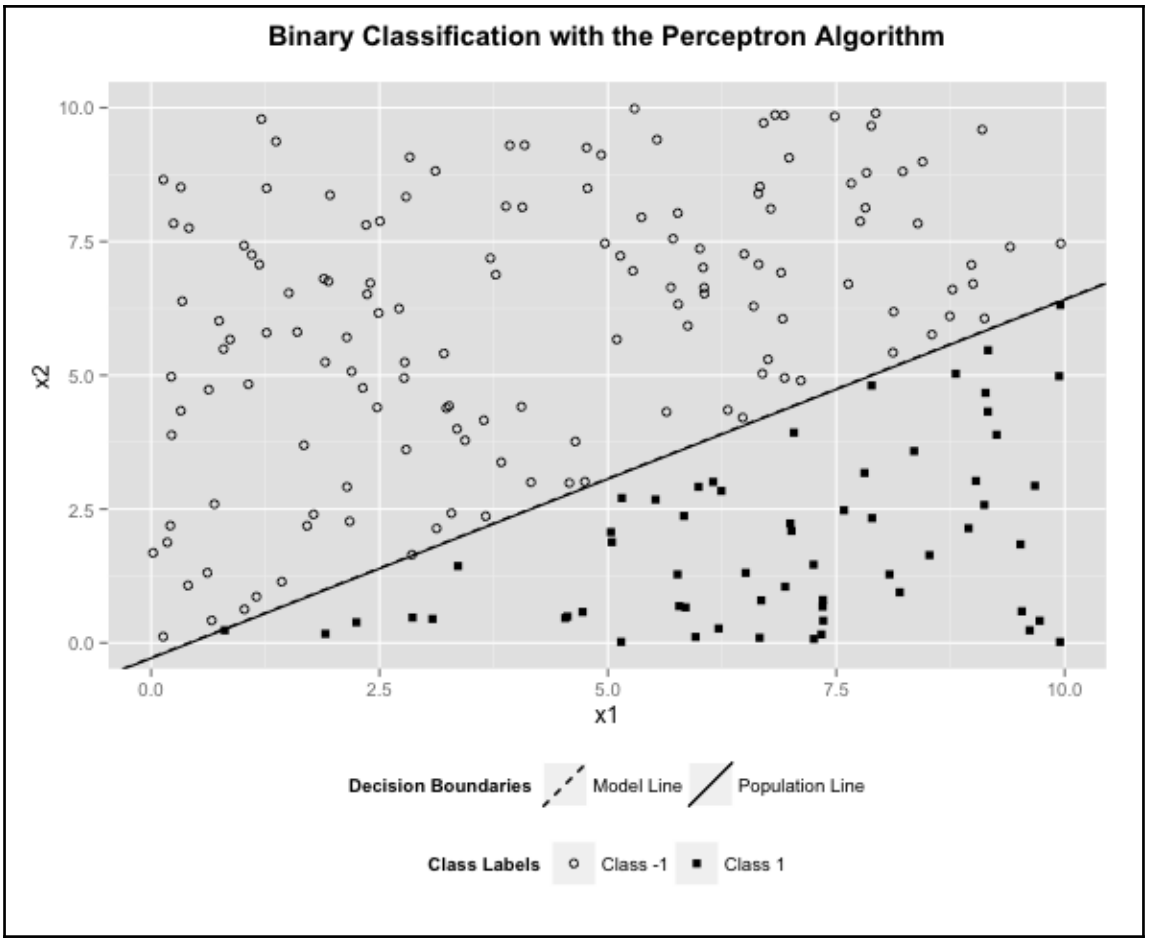
$$\frac{\partial J\left(\vec{w}\right)}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) x_{ik}$$

$$w_k \leftarrow w_k - \eta \frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right) x_{ik}$$



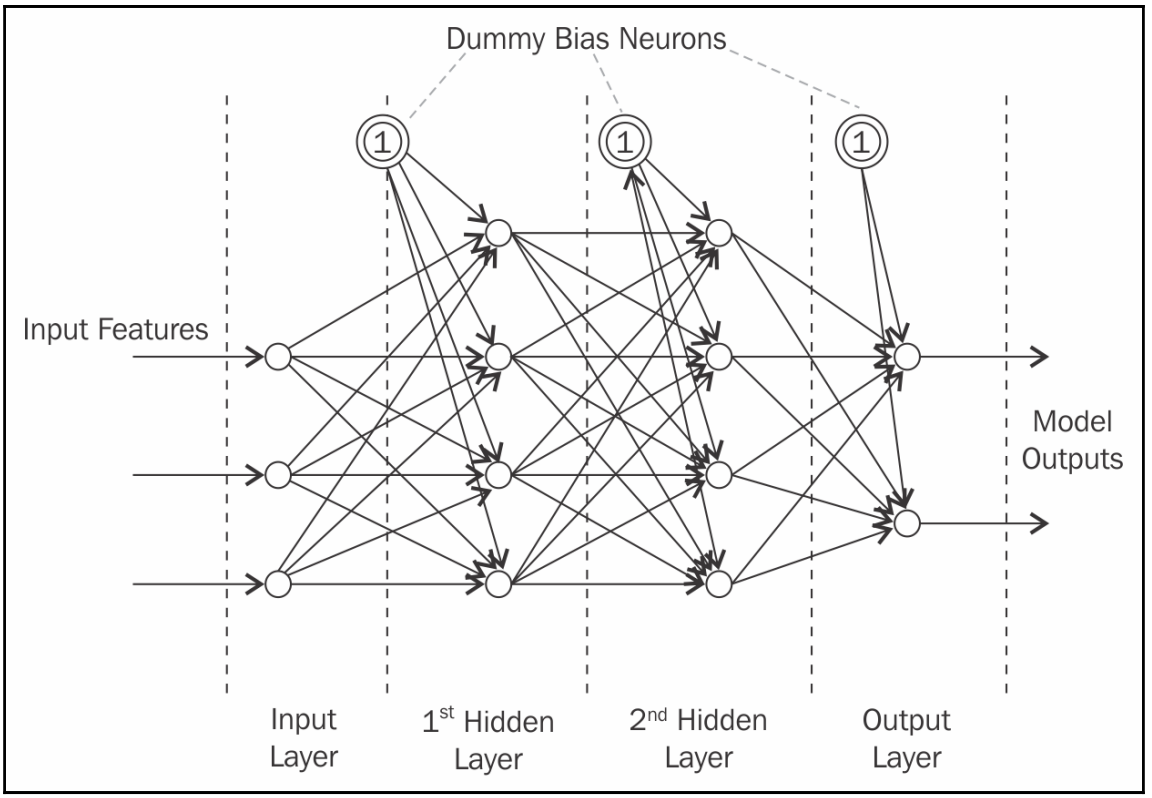
$$\hat{y}_i$$

$$y = \begin{cases} -1, & -0.89 + 2.07x_1 - 3.09x_2 < 0 \\ 1, & otherwise \end{cases}$$



$$w_k \leftarrow w_k - \eta \frac{\partial J\left(\vec{w}\right)}{\partial w_k}$$

$$w_k \leftarrow w_k - \eta\left(\hat{y}_i - y_i\right)x_{ik}$$

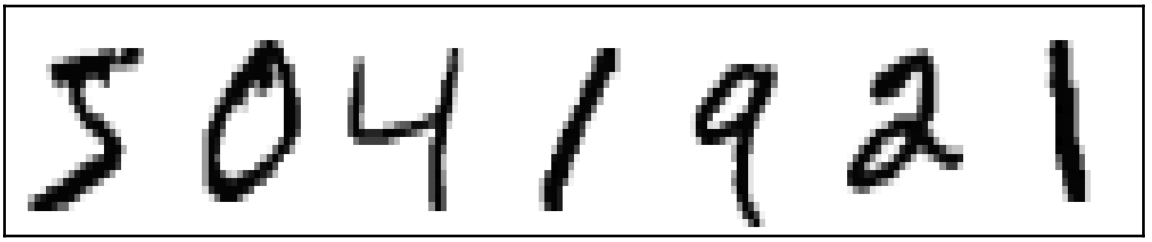


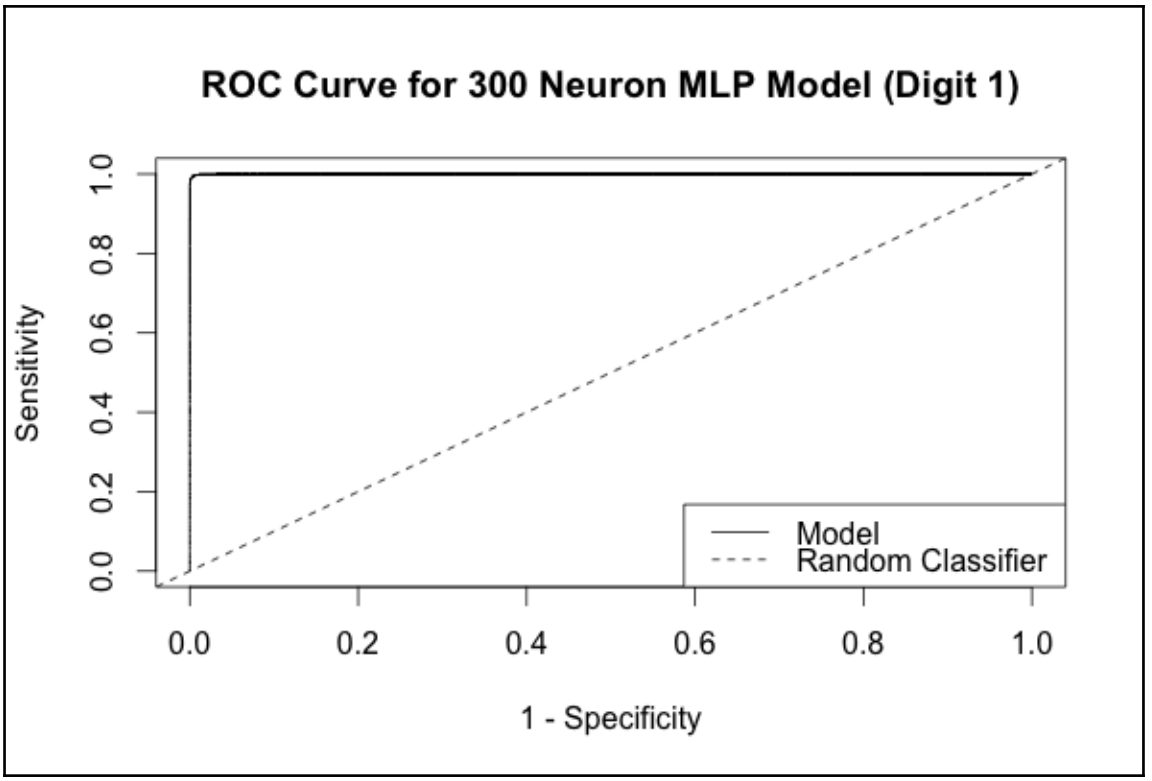
$$w_{ji}^{(n)} \leftarrow w_{ji}^{(n)} + \eta \cdot \delta_j^{(n)} \cdot y_i^{(n)}$$

$$y_j = g\left(z_j\right)$$

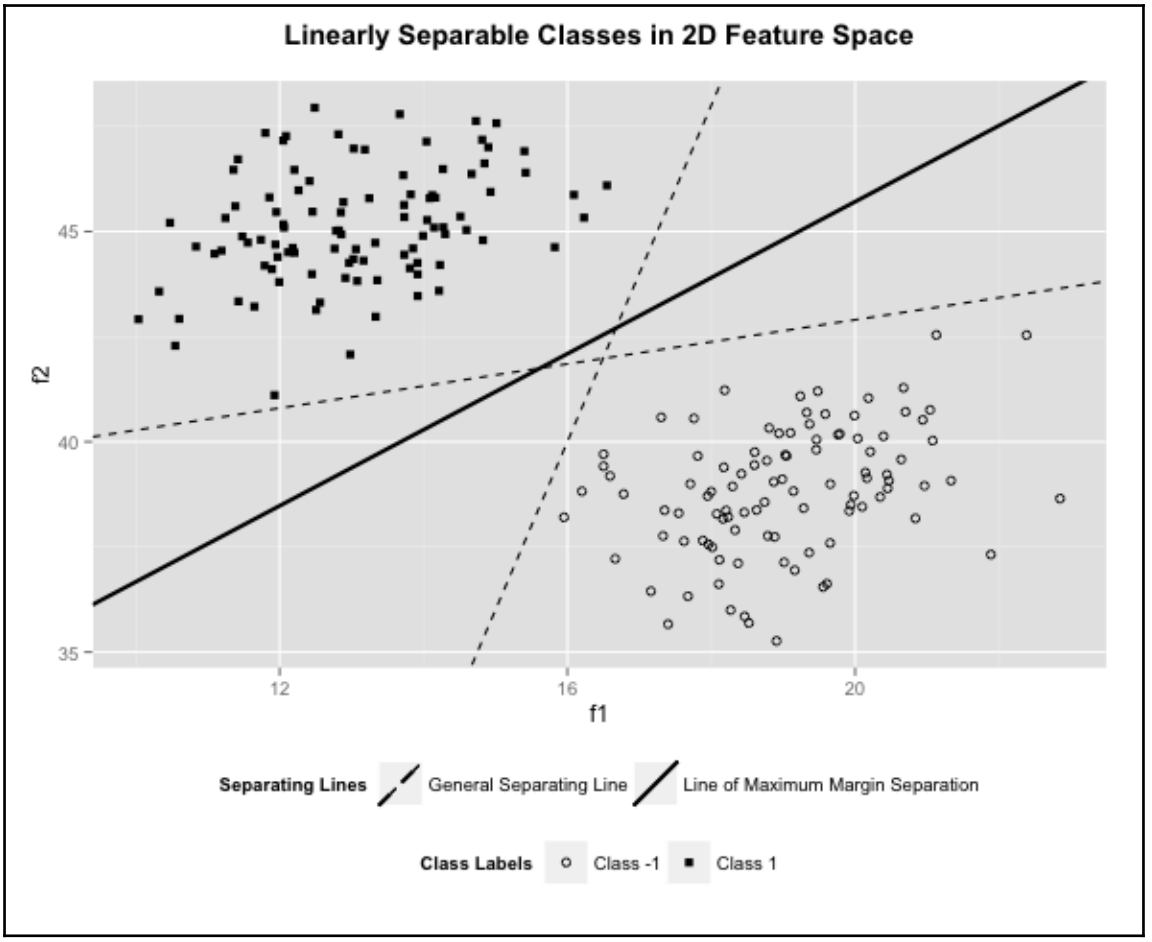
$$\delta_j = \left(t_j - y_j\right) \cdot y_j \cdot \left(1 - y_j\right)$$

$$\delta_j = \left(\sum_k \delta_k \cdot w_{kj}\right) \cdot y_j \cdot \left(1 - y_j\right)$$




ROC Curve for 300 Neuron MLP Model (Digit 1)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Model
Random Classifier

Chapter 6: Support Vector Machines

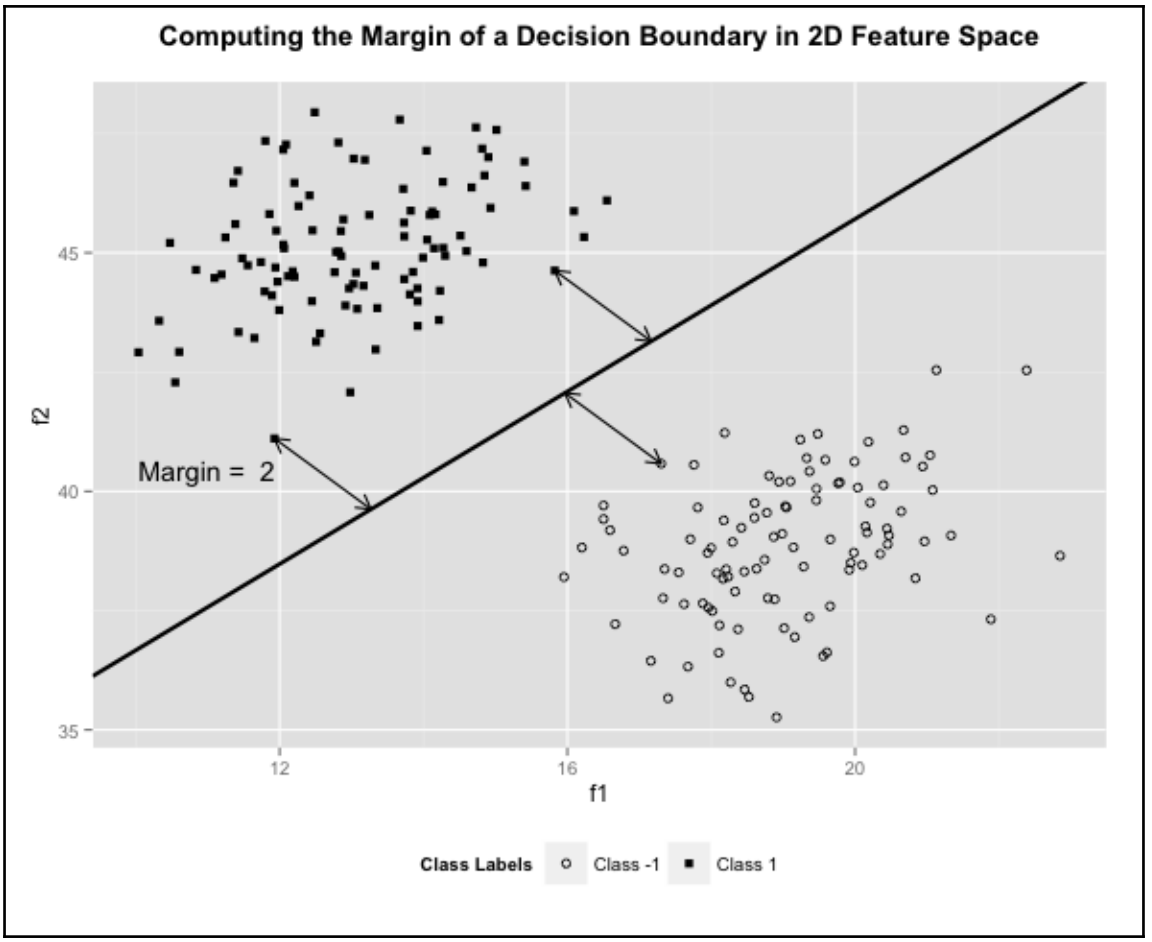


$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$$

$$\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} > 0 \ \ if \ \ y_i = 1$$

$$\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} < 0 \ \ if \ \ y_i = -1$$

