Chapter 24

Introduction to Data Mining

Big data typically involves huge *relational databases* called digital *data warehouses* which deal with millions if not billions of records. *Data mining* of the data warehouses are used to make predictions and decisions; *business analytics* describes the use of data to make business decisions.

24.1 The Big Data Revolution

Exercise 24.1 (The Big Data Revolution)

24.2 Direct Marketing

Exercise 24.2 (Direct Marketing)

24.3 The Goals of Data Mining

Exercise 24.3 (The Goals of Data Mining)

24.4 Direct Mining Myths

Exercise 24.4 (Direct Mining Myths)

24.5 Successful Data Mining

Exercise 24.5 (Successful Data Mining)

24.6 Data Mining Problems

There are two different types of data mining problems: the *un*supervised (learning) problem and the *supervised* (learning) problem. After clarifying the difference between these two problems, we will describe one new *un*supervised problem technique called *k-means clustering* and one new supervised problem technique called the *regression trees*.

Exercise 24.6 (Data Mining Problems)

1. *Unsupervised versus supervised (learning) problems.*
   Unsupervised learning and supervised learning are somewhat analogous to the relationship between correlation and regression, respectively. Correlation describes the (possibly linear) association between two variables whereas (possibly linear) regression describes how one variable *predicts* or describes another variable often called the response. Correlation **is** / **is not** involved in prediction whereas regression **is** / **is not** involved in prediction. Most of the methods we have looked at in this course are involved in predicting one variable from another and are **unsupervised** / **supervised** learning problems.

2. *Unsupervised (learning) problem: K-means clustering*
   K-means clustering involves dividing the observations into K clusters where the within-cluster variation (typically, and in the case given here, where variation is measured by the sum of squared Euclidean distances between observations) is as small as possible:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

   A computer algorithm first randomly assigns the observations into K clusters, K cluster *centroids* are calculated, the observations are reassigned *closest* to these centroids which forms new clusters, new K centroids are calculated and the process repeats until cluster assignments stop changing.

   (a) *Reading versus illumination, with levels of education*
   K-means clustering may suggest possible multiple linear regressions to use.

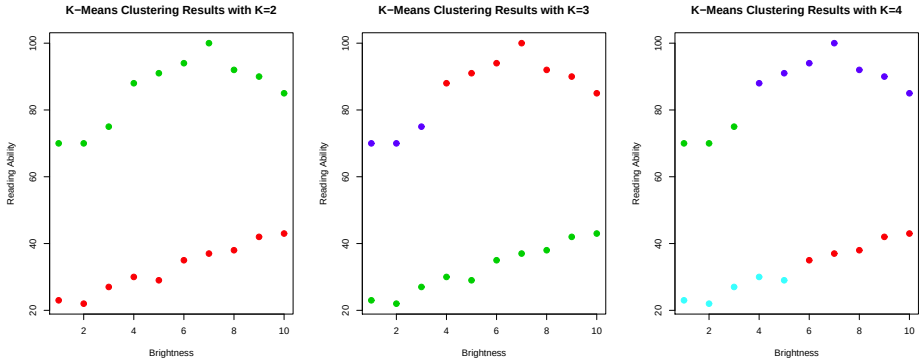

Figure 24.1: Scatterplots with various-means clustering

illumination, x_1	1	2	3	4	5	6	7	8	9	10
ability to read, y	23	22	27	30	29	35	37	38	42	43

illumination, x_1	1	2	3	4	5	6	7	8	9	10
ability to read, y	70	70	75	88	91	94	100	92	90	85

```
K-means clustering with 2 clusters of sizes 10, 10

Cluster means:
  brightness reading
1        5.5    85.5
2        5.5    32.6

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 1055.0  588.9
 (between_SS / total_SS =  89.5 %)
```

```
model.reading <- lm(reading~brightness+education); summary(model.reading)
predict(model,list(education=0,brightness=7))

data <- chapter18.reading.groups; attach(data.reading); head(data)
brightness <- data$brightness; reading <- data$reading; education <- data$education

par(mfrow=c(1,3))
data.new <- data[,-2]; data.new
km.out <- kmeans(data.new,2,nstart=20); km.out
plot(data.new, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=2",
     xlab="Brightness", ylab="Reading Ability", pch =20, cex =2)
km.out <- kmeans(data.new,3,nstart=20); km.out
plot(data.new, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=3",
     xlab="Brightness", ylab="Reading Ability", pch =20, cex =2)
km.out <- kmeans(data.new,4,nstart=20); km.out
plot(data.new, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=4",
     xlab="Brightness", ylab="Reading Ability", pch =20, cex =2)
par(mfrow=c(1,1))
```