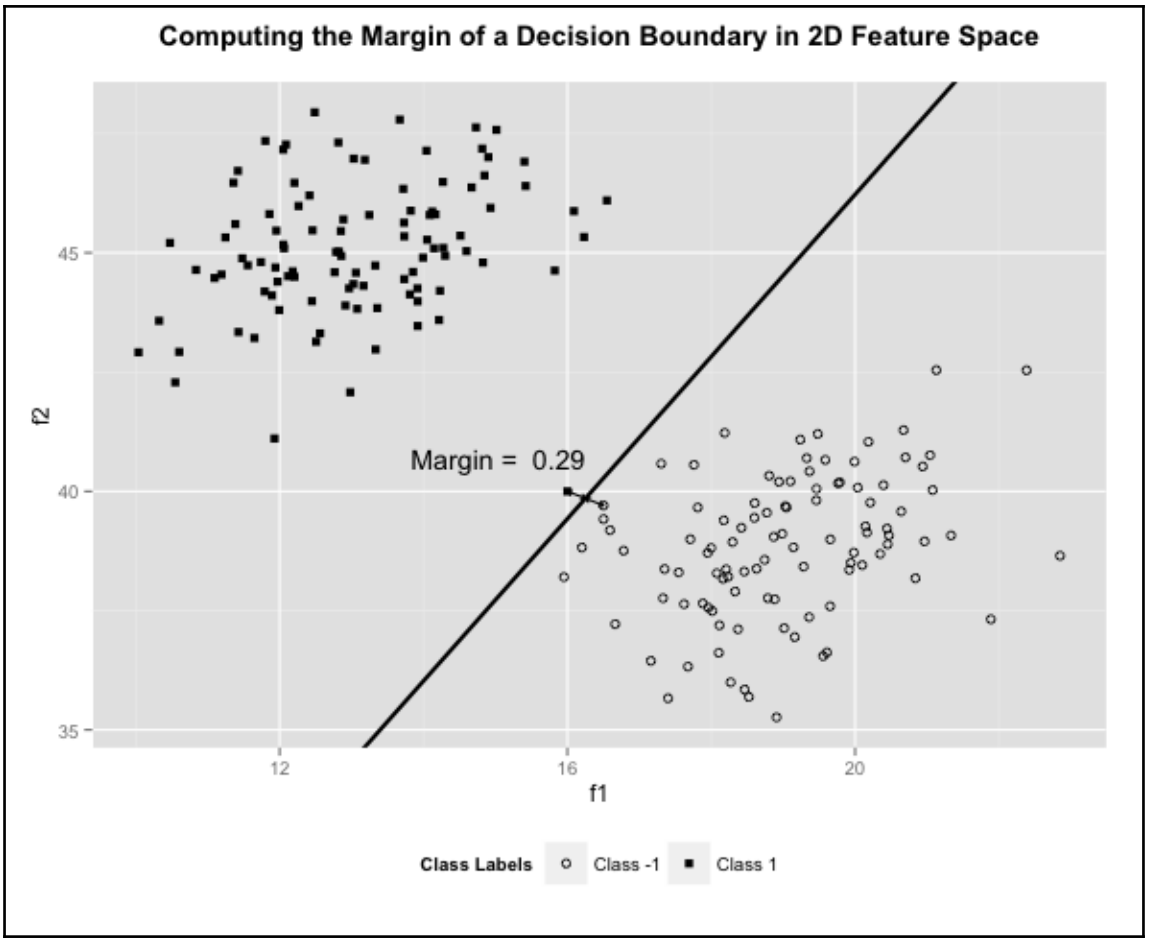
$$y_i \cdot \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) > 0, \forall i$$


Computing the Margin of a Decision Boundary in 2D Feature Space
Margin = 2
f2
f1
45
40
35
12
16
20
Class Labels
Class -1
Class 1

$$Select\ \beta_0, \beta_1 \ldots \beta_p\ that\ maximize\ M$$

$$such\ that: \sum_{k=1}^{p} \beta_k^2 = 1$$

$$and\ \forall i: y_i \cdot \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) \geq M$$


Computing the Margin of a Decision Boundary in 2D Feature Space
Margin = 0.29
f1
f2
12
16
20
35
40
45
Class Labels
Class -1
Class 1

$$Select\ \beta_0, \beta_1 \ldots \beta_p\ that\ maximize\ M$$

$$such\ that: \sum_{k=1}^{p} \beta_k^2 = 1$$

$$and\ \forall i: y_i \cdot \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) \geq M\left(1 - \xi_i\right)$$

$$and\ \forall i: \xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq C$$

$\xi_i = 0,$	x_i *is correctly classified, and outside the margin*
$0 < \xi_i \leq 1,$	x_i *is correctly classified, but falls inside the margin*
$\xi_i > 1$	x_i *is incorrectly correctly classified*

$$\langle v_1, v_2 \rangle = \sum_{i=1}^{p} v_{1i} \cdot v_{2i}$$

$$y = \beta_0 + \sum_{i=1}^{p} \beta_k x_k$$

$$y(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

$$y(x) = \beta_0 + \sum_{s \in S} \alpha_s \langle x, x_s \rangle$$

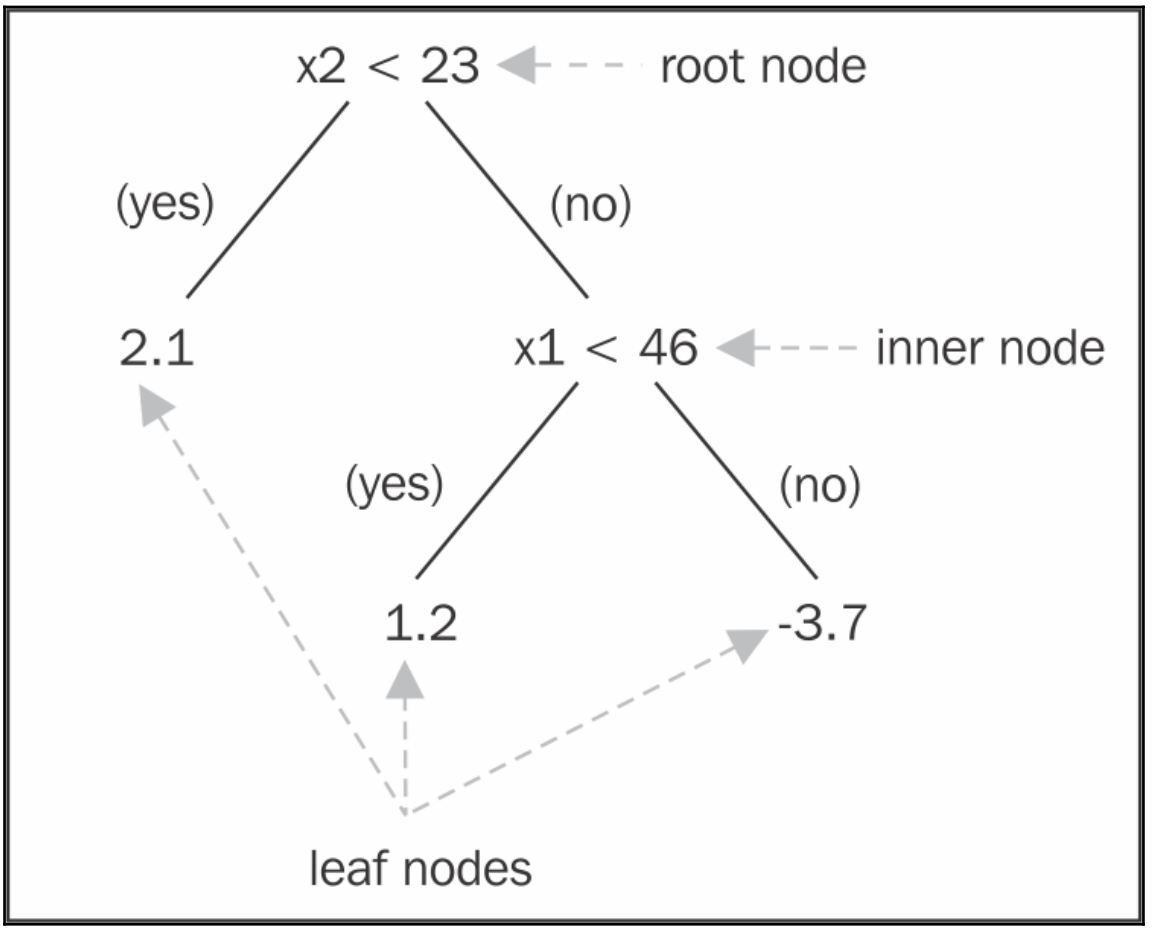
$$y(x) = \beta_0 + \sum_{s \in S} \alpha_s K\langle x, x_s \rangle$$

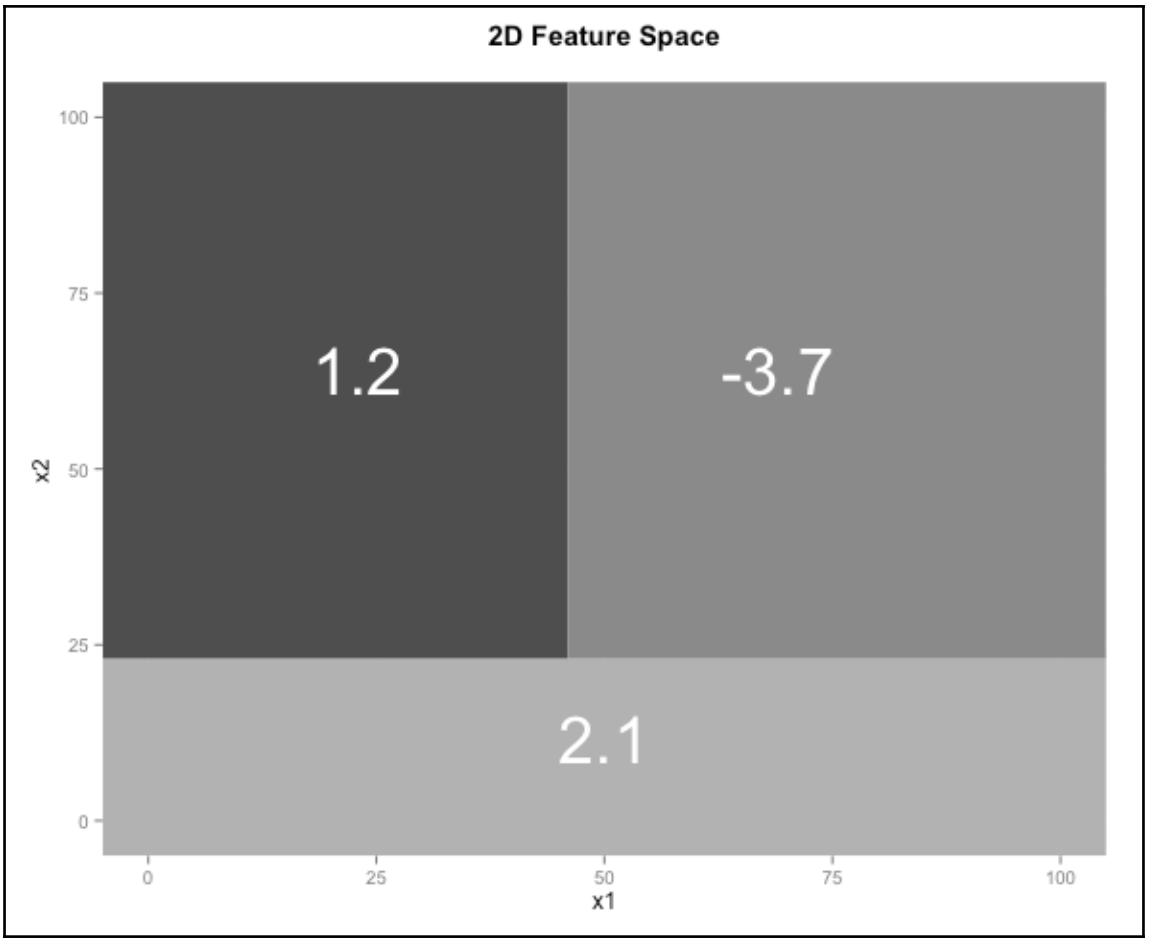
$$K_{linear}\left(x_i, x_j\right) = \sum_{k=1}^{p} x_{ik} x_{jk}$$

$$K_{polynomial}\left(x_i, x_j\right) = \left(1 + \sum_{k=1}^{p} x_{ik} x_{jk}\right)^{d}$$

$$K_{radial}\left(x_i, x_j\right) = e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}$$

Chapter 7: Tree-Based Methods

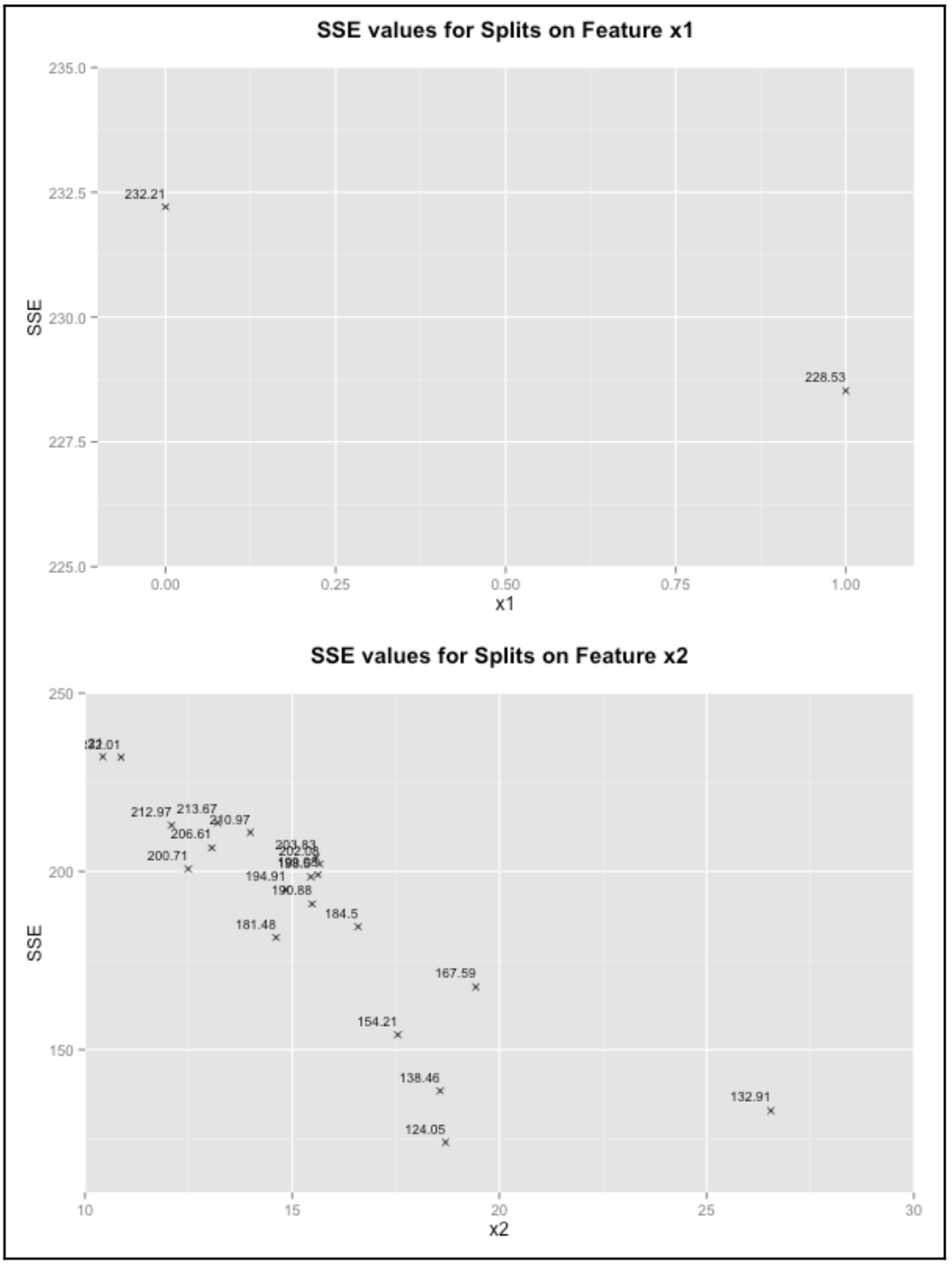




$$\overline{y}$$

$$SSE = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$$

$$SSE = \sum_{j=1}^{n_1} \left(y_j - \overline{y}_1 \right)^2 + \sum_{k=1}^{n_2} \left(y_k - \overline{y}_2 \right)^2$$


SSE values for Splits on Feature x1
SSE
x1
235.0
232.5
230.0
227.5
225.0
0.00
0.25
0.50
0.75
1.00
232.21
228.53
SSE values for Splits on Feature x2
SSE
x2
250
200
150
10
15
20
25
30
212.97
213.67
206.61
200.71
194.91
190.88
184.5
181.48
167.59
154.21
138.46
132.91
124.05

$$SSE_{penalized} = SSE + \alpha \cdot T_p$$

$$\sigma_{reduced} = \sigma_{initial} - \sum_{i=1}^{p} \frac{n_i}{n} \cdot \sigma_i$$

$$G = \sum_{k=1}^{K} \hat{p}_k \cdot \left(1 - \hat{p}_k\right)$$

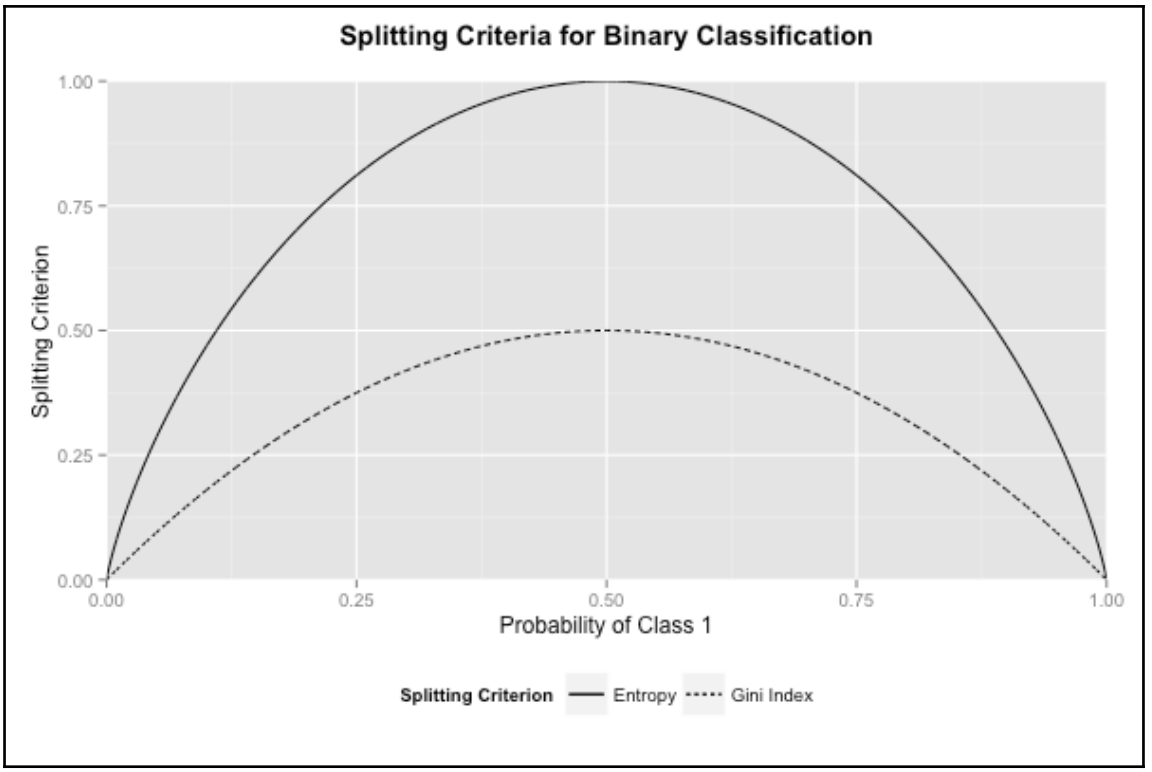
$$\hat{p}$$

$$Gini_{reduced} = Gini_{initial} - \sum_{i=1}^{p} \frac{n_i}{n} \cdot Gini_i$$

$$D = -2\sum_{k=1}^{K} n_k \cdot \log\left(\hat{p}_k\right)$$

$$Entropy = -\sum_{k=1}^{K} p_k \cdot \log_2 p_k$$

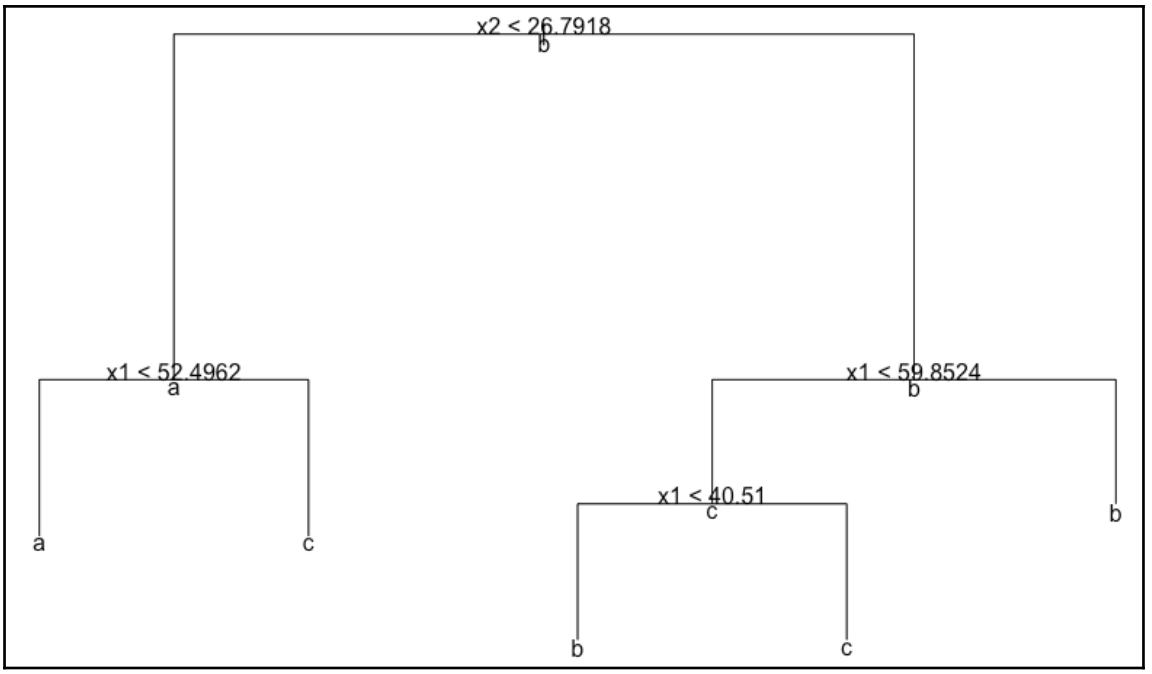
$$Binary\ Class\ Entropy = -\left[p \cdot \log_2 p + (1-p) \cdot \log_2 (1-p)\right]$$

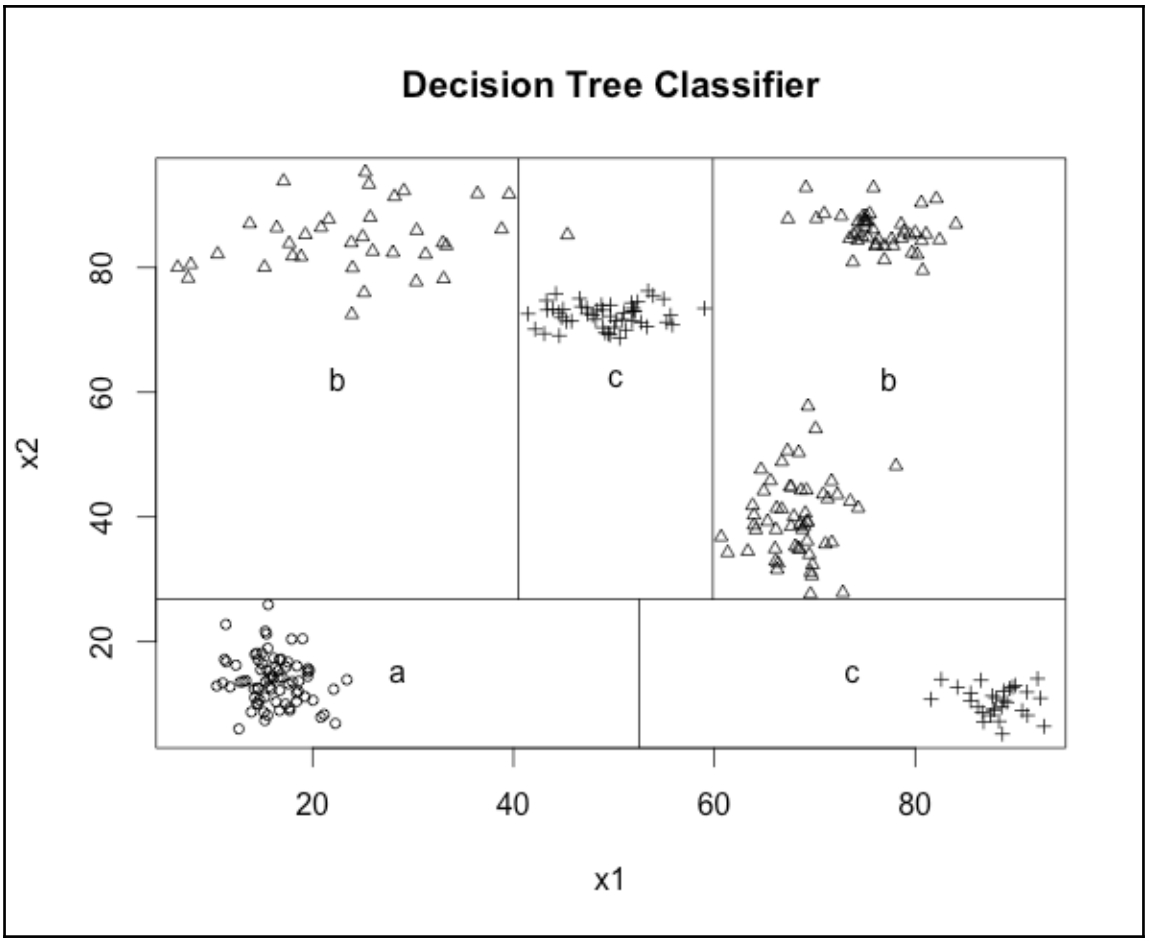


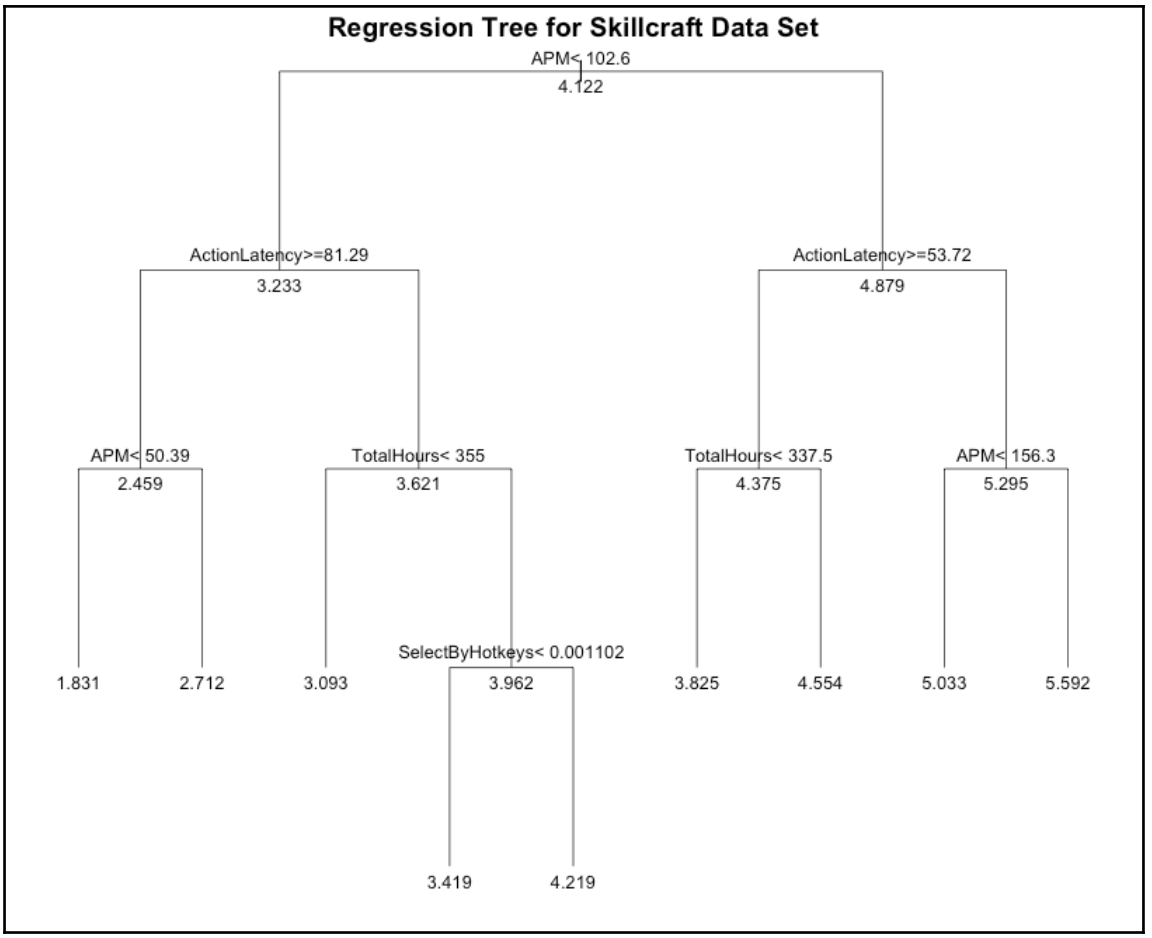
$$Information\,Gain = Entropy_{initial} - \sum_{i=1}^{p} \frac{n_i}{n} \cdot Entropy_i$$

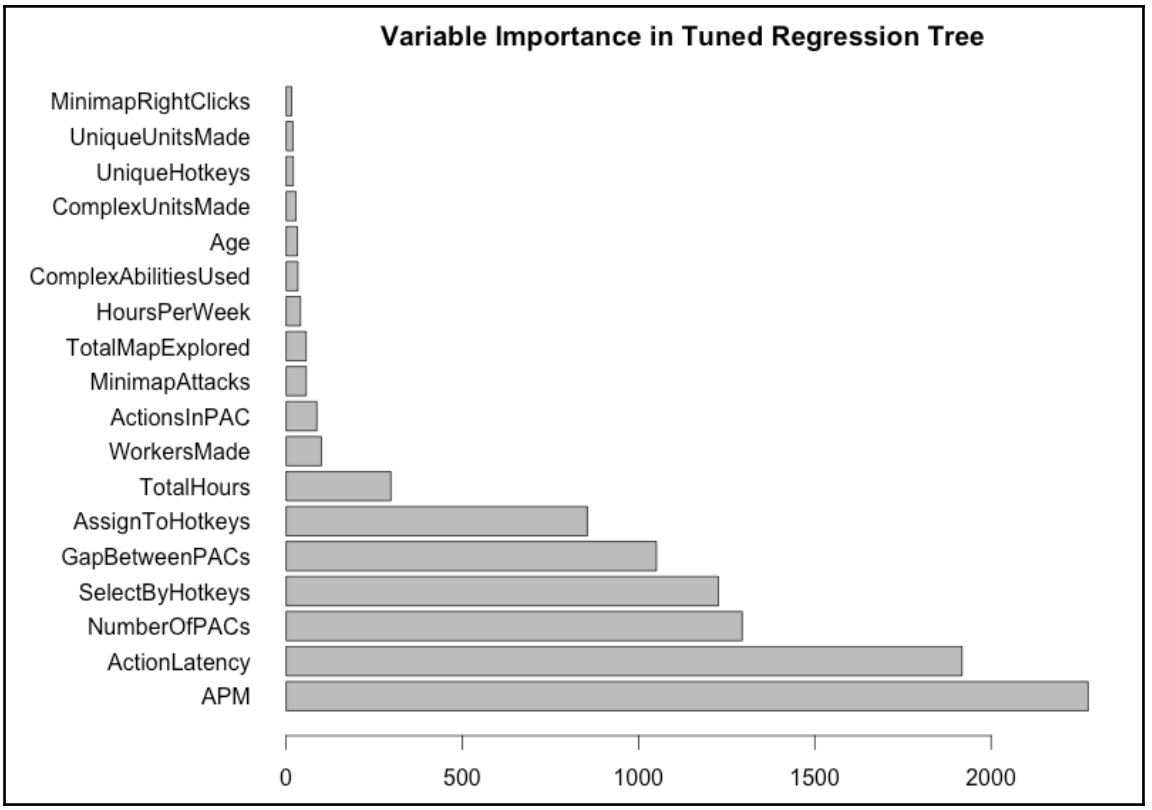
$$Information\,Gain\,Ratio = \frac{Information\,Gain}{Split\,Information\,Value}$$

$$Split\,Information\,Value = -\sum_{i=1}^{p} \frac{n_i}{n} \cdot \log_2\left(\frac{n_i}{n}\right)$$


x2 < 26.7918
b
x1 < 52.4962
a
x1 < 59.8524
b
x1 < 40.51
c
a
c
b
b
c

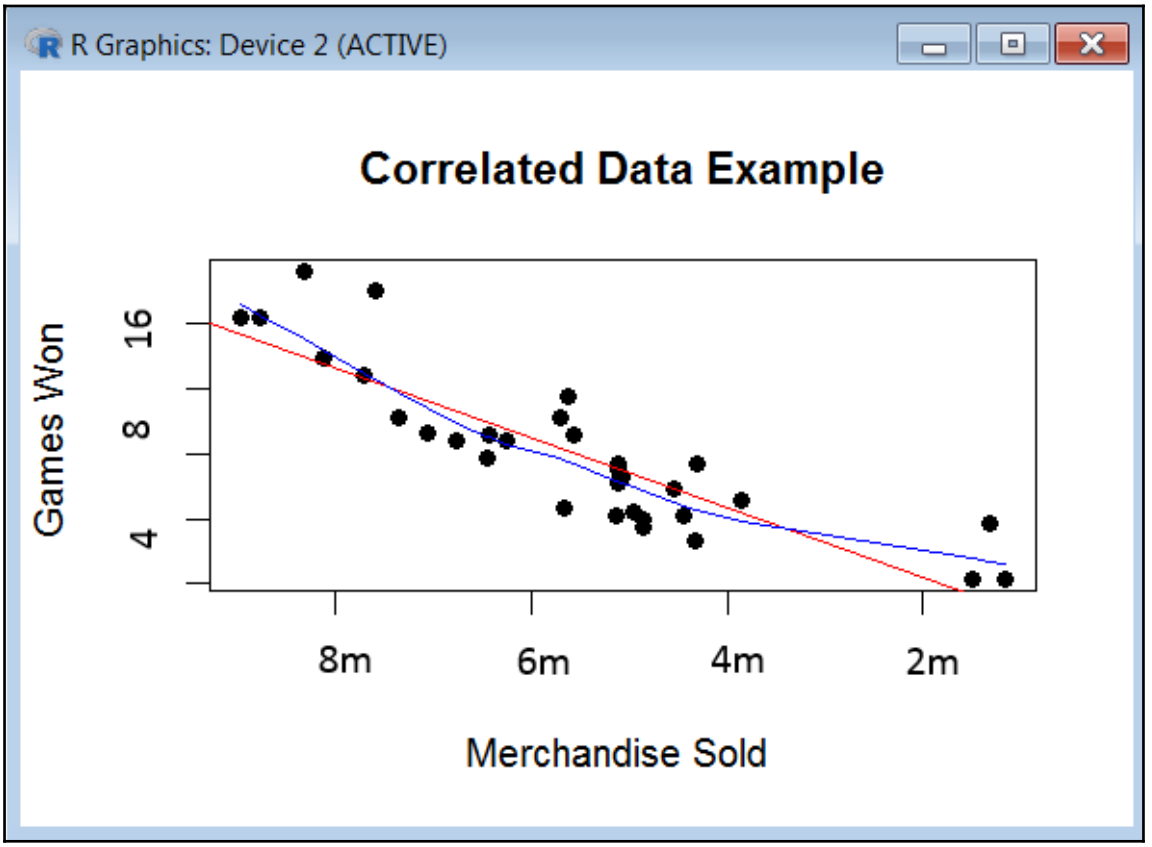

Decision Tree Classifier
x2
x1
20
40
60
80
b
c
b
a
c


Regression Tree for Skillcraft Data Set
APM< 102.6
4.122
ActionLatency>=81.29
3.233
ActionLatency>=53.72
4.879
APM< 50.39
2.459
TotalHours< 355
3.621
TotalHours< 337.5
4.375
APM< 156.3
5.295
1.831
2.712
3.093
SelectByHotkeys< 0.001102
3.962
3.825
4.554
5.033
5.592
3.419
4.219