Best number of clusters in Figure 24.1 occurs when $K = \mathbf{2}$ / $\mathbf{3}$ / $\mathbf{4}$
with sizes $\mathbf{(10, 10)}$ / $\mathbf{(7, 10, 3)}$ / $\mathbf{(5, 2, 7, 5)}$
with cluster means $\mathbf{(5.5, 32.6), (5.5, 85.5)}$ / $\mathbf{(5, 32.6), (5, 85.5)}$

where $\frac{\text{between SS}}{\text{total SS}} = \mathbf{89.5\%}$ / $\mathbf{90.2\%}$ / $\mathbf{91.3\%}$
meaning $100 - 89.5 = 10.5\%$ within SS, observations within clusters very close to one another
which researchers find out later represent high school and college students; consequently, introducing an education indicator x_2 to represent these two levels of education, we would now perform a multiple linear regression.

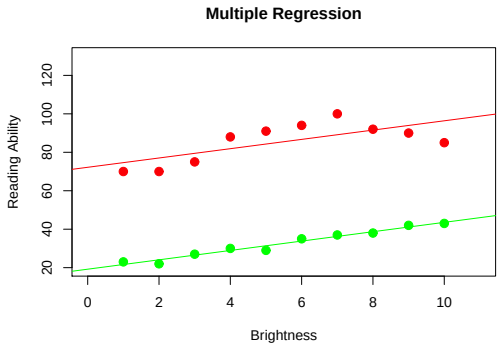

Figure 24.2: Comparing linear regression with various K-means clustering

illumination, x_1	1	2	3	4	5	6	7	8	9	10
education x_2	0	0	0	0	0	0	0	0	0	0
ability to read, y	23	22	27	30	29	35	37	38	42	43
illumination, x_1	1	2	3	4	5	6	7	8	9	10
education x_2	1	1	1	1	1	1	1	1	1	1
ability to read, y	70	70	75	88	91	94	100	92	90	85

```
data <- chapter18.reading.groups; attach(data.reading); head(data)
brightness <- data$brightness; reading <- data$reading; education <- data$education
x1 <- brightness[education=="0"]; y1 <- reading[education=="0"]
x2 <- brightness[education=="1"]; y2 <- reading[education=="1"]
plot(brightness, reading, xlim=c(0,11), ylim=c(20,130), xlab="Brightness", ylab="Reading Ability",
     main="Multiple Regression",type="n")
points(x1,y1,col="green",pch=20,cex=2); abline(lm(y1~x1),col="green")
points(x2,y2,col="red",pch=20,cex=2); abline(lm(y2~x2),col="red")
legend(topleft,c("1: college","0: high.school"),pch=c(16,16),col=c("red","green"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2500     2.9062   6.624 4.31e-06 ***
brightness    2.4273     0.4251   5.711 2.55e-05 ***
education    52.9000     2.4417  21.665 8.07e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.46 on 17 degrees of freedom
Multiple R-squared:  0.9672,        Adjusted R-squared:  0.9634
F-statistic:   251 on 2 and 17 DF,  p-value: 2.399e-13

model.reading <- lm(reading~brightness+education); summary(model.reading)
predict(model,list(education=0,brightness=7))
```

The multiple linear regression for high school and college scatterplot is
$\hat{\boldsymbol{y}} = \mathbf{19.2} + \mathbf{2.44}\boldsymbol{x_1}$
$\hat{\boldsymbol{y}} = \mathbf{72.2} + \mathbf{2.42}\boldsymbol{x_2}$
$\hat{\boldsymbol{y}} = \mathbf{19.25} + \mathbf{2.43}\boldsymbol{x_1} + \mathbf{52.9}\boldsymbol{x_2}$

One indicator, education x_2, is used
where $x_2 = 0$ represents **high school** / **college**
and $x_2 = 1$ represents **high school** / **college**

Overall model-fit statistics
$R^2 =$ **0.367** / **0.496** / **0.967** variation y explained by regression
$F =$ **7.88** / **234.7** / **251** with p-value = **0.03** / **0.06** / **0.08**
so to predict reading for high schoolers, $x_2 = 0$, at brightness $x_1 = 7$:

$$\hat{y} = 19.25 + 2.43(7) + 52.9(0) =$$

36.26 / **37** / **89.16** / **100**

(b) *Reading versus illumination*
K-means clustering may suggest interesting avenues for further research.