Variable Importance in Tuned Regression Tree
MinimapRightClicks
UniqueUnitsMade
UniqueHotkeys
ComplexUnitsMade
Age
ComplexAbilitiesUsed
HoursPerWeek
TotalMapExplored
MinimapAttacks
ActionsInPAC
WorkersMade
TotalHours
AssignToHotkeys
GapBetweenPACs
SelectByHotkeys
NumberOfPACs
ActionLatency
APM
0
500
1000
1500
2000

$$y_i$$

Chapter 8: Dimensionality Reduction



R Console

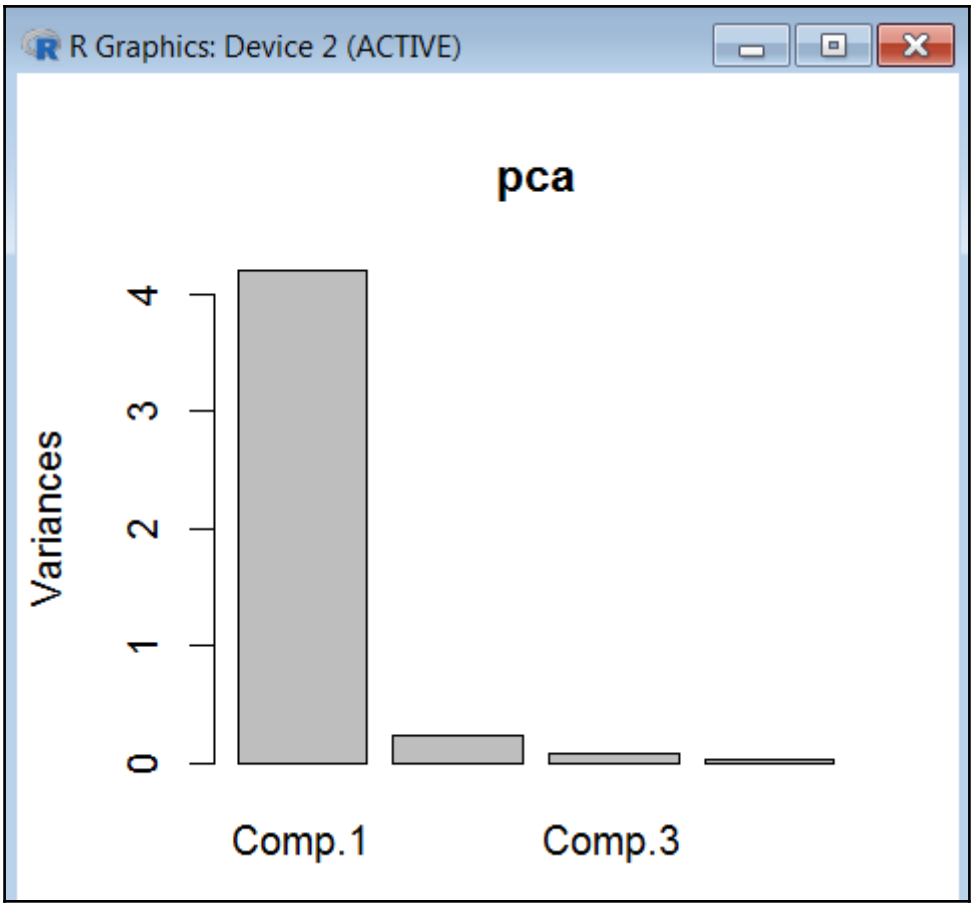
```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
```

R Console

```
> pca<-princomp(iris[-5])
> summary(pca)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
>
```

R Console

```
> pca<-princomp(iris[-5])
> summary(pca)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
> screeplot(pca)
>
```


R Graphics: Device 2 (ACTIVE)
pca
Variances
0
1
2
3
4
Comp.1
Comp.3

```
> pca$loadings

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.361 -0.657 -0.582  0.315
Sepal.Width         -0.730  0.598 -0.320
Petal.Length  0.857  0.173        -0.480
Petal.Width   0.358         0.546  0.754

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> pca$scores
             Comp.1       Comp.2       Comp.3        Comp.4
  [1,] -2.684125626 -0.319397247 -0.027914828  0.0022624371
  [2,] -2.714141687  0.177001225 -0.210464272  0.0990265503
  [3,] -2.888990569  0.144949426  0.017900256  0.0199683897
  [4,] -2.745342856  0.318298979  0.031559374 -0.0755758166
  [5,] -2.728716537 -0.326754513  0.090079241 -0.0612585926
  [6,] -2.280859633 -0.741330449  0.168677658 -0.0242008576
  [7,] -2.820537751  0.089461385  0.257892158 -0.0481431065
  [8,] -2.626144973 -0.163384960 -0.021879318 -0.0452978706
  [9,] -2.886382732  0.578311754  0.020759570 -0.0267447358
 [10,] -2.672755798  0.113774246 -0.197632725 -0.0562954013
```

```
> v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
> v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
> v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
> v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
> v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
> v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
> m1 <- cbind(v1,v2,v3,v4,v5,v6)
> summary(m1)
       v1              v2              v3              v4              v5
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
 Median :1.000   Median :2.000   Median :1.000   Median :2.000   Median :1.000
 Mean   :2.222   Mean   :2.444   Mean   :2.222   Mean   :2.389   Mean   :2.222
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
 Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
       v6
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.389
 3rd Qu.:3.000
 Max.   :6.000
>
```

```
> cor(m1)
          v1        v2        v3        v4        v5        v6
v1 1.0000000 0.9393083 0.5128866 0.4320310 0.4664948 0.4086076
v2 0.9393083 1.0000000 0.4124441 0.4084281 0.4363925 0.4326113
v3 0.5128866 0.4124441 1.0000000 0.8770750 0.5128866 0.4320310
v4 0.4320310 0.4084281 0.8770750 1.0000000 0.4320310 0.4323259
v5 0.4664948 0.4363925 0.5128866 0.4320310 1.0000000 0.9473451
v6 0.4086076 0.4326113 0.4320310 0.4323259 0.9473451 1.0000000
>
```

```
> factanal(m1, factors = 3)

Call:
factanal(x = m1, factors = 3)

Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005

Loadings:
   Factor1 Factor2 Factor3
v1 0.944   0.182   0.267
v2 0.905   0.235   0.159
v3 0.236   0.210   0.946
v4 0.180   0.242   0.828
v5 0.242   0.881   0.286
v6 0.193   0.959   0.196

                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
Proportion Var    0.316   0.314   0.300
Cumulative Var    0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
>
```

```
> factanal(m1, factors = 4)
Error in factanal(m1, factors = 4) :
  4 factors are too many for 6 variables
>
```

```
> factanal(m1, factors = 2)

Call:
factanal(x = m1, factors = 2)

Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.114 0.642 0.742 0.005 0.097

Loadings:
   Factor1 Factor2
v1 0.971   0.228
v2 0.917   0.213
v3 0.429   0.418
v4 0.363   0.355
v5 0.254   0.965
v6 0.205   0.928

                Factor1 Factor2
SS loadings       2.206   2.190
Proportion Var    0.368   0.365
Cumulative Var    0.368   0.733

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 23.14 on 4 degrees of freedom.
The p-value is 0.000119
> 
```

Chapter 9: Ensemble Methods

$$\left(1-\frac{1}{n}\right)^n$$

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{n} w_i}$$

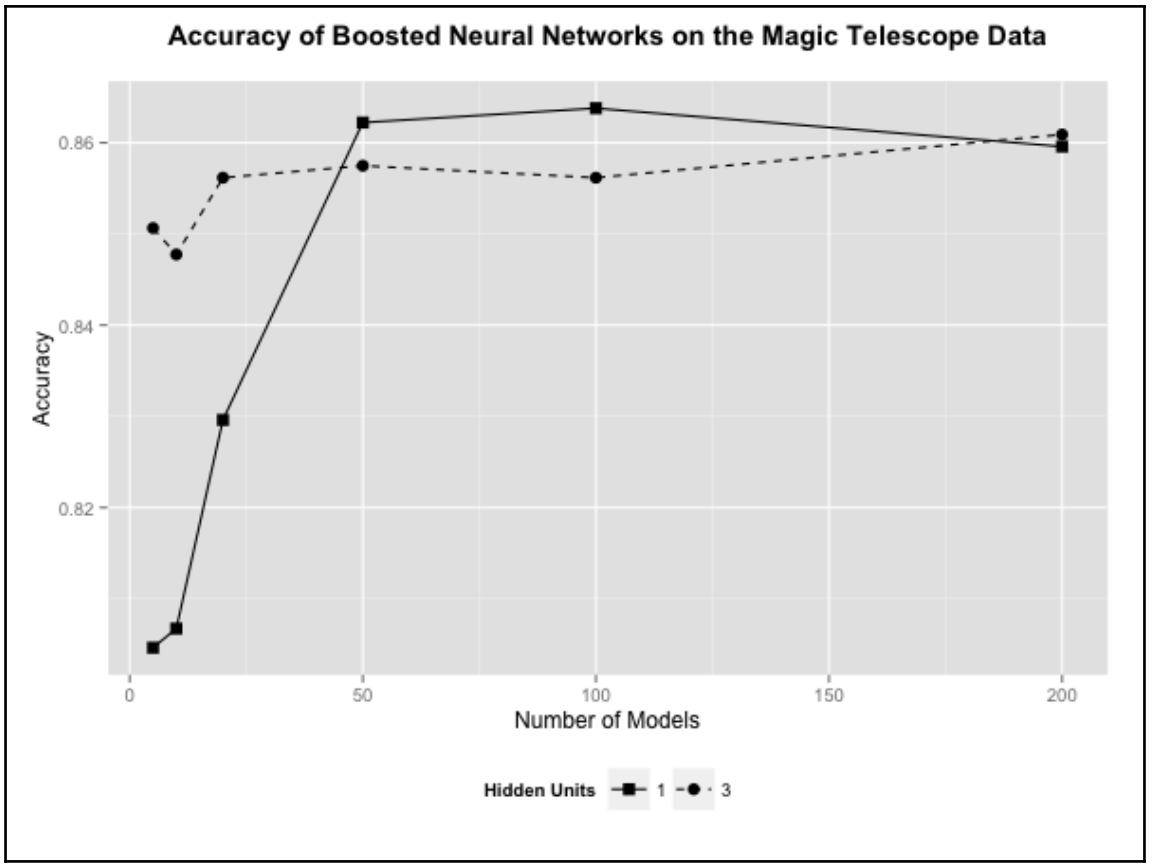
$$\alpha_m = \frac{1}{2}\log_e\left(\frac{1-err_m}{err_m}\right)$$

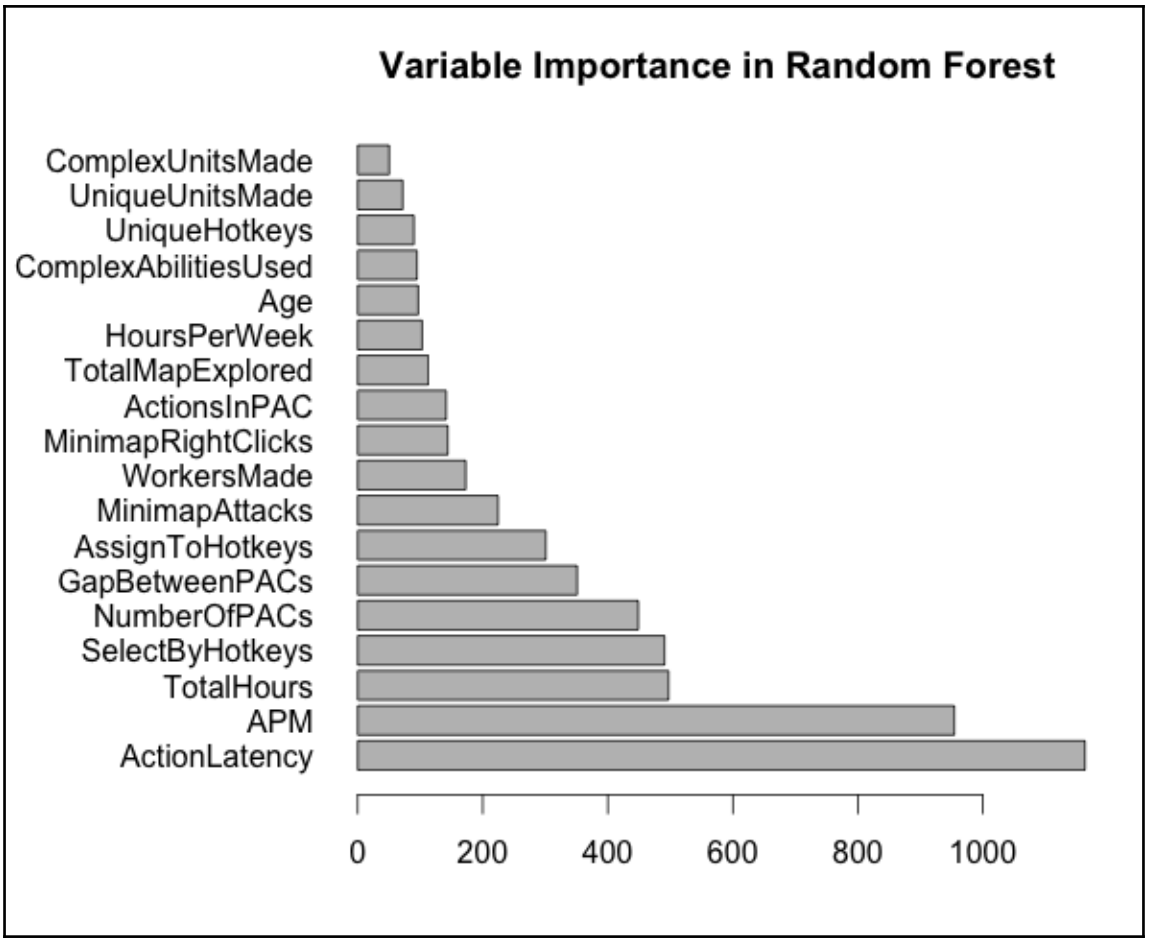
$$e^{\alpha_m}$$

$$e^{-\alpha_m}$$

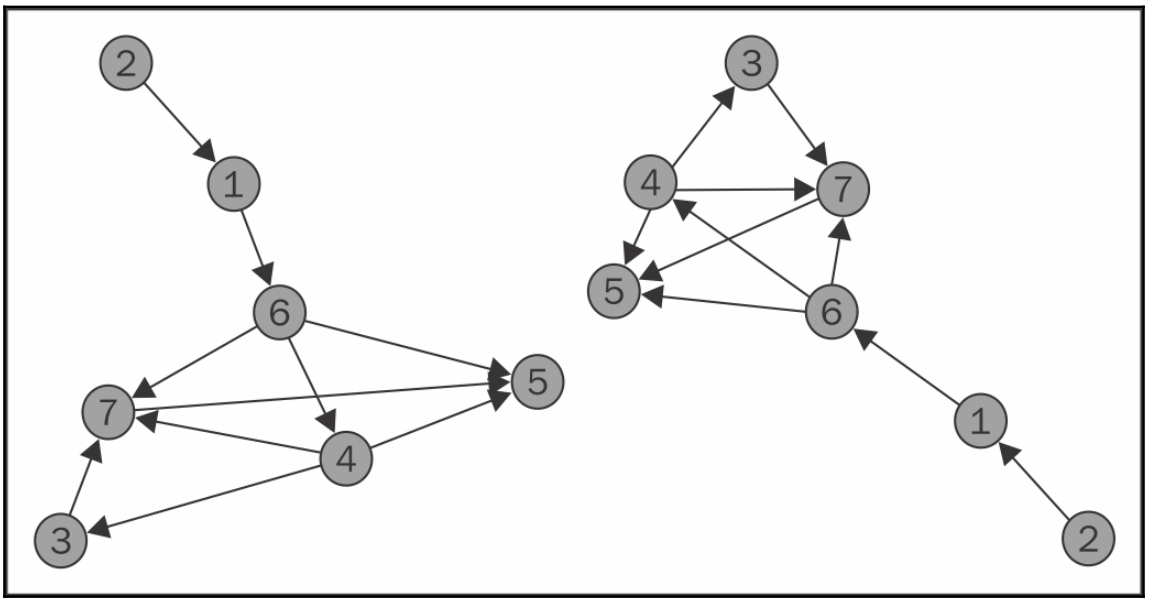
$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

$$\sigma^2 / n$$


Accuracy of Boosted Neural Networks on the Magic Telescope Data
Accuracy
0.86
0.84
0.82
0
50
100
150
200
Number of Models
Hidden Units
1
3


Variable Importance in Random Forest
ComplexUnitsMade
UniqueUnitsMade
UniqueHotkeys
ComplexAbilitiesUsed
Age
HoursPerWeek
TotalMapExplored
ActionsInPAC
MinimapRightClicks
WorkersMade
MinimapAttacks
AssignToHotkeys
GapBetweenPACs
NumberOfPACs
SelectByHotkeys
TotalHours
APM
ActionLatency
0
200
400
600
800
1000

Chapter 10: Probabilistic Graphical Models



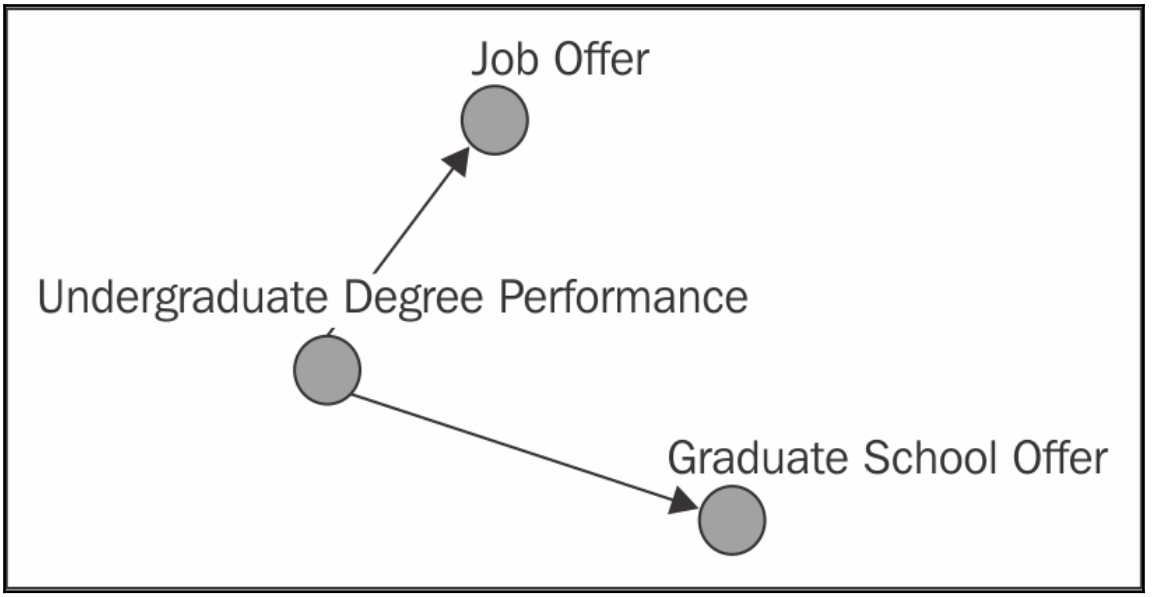
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P_{independent\ events}(A \mid B) = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

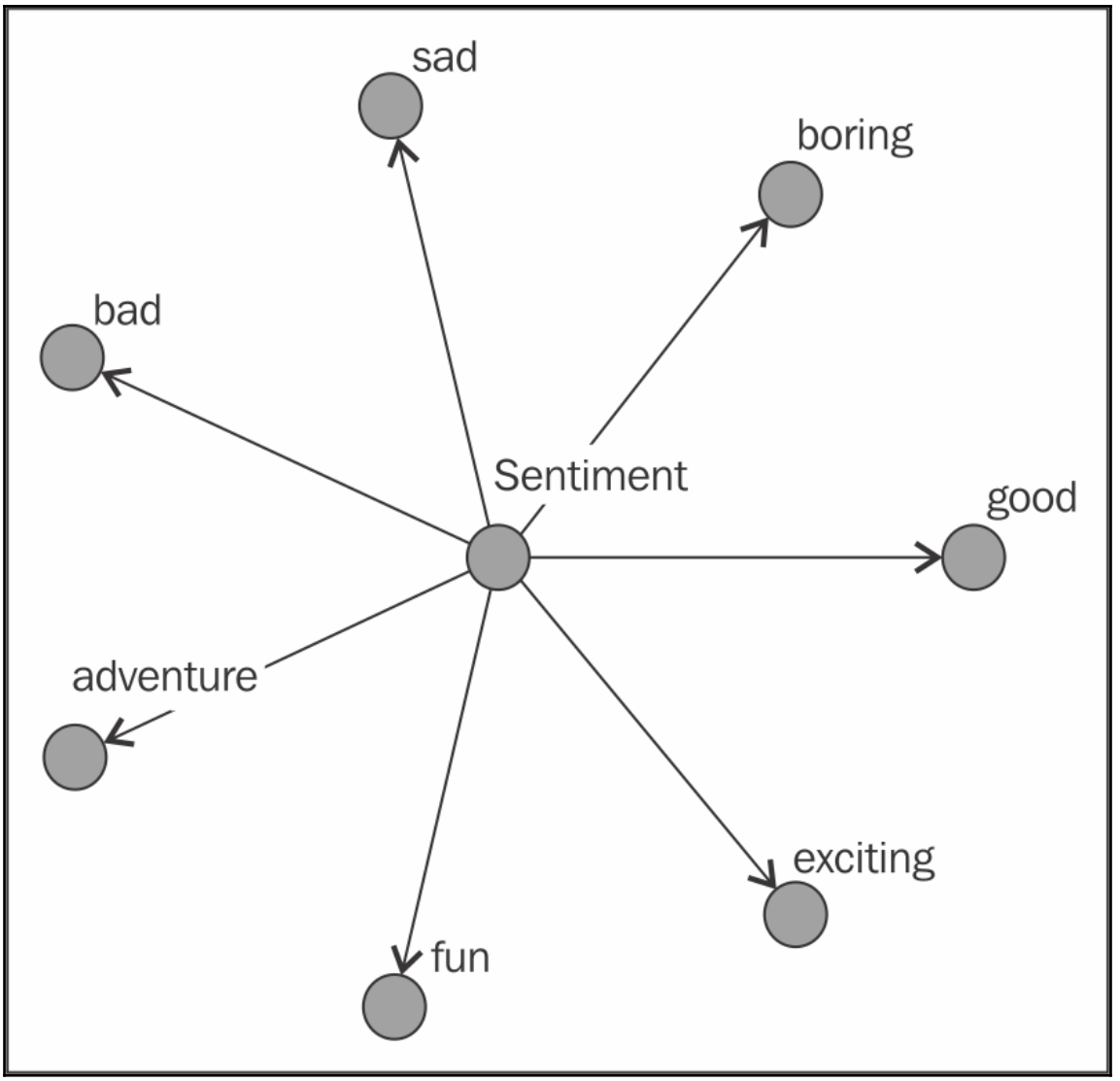
$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$$



$$P(G, J, U) = P(G \mid J, U) \cdot P(J, U) = P(G \mid J, U) \cdot P(J \mid U) \cdot P(U)$$

$$P(G,J,U) = P(G \mid U) \cdot P(J \mid U) \cdot P(U)$$

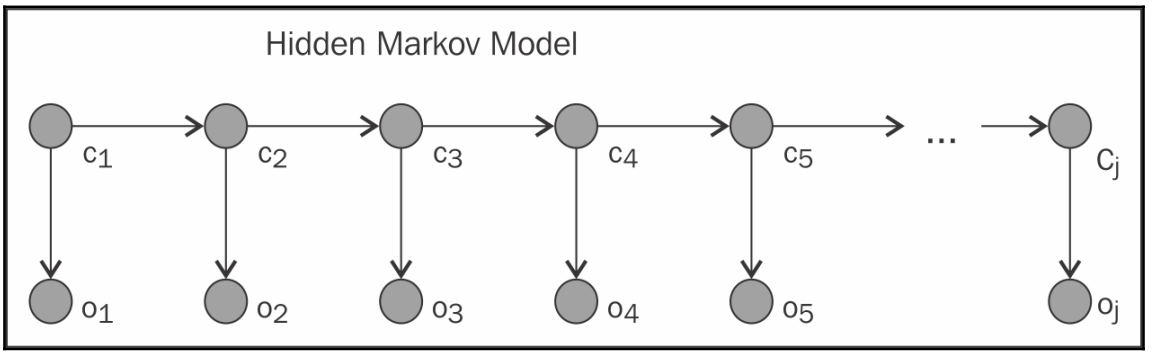
$$P(A_1, A_2, A_3, A_4, A_5, A_6, A_7) =$$
$$P(A_2) \cdot P(A_1 \mid A_2) \cdot P(A_6 \mid A_1) \cdot P(A_4 \mid A_6) \cdot$$
$$P(A_3 \mid A_4) \cdot P(A_7 \mid A_3, A_4, A_6) \cdot P(A_5 \mid A_4, A_6, A_7)$$

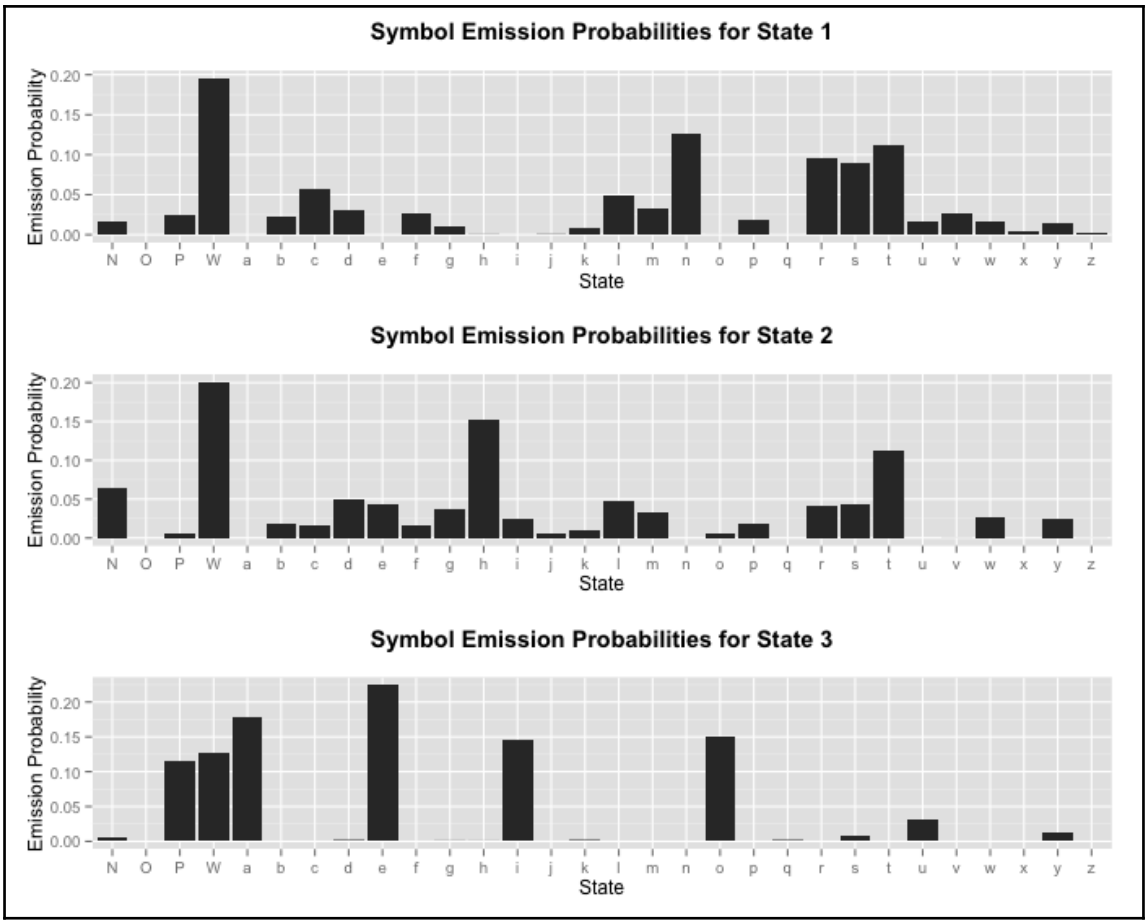


$$P(C \mid F_1, \ldots, F_n) = \frac{P(C) \cdot P(F_1, \ldots, F_n \mid C)}{P(F_1, \ldots, F_n)}$$

$$P(C \mid F_1, \ldots, F_n) = \frac{P(C) \cdot P(F_1 \mid C) \cdot \ldots \cdot P(F_n \mid C)}{P(F_1, \ldots, F_n)}$$

$$Classify\, C_i : \underset{c}{\operatorname{argmax}}\, P(C) \cdot \prod_{i=1}^{n} P(F_i \mid C)$$