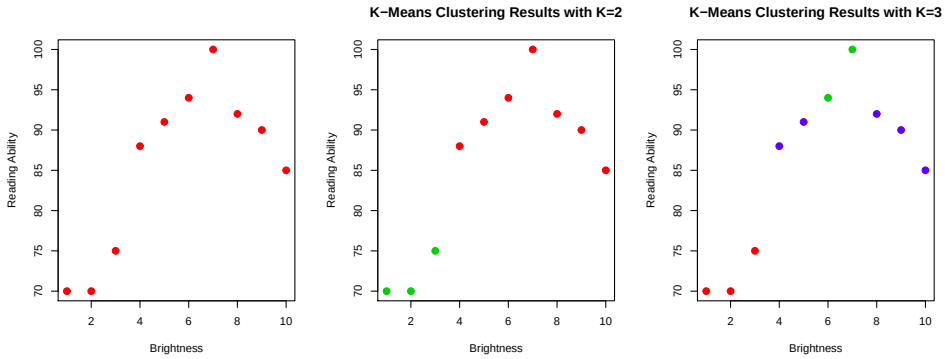

Figure 24.3: Scatterplots with 2-means and 3-means clustering

brightness, x	1	2	3	4	5	6	7	8	9	10
ability to read, y	70	70	75	88	91	94	100	92	90	85

```
data <- chapter4.reading.brightness; attach(data.reading); head(data.reading)
brightness <- data$brightness; reading <- data$reading

par(mfrow=c(1,3))
plot(brightness,reading,pch=20,cex=2,col="red",xlab="Brightness",ylab="Reading Ability")
km.out <- kmeans(data,2,nstart=20); km.out
plot(data, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=2", xlab="Brightness", ylab="Reading
km.out <- kmeans(data,3,nstart=20); km.out
plot(data, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=3", xlab="Brightness", ylab="Reading
par(mfrow=c(1,1))
```

It **is** / **is not** immediately clear how many clusters is best in Figure 24.3; however, the clustering does suggest possible avenues of exploration, to find out why the clustering occurs the way it does, possibly due to education level, age, gender and so on.
K-means clustering **is** / **is not** used for prediction: reading ability **is** / **is not** treated as a predictor variable in this method

3. *Supervised (learning) problem: regression trees*

   (a) *Book expenditure versus reading, education and income*
   Regression trees method involves dividing the predictors into J distinct non-overlapping regions, $R_1, \ldots, R_J$ that minimize the residual sum of squares (RSS):

$$\min_{R_1,\ldots,R_J} \left\{ \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \right\}$$

   Recursive binary splitting is used to construct $R_1, \ldots, R_J$, where each of the predictors are split into two regions and the split which minimizes the RSS is chosen as the first split and then a similar procedure occurs for each resulting split region until a stopping criterion is reached such as no region contains more than 4 observations.

reading ability, x_1	60	65	47	57	$\cdots$	54	89	64	79	75
education, x_2	14	17	9	15	$\cdots$	15	12	11	15	5
income, x_3	58	66	58	72	$\cdots$	41	36	41	54	25
book expenditure, y	78	105	100	76	$\cdots$	76	122	119	66	81

```
Regression tree:
tree(formula = book.expenditure ~ ., data = data)
Variables actually used in tree construction:
[1] "reading" "income"
Number of terminal nodes:  5
Residual mean deviance:  527.9 = 13720 / 26
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-40.710 -14.760   1.333   0.000  16.400  49.290
```

```
library(tree)
data <- chapter24.book.expenditure; attach(data); head(data)
book.expenditure <- data$book.expenditure; education <- data$education
reading <- data$reading; income <- data$income
tree.data <- tree(book.expenditure~., data)
summary (tree.data)
plot(tree.data)
text(tree.data, pretty=0)
```

Regression tree method *predicts* average book expenditure of
\$53.33 /\$66.40 / \$95.67 if reading ability < 49 units
\$53.33 /\$66.40 / \$72.71 if reading ability < 63 units and income $< \$62,000$
\$53.33 /\$66.40 / \$72.71 if reading ability < 63 units and income $> \$62,000$