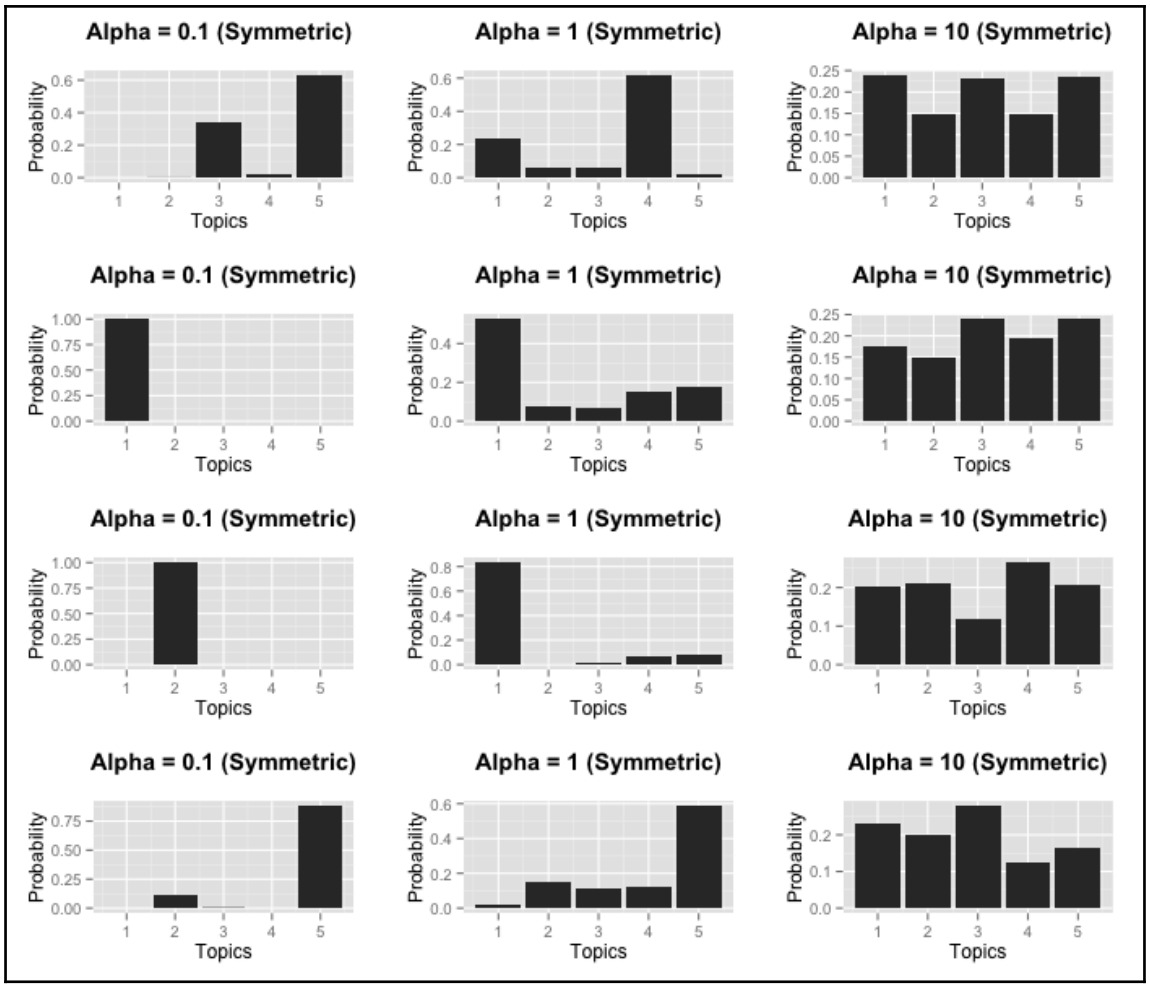


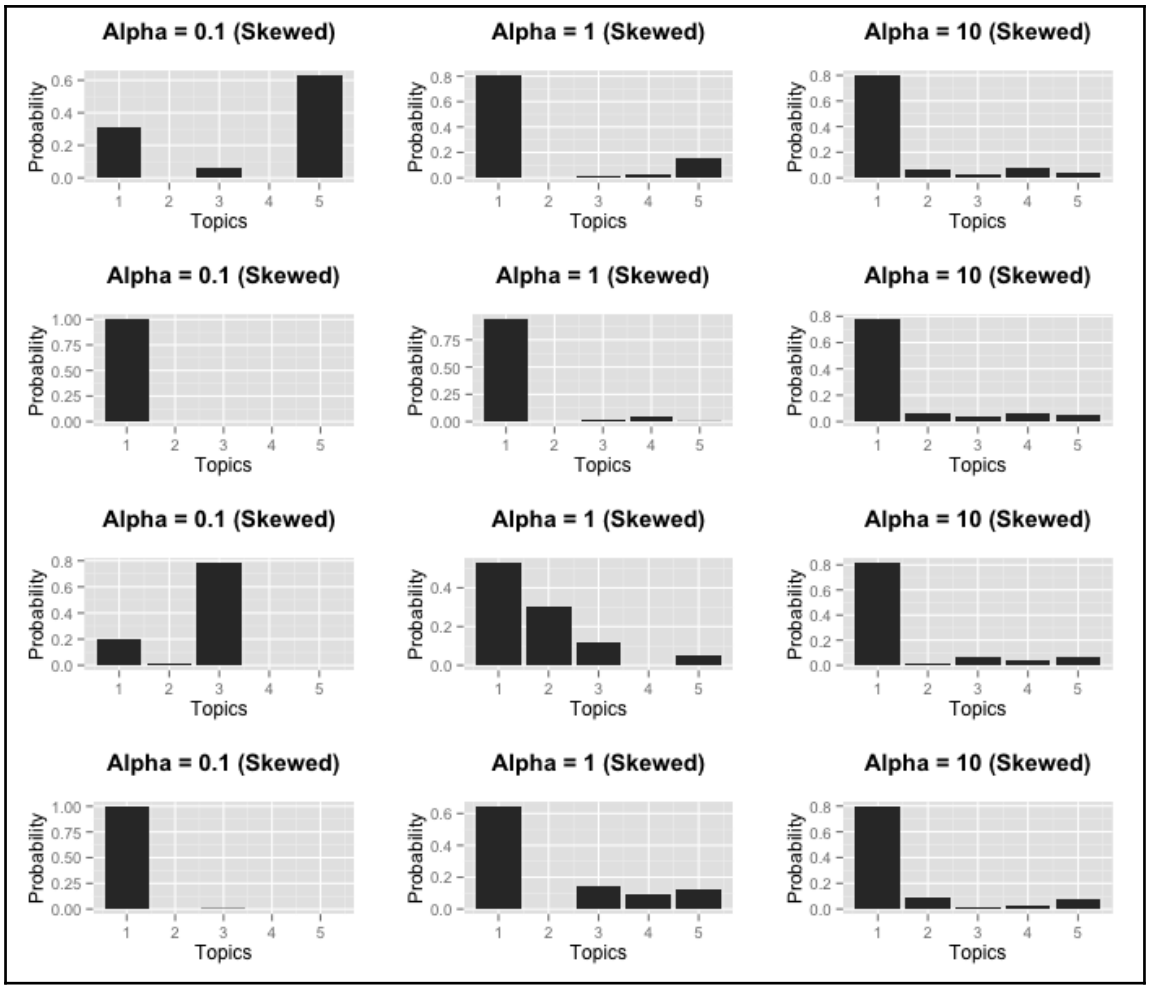

Symbol Emission Probabilities for State 1
Emission Probability
0.20
0.15
0.10
0.05
0.00
N O P W a b c d e f g h i j k l m n o p q r s t u v w x y z
State
Symbol Emission Probabilities for State 2
Emission Probability
0.20
0.15
0.10
0.05
0.00
N O P W a b c d e f g h i j k l m n o p q r s t u v w x y z
State
Symbol Emission Probabilities for State 3
Emission Probability
0.20
0.15
0.10
0.05
0.00
N O P W a b c d e f g h i j k l m n o p q r s t u v w x y z
State

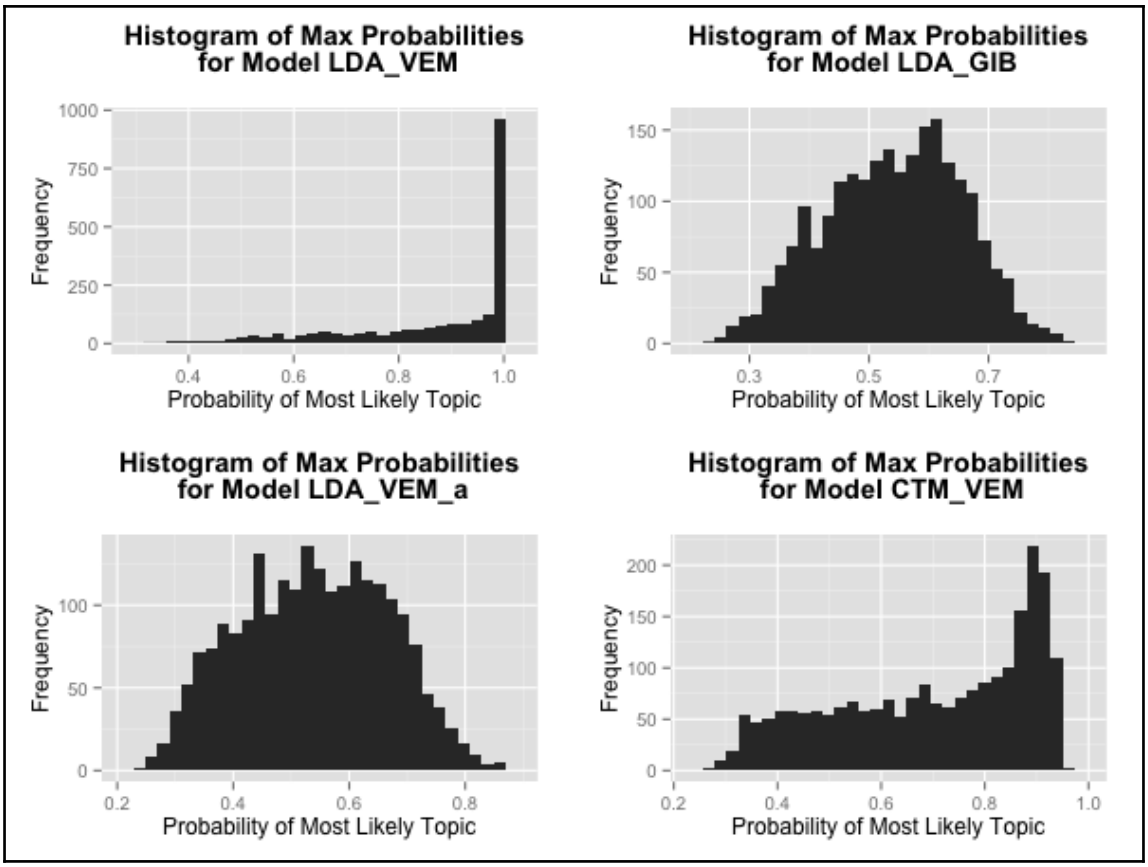
Chapter 11: Topic Modeling

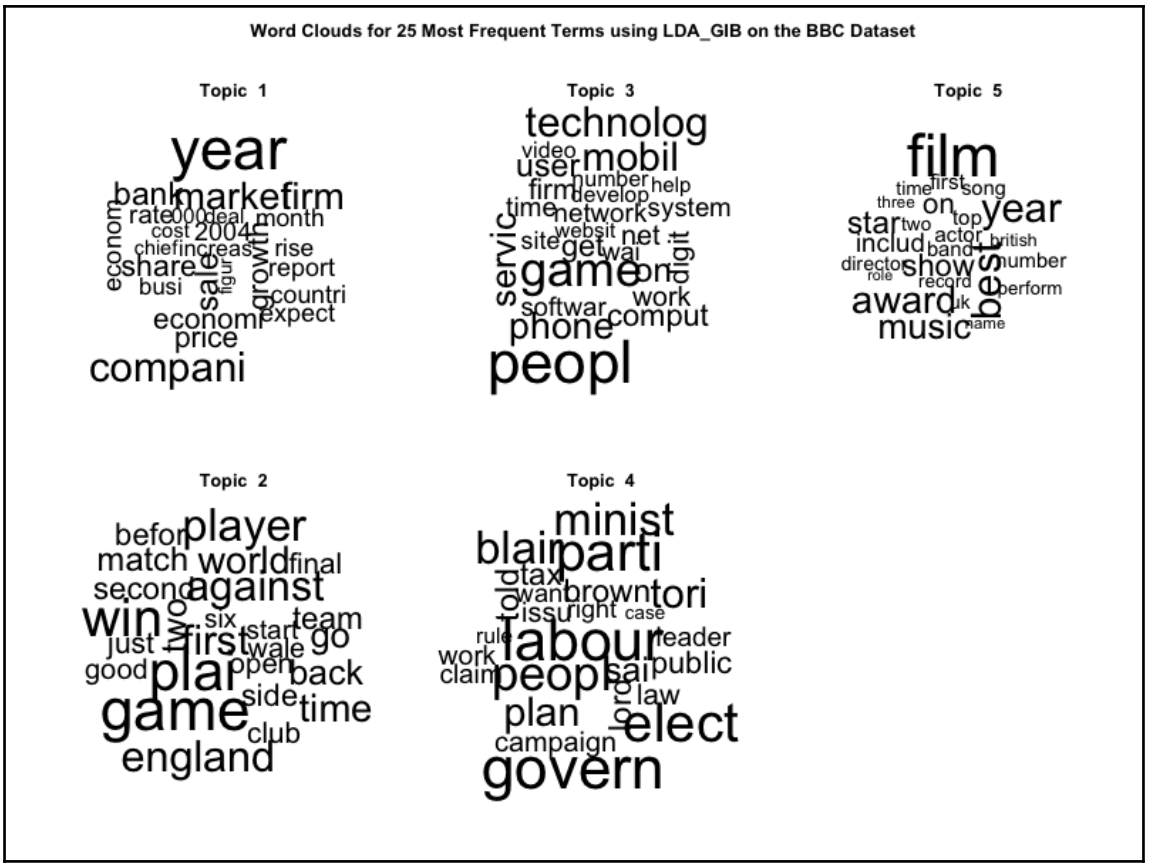
$$P(\overline{x} \mid \overline{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \prod_{k=1}^{K} x_k^{\alpha_{k-1}}$$

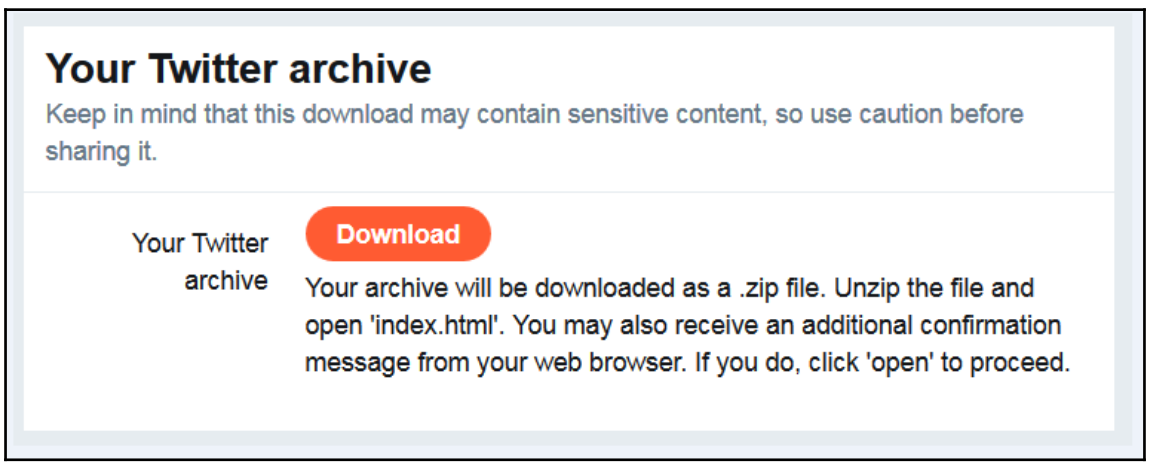
$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$

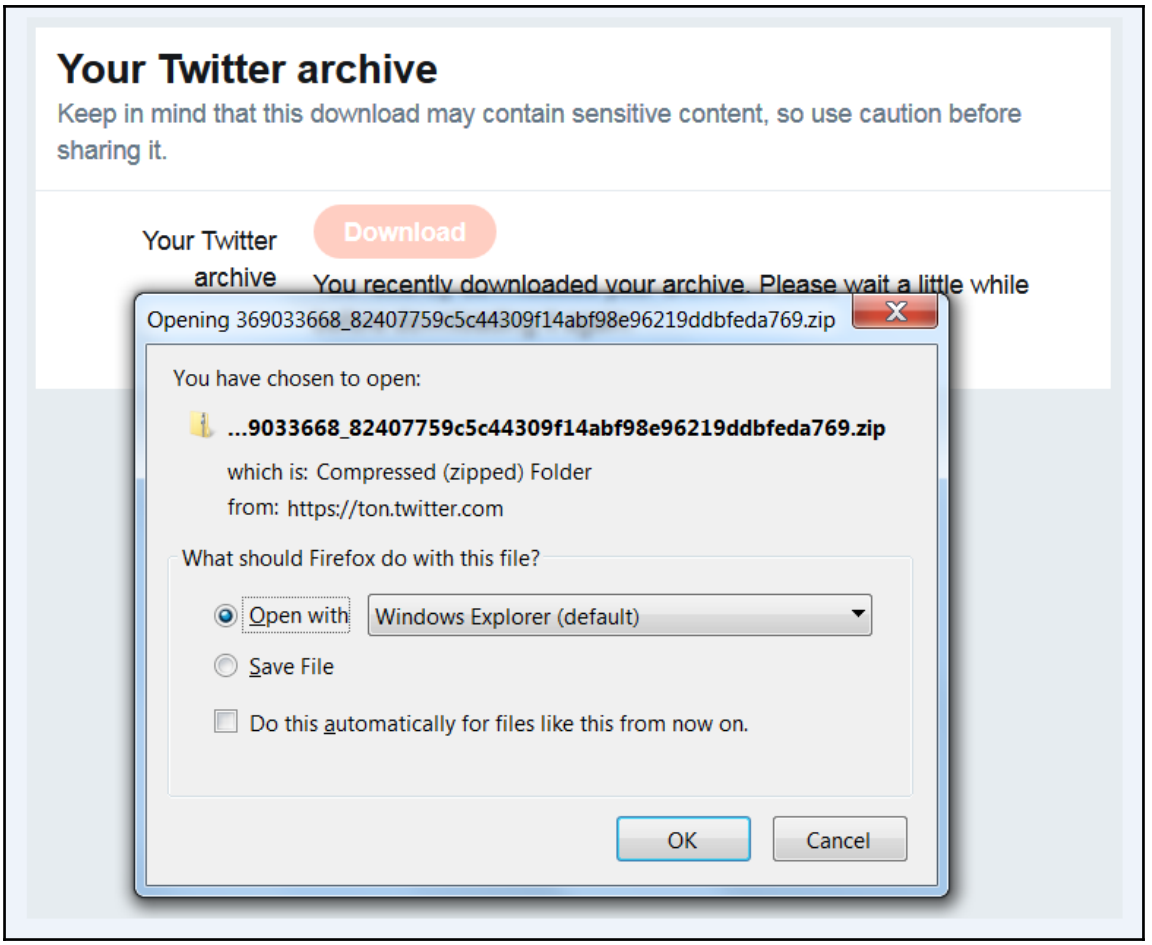

Alpha = 0.1 (Symmetric)
Alpha = 1 (Symmetric)
Alpha = 10 (Symmetric)
Alpha = 0.1 (Symmetric)
Alpha = 1 (Symmetric)
Alpha = 10 (Symmetric)
Alpha = 0.1 (Symmetric)
Alpha = 1 (Symmetric)
Alpha = 10 (Symmetric)
Alpha = 0.1 (Symmetric)
Alpha = 1 (Symmetric)
Alpha = 10 (Symmetric)
Probability
Topics


Alpha = 0.1 (Skewed)
Alpha = 1 (Skewed)
Alpha = 10 (Skewed)
Alpha = 0.1 (Skewed)
Alpha = 1 (Skewed)
Alpha = 10 (Skewed)
Alpha = 0.1 (Skewed)
Alpha = 1 (Skewed)
Alpha = 10 (Skewed)
Alpha = 0.1 (Skewed)
Alpha = 1 (Skewed)
Alpha = 10 (Skewed)
Probability
Topics


Histogram of Max Probabilities for Model LDA_VEM
Frequency
1000
750
500
250
0
0.4
0.6
0.8
1.0
Probability of Most Likely Topic
Histogram of Max Probabilities for Model LDA_GIB
Frequency
150
100
50
0
0.3
0.5
0.7
Probability of Most Likely Topic
Histogram of Max Probabilities for Model LDA_VEM_a
Frequency
100
50
0
0.2
0.4
0.6
0.8
Probability of Most Likely Topic
Histogram of Max Probabilities for Model CTM_VEM
Frequency
200
150
100
50
0
0.2
0.4
0.6
0.8
1.0
Probability of Most Likely Topic


Word Clouds for 25 Most Frequent Terms using LDA_GIB on the BBC Dataset
Topic 1
year
bank
market
firm
rate
deal
month
cost
2004
increas
share
sale
rise
report
busi
countri
expect
economi
price
compani
Topic 3
technolog
video
mobil
user
number
help
develop
firm
time
network
system
websit
net
site
get
game
phone
softwar
comput
work
peopl
Topic 5
film
time
first
song
three
on
year
star
two
top
includ
actor
band
british
director
show
number
role
record
award
perform
uk
music
name
Topic 2
player
match
world
final
against
win
six
team
just
first
start
wale
go
good
open
back
play
side
game
time
club
england
Topic 4
minist
blair
parti
tax
brown
tori
issu
right
case
labour
reader
work
claim
peopl
public
law
plan
elect
campaign
govern


Your Twitter archive
Keep in mind that this download may contain sensitive content, so use caution before sharing it.
Your Twitter archive
Download
Your archive will be downloaded as a .zip file. Unzip the file and open 'index.html'. You may also receive an additional confirmation message from your web browser. If you do, click 'open' to proceed.

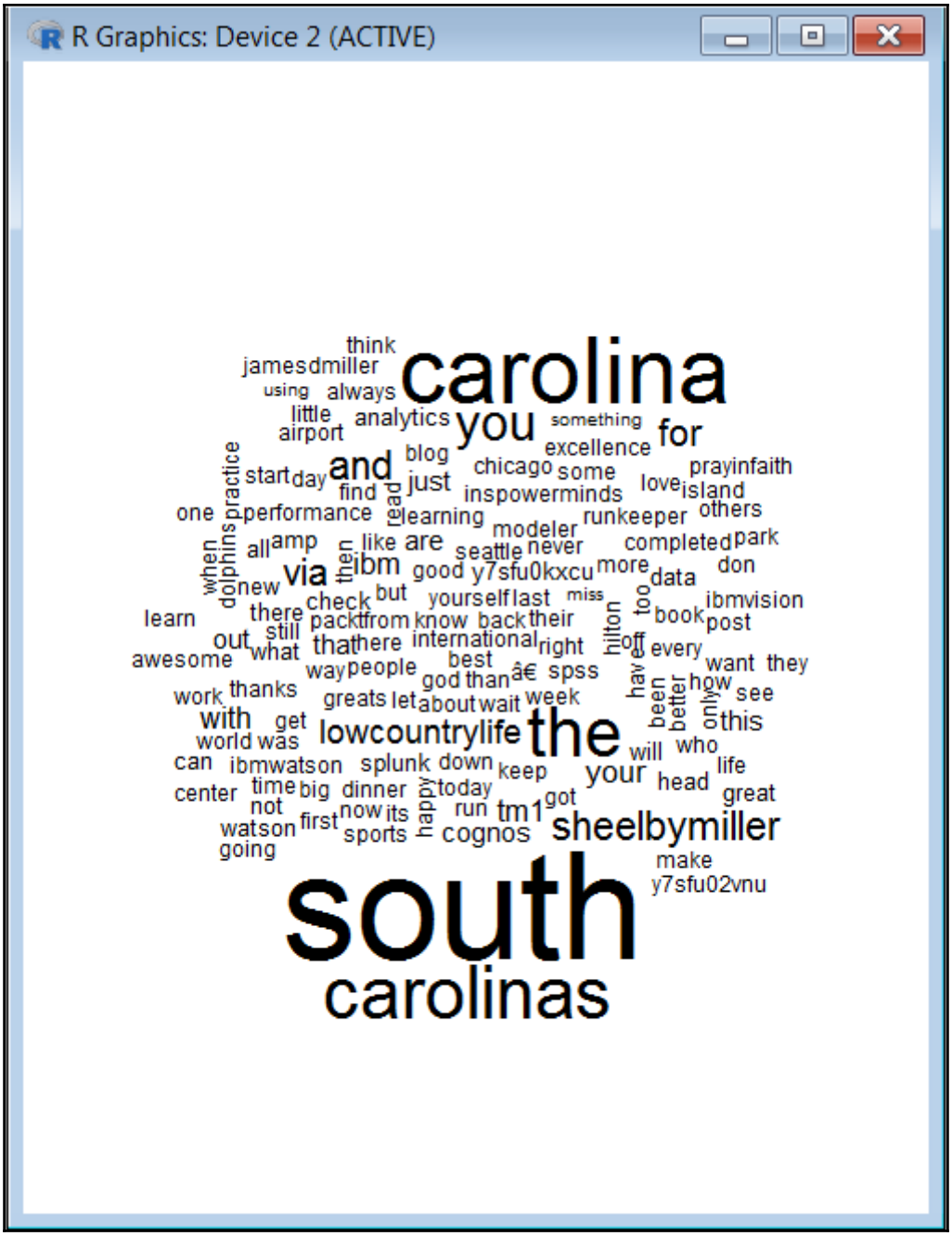


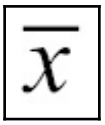

```
>     x <- read.csv("c:/Worker/mytweets2.csv")
>     y <- list(x)
>     df1 <- do.call("rbind", lapply(y, as.data.frame))
```

```
> library(tm)
> myCorpus<-Corpus(VectorSource(df1$text))
>
```

```
> myDtm <- TermDocumentMatrix(myCorpus, control = list(minWordLength = 1))
>
```

```
> myDtm <- TermDocumentMatrix(myCorpus, control = list(minWordLength = 1))
> m <- as.matrix(myDtm)
> v <- sort(rowSums(m), decreasing=TRUE)
> myNames <- names(v)
> k <- which(names(v)=="miners")
> myNames[k] <- "mining"
> d <- data.frame(word=myNames, freq=v)
>
> library(wordcloud)
> wordcloud(d$word, d$freq, min.freq=17)
>
```


R Graphics: Device 2 (ACTIVE)
carolina
south
carolinas
you
the
via
and
for
sheelbymiller
lowcountrylife
tm1
think
jamesdmiller
using
always
little
airport
analytics
something
excellence
blog
chicago
some
prayinfaith
practice
start
day
find
read
just
inspowerminds
love
island
one
performance
learning
runkeeper
others
modeler
all
amp
like
are
seattle
never
completed
park
when
dolphins
ibm
good
y7sfu0kxcu
more
data
don
new
check
but
yourself
last
miss
too
ibmvision
learn
there
packtfrom
know
back
their
hilton
book
post
out
still
thatthere
international
right
off
every
awesome
what
way
people
best
god
than
â€
spss
have
want
they
work
thanks
greats
let
about
wait
week
been
better
how
only
see
with
get
world
was
this
can
ibmwatson
splunk
down
keep
will
who
life
center
time
big
dinner
today
happy
got
your
head
great
not
watson
first
now
its
run
sports
cognos
going
make
y7sfu02vnu



Chapter 12: Recommendation Systems

$$d_{euclidean}(a,b) = \sqrt{\sum_{i=1}^{i=n} (a_i - b_i)^2}$$

$$d_{cosine}(a,b) = \cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|}$$

$$d_{jaccard}(a,b) = 1 - \frac{|a \cap b|}{|a \cup b|}$$

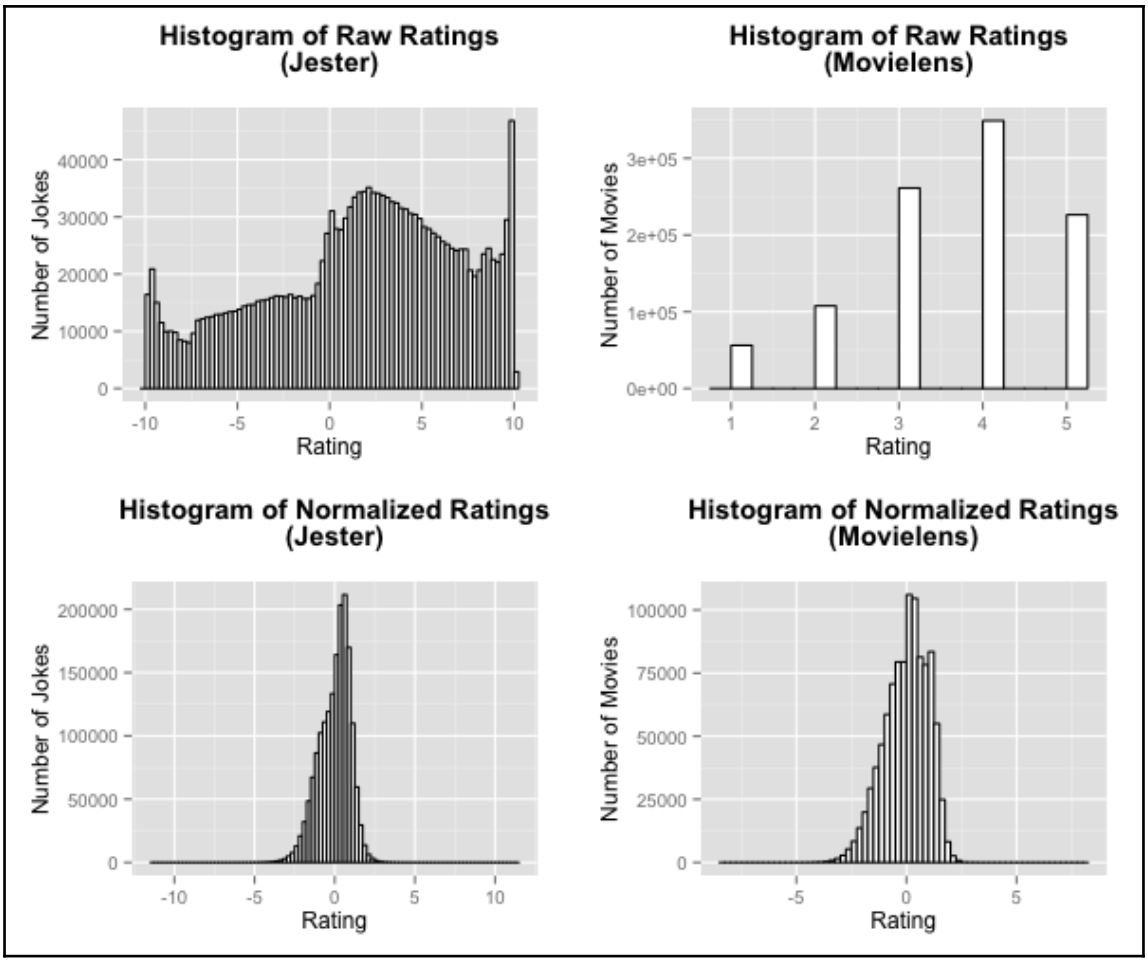
$$\hat{r}_{tj} = \frac{1}{|N(t)|} \sum_{u \in N(t)} r_{uj}$$

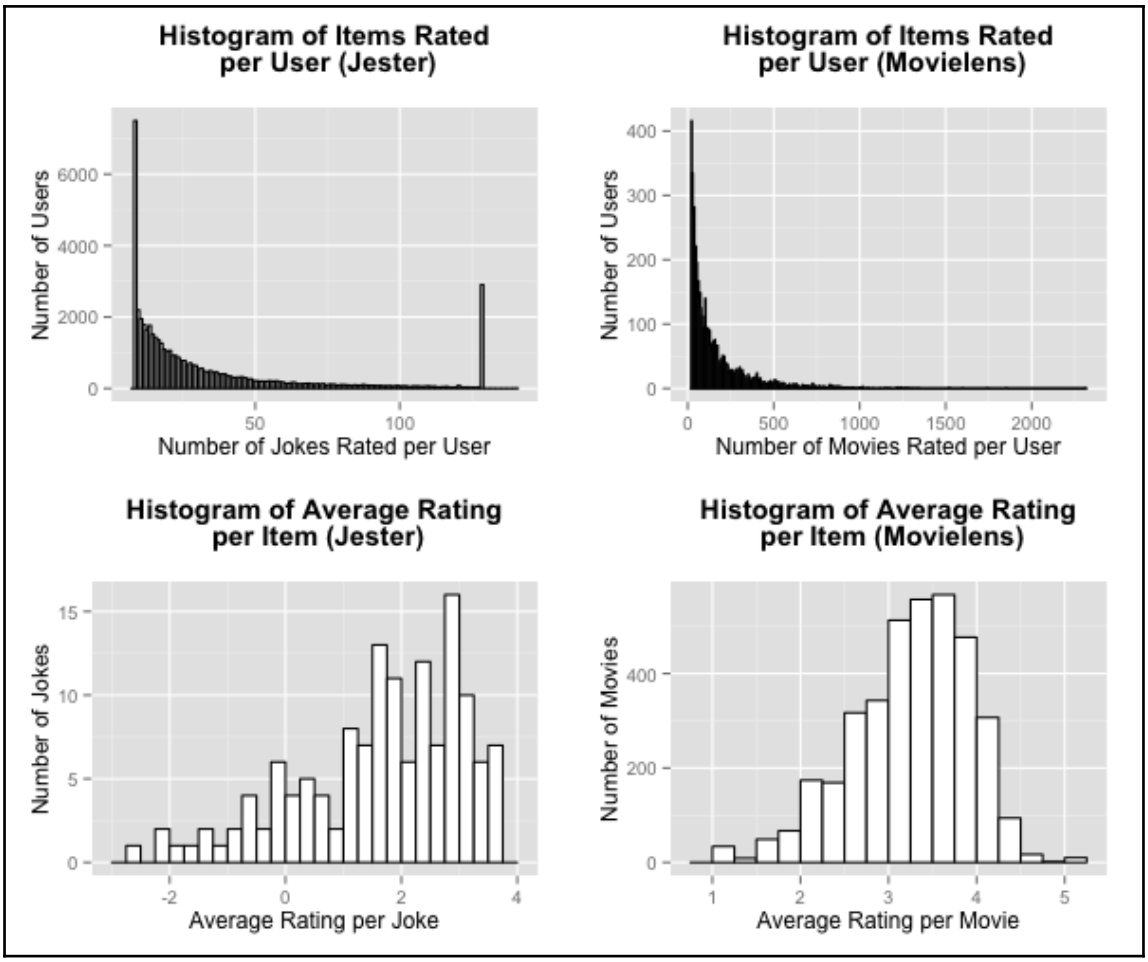
$$\hat{r}_{tj} = \frac{1}{\sum_{u \in N(t)} sim(u,t)} \sum_{u \in N(t)} sim(u,t) \cdot r_{uj}$$

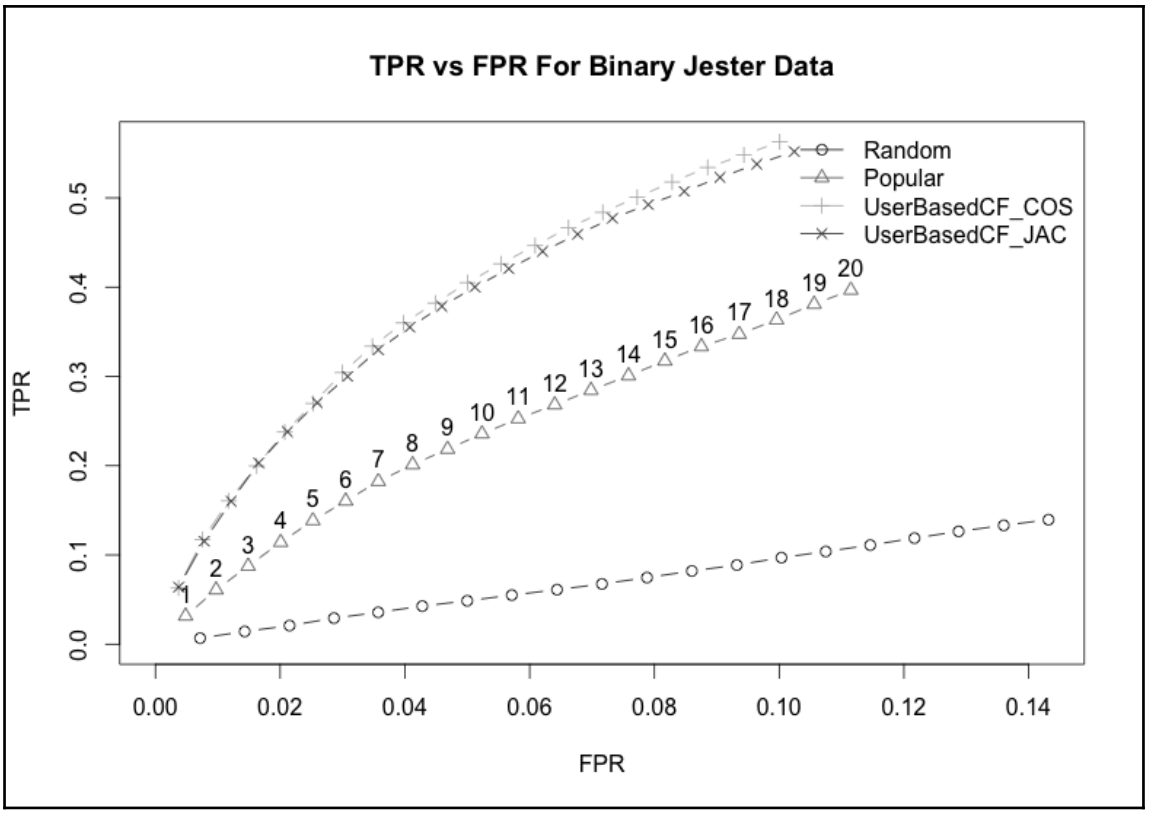
$$\hat{r}_{ti} = \frac{1}{\sum_{j \in S(i)} sim(i,j)} \sum_{j \in S(i)} sim(i,j) \cdot r_{tj}$$

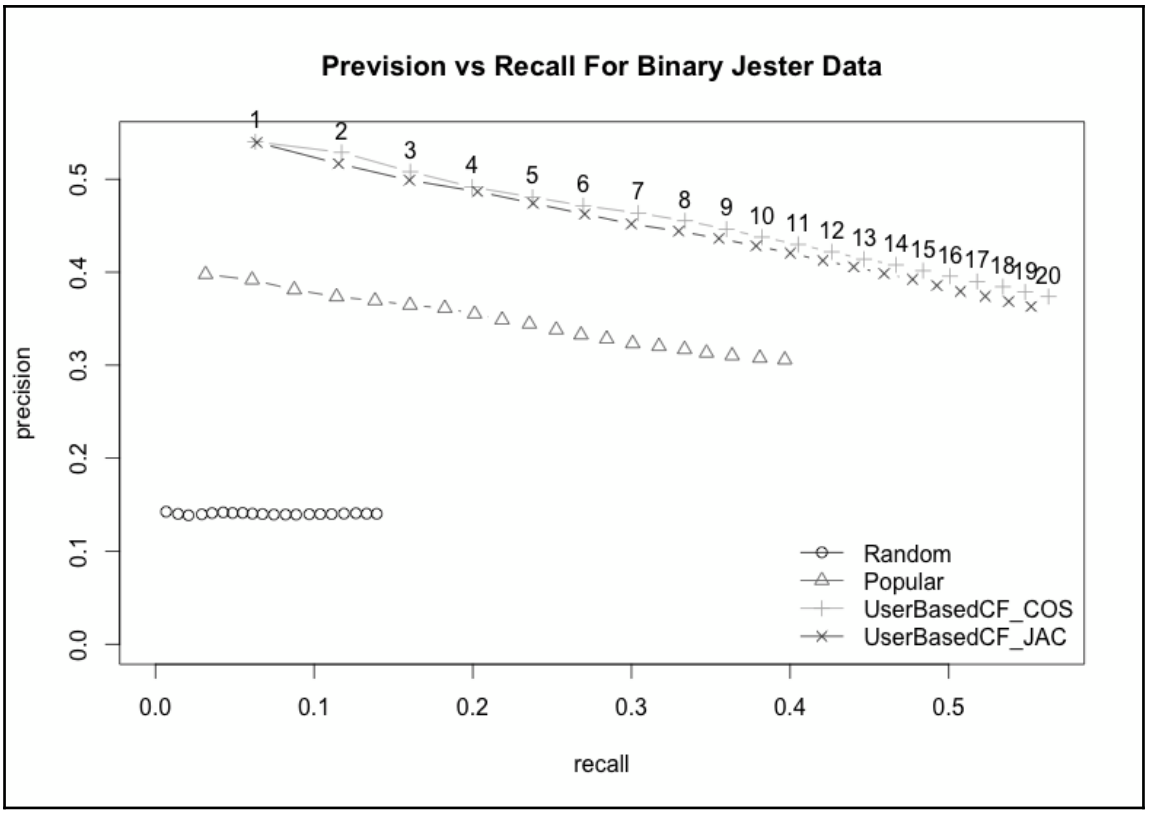
$$D_{m \times n} = U_{m \times n} \cdot \sum_{m \times n} \cdot V^{T}_{n \times n}$$

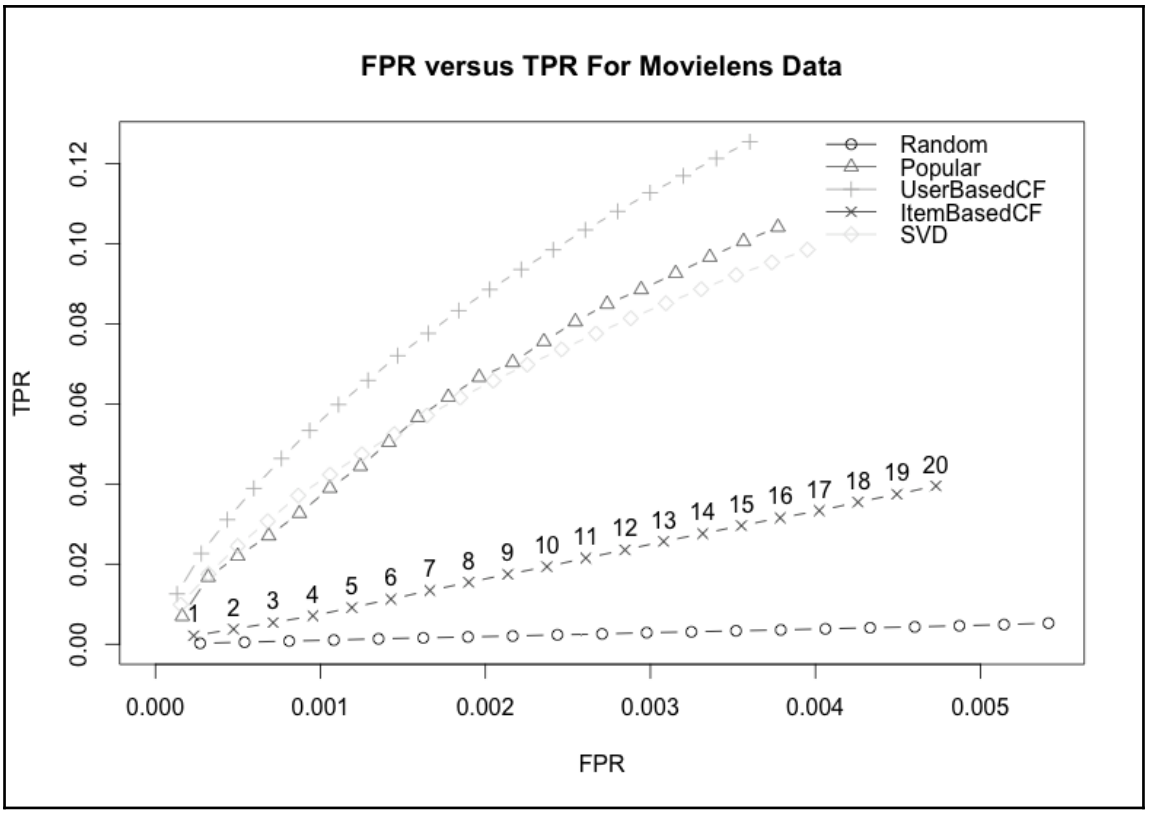
$$\hat{r}_{tj}$$

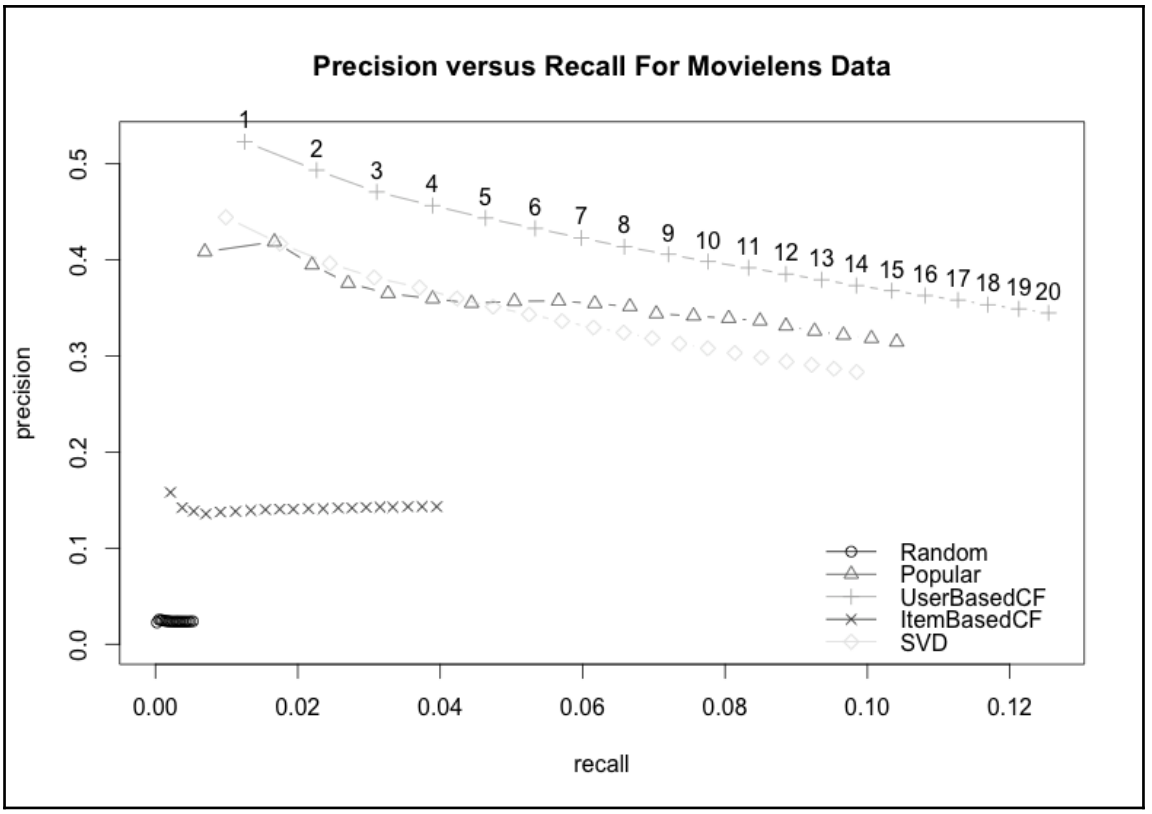

Histogram of Raw Ratings
(Jester)
Number of Jokes
40000
30000
20000
10000
0
-10
-5
0
5
10
Rating
Histogram of Raw Ratings
(Movielens)
Number of Movies
3e+05
2e+05
1e+05
0e+00
1
2
3
4
5
Rating
Histogram of Normalized Ratings
(Jester)
Number of Jokes
200000
150000
100000
50000
0
-10
-5
0
5
10
Rating
Histogram of Normalized Ratings
(Movielens)
Number of Movies
100000
75000
50000
25000
0
-5
0
5
Rating


Histogram of Items Rated per User (Jester)
Number of Users
6000
4000
2000
0
50
100
Number of Jokes Rated per User
Histogram of Items Rated per User (Movielens)
Number of Users
400
300
200
100
0
0
500
1000
1500
2000
Number of Movies Rated per User
Histogram of Average Rating per Item (Jester)
Number of Jokes
15
10
5
0
-2
0
2
4
Average Rating per Joke
Histogram of Average Rating per Item (Movielens)
Number of Movies
400
200
0
1
2
3
4
5
Average Rating per Movie

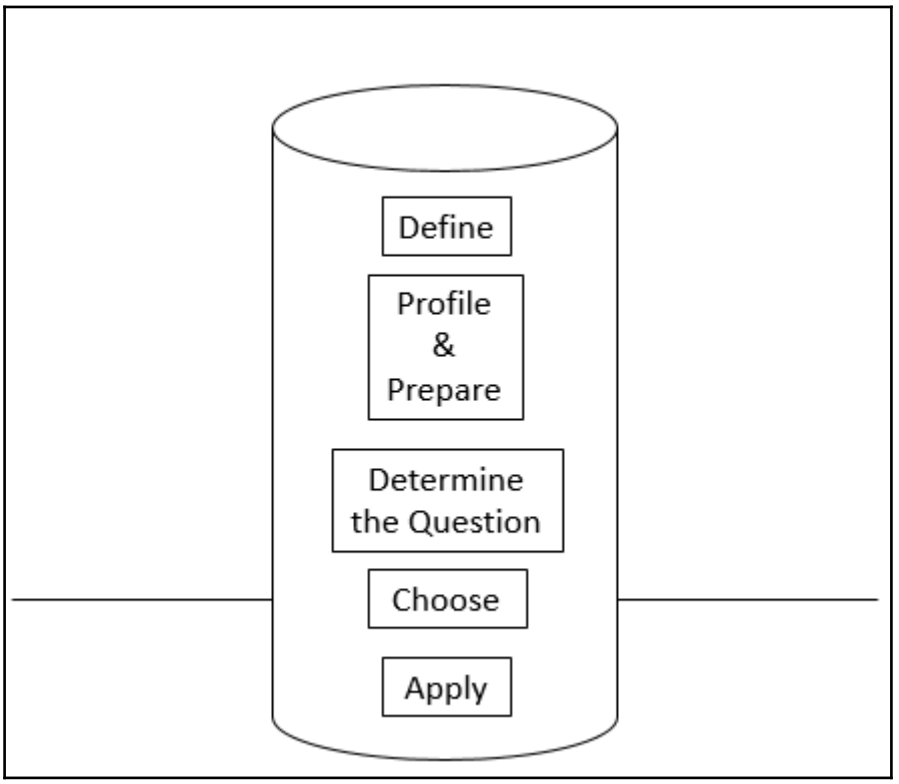

TPR vs FPR For Binary Jester Data
Random
Popular
UserBasedCF_COS
UserBasedCF_JAC
TPR
FPR
0.0
0.1
0.2
0.3
0.4
0.5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

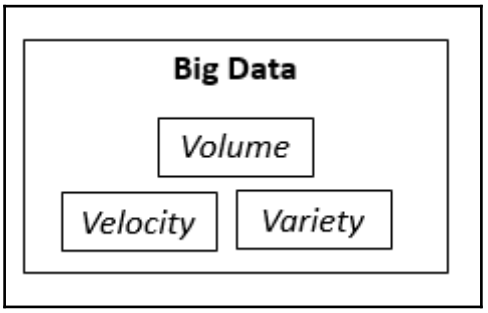

Prevision vs Recall For Binary Jester Data
precision
recall
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Random
Popular
UserBasedCF_COS
UserBasedCF_JAC


FPR versus TPR For Movielens Data
Random
Popular
UserBasedCF
ItemBasedCF
SVD
TPR
FPR
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.000
0.001
0.002
0.003
0.004
0.005
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


Precision versus Recall For Movielens Data
precision
recall
0.0
0.1
0.2
0.3
0.4
0.5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Random
Popular
UserBasedCF
ItemBasedCF
SVD

Chapter 13: Scaling Up





```
> summary(x[9])
          V9
 Mississippi: 12
 Utah       : 12
 Alabama    :  6
 Alaska     :  6
 Arkansas   :  6
 Delaware   :  6
 (Other)    :102
>
```

```
> summary(x[5])
       V5
 weight: 1
 170    :47
 177.00 :24
 65     :72
 72.00  : 6
>
```

```
> summary(x[79])
                    V79
 1_year_ago_weight: 1
 40                 :24
 56                 :30
 90                 :24
 161                :65
 195                : 6
>
```

```
> HCDF = read.csv("c:/Worker/HCSurvey20170202.txt")
> for (name in levels(HCDF$state)){
+   if (HCDF[5]>HCDF[79]+5) {
+   tmp=subset(HCDF,state==name)
+   fn=paste('c:/Worker/',gsub(' ','',name),sep='')
+   write.csv(tmp,fn,row.names=FALSE) }
+ }
```

```
> x<-read.csv("c:/Worker/HCSurvey20170202.txt")
> mysample <- x[sample(1:nrow(x), 500,
+    replace=FALSE),]
> nrow(x)
[1] 5994
> nrow(mysample)
[1] 500
>
```

```
> z <-aggregate(x, by=x["state"], FUN=mean, na.rm=TRUE)
There were 50 or more warnings (use warnings() to see the first 50)
> nrow(z)
[1] 50
> nrow(x)
[1] 5994
>
```

Chapter 14: Deep Learning

```
218316 250
</s> 0.001361 -0.000269 0.000991 -0.000002 -0.001748 0.001911 -0.001624 0.000845 -0.001084 -0.001371 0.001858 0.000468
-0.001717 0.001222 0.000752 0.000827 0.000778 -0.001539 -0.000694 -0.001299 -0.001736 -0.000549 -0.000933 -0.001127
0.001041 0.000547 0.001298 -0.001436 -0.000726 0.000316 -0.000075 0.000028 -0.001368 0.001804 0.000544 0.001903
-0.001856 -0.001760 0.001048 -0.001219 -0.000760 0.000704 -0.001256 -0.001917 -0.000637 -0.001063 0.000636 -0.001741
-0.001899 0.001606 -0.000662 0.000743 0.000344 0.001428 0.001738 0.001092 -0.000178 0.001399 -0.001932 -0.000540
-0.000081 0.000019 -0.000053 0.000389 0.000120 0.001389 0.001462 0.001230 -0.001047 -0.001650 0.001815 -0.001709
0.001795 -0.001571 0.001816 0.000776 -0.001863 -0.000841 0.000314 0.001601 0.000178 0.001402 -0.001247 -0.000623
0.001020 -0.000546 -0.000611 0.001851 0.000867 0.000183 -0.001398 0.000941 0.000471 -0.000126 -0.000823 -0.000859
0.000898 -0.001370 0.000296 -0.001508 -0.000944 -0.001548 0.000750 -0.000559 0.000066 -0.001170 -0.000592 0.000628
-0.001558 0.000858 -0.001044 0.000187 0.001062 0.000904 0.001245 -0.001303 -0.001731 -0.001764 -0.001953 0.000416
0.001849 -0.000813 0.000847 0.000360 -0.001535 0.001423 0.001300 -0.001805 -0.001197 0.001550 -0.001740 0.000734
0.001041 0.000660 0.001190 -0.001997 0.001445 -0.000100 -0.000155 0.000598 -0.000175 0.001805 0.000136 0.000794 0.000624
-0.001749 -0.001758 0.000384 0.001856 -0.001106 0.001825 -0.000541 -0.001234 0.000079 0.000765 0.001023 -0.000582
0.000896 -0.000143 -0.000640 -0.000479 -0.001032 0.000660 0.000960 0.001124 -0.001459 -0.000754 -0.001915 -0.000093
0.001137 -0.000085 0.000784 0.000111 0.000161 -0.001325 0.001055 0.000989 0.001289 -0.001148 0.000377 -0.001908
-0.000899 -0.000558 -0.000358 -0.001124 0.001694 0.001534 0.001914 -0.000715 0.001710 -0.000619 -0.001916 0.001301
0.001779 0.000963 -0.001936 -0.000376 0.001611 -0.001060 -0.001136 -0.000141 0.000810 -0.001208 0.001496 -0.001037
0.000351 0.001342 0.000003 -0.001435 0.000706 -0.001524 0.001329 0.001594 -0.001971 0.001831 -0.001693 -0.000114
-0.001384 0.000518 -0.000002 -0.000328 0.001973 -0.001251 0.001166 0.001675 0.000594 -0.000198 0.001470 0.001719
-0.001428 -0.001084 0.000456 0.001726 -0.001907 0.000304 -0.001583 -0.000408 0.000845 0.001401 0.001077 -0.001527
-0.001469 0.001607 -0.000253 -0.000810 -0.001671 0.000243 0.000228 0.001371 0.001966
the 0.021372 -0.111613 0.087509 0.014064 -0.017226 -0.023956 -0.095988 0.051796 0.023237 0.034806 0.007244 0.010209
0.014740 -0.000513 0.083205 0.050156 0.048451 -0.000807 0.129050 -0.143777 0.114502 -0.023673 0.013746 0.082953 0.011525
-0.034647 0.100390 -0.096474 0.048826 0.156398 -0.105746 -0.087686 -0.054978 -0.118068 0.080848 -0.019460 -0.071667
```