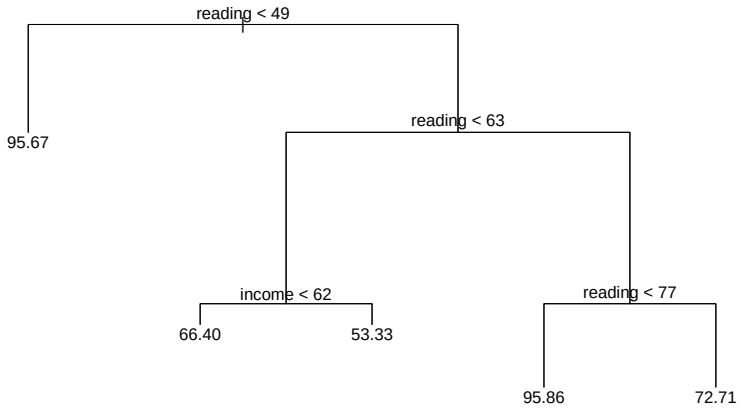

Figure 24.4: Regression Tree

\$53.33 /\$66.40 / \$95.86 if reading ability $>$ 63 units but also $<$ 77
\$53.33 /\$66.40 / \$72.71 if reading ability $>$ 77 units

One predictor not used in predicting average book expenditure:
reading ability / **education** / **income**

number of terminal nodes: **4** / **5** / **6**
residual mean deviance: **26** / **527.9** / **13720**

(b) *Book expenditure versus reading, education and income with minor change*
Same example, but first two (bolded) reading ability observations changed:

reading ability, x_1	**80**	**85**	47	57	$\cdots$	54	89	64	79	75
education, x_2	14	17	9	15	$\cdots$	15	12	11	15	5
income, x_3	58	66	58	72	$\cdots$	41	36	41	54	25
book expenditure, y	78	105	100	76	$\cdots$	76	122	119	66	81

```
Regression tree:
tree(formula = book.expenditure ~ ., data = data)
Variables actually used in tree construction:
[1] "reading" "income"
Number of terminal nodes:  5
Residual mean deviance:  537.1 = 13970 / 26
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-42.2500 -14.1200   0.4286   0.0000  13.1100  50.7500
```

```
library(tree)
data <- chapter24.book.expenditure2; attach(data); head(data)
```

```
book.expenditure <- data$book.expenditure; education <- data$education
reading <- data$reading; income <- data$income; coupons <- data$coupons; gender <- data$gender

tree.data <- tree(book.expenditure~., data)
summary (tree.data)
plot(tree.data)
text(tree.data, pretty=0)
```

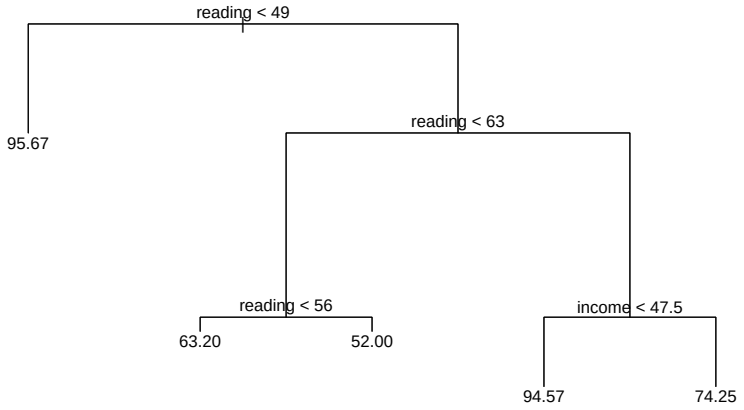

Figure 24.5: Regression Tree

Regression tree method *predicts* average book expenditure of
\$52.00 /\$63.20 / \$95.67 if reading ability < 49 units
\$52.00 /\$63.20 / \$72.71 if reading ability > 49 units and also < 56
\$52.00 /\$63.20 / \$74.25 if reading ability > 56 units and also < 63
\$53.33 /\$66.40 / \$94.57 if reading ability > 63 units and income $< \$47,500$
\$53.33 /\$66.40 / \$74.25 if reading ability > 63 units and income $> \$47,500$

One predictor not used in predicting average book expenditure:
reading ability / **education** / **income**

number of terminal nodes: **4** / **5** / **6**
residual mean deviance: **26** / **537.1** / **13970**

(c) *Advantages and disadvantages of regression trees*
advantage / **disadvantage**
easier to understand than linear regression
advantage / **disadvantage**
can be displayed graphically
advantage / **disadvantage**

not as predictively accurate as linear regression
advantage / disadvantage
non-robust: small change in data causes big change in estimated tree

24.7 Data Mining Algorithms

Exercise 24.7 (Data Mining Algorithms)

24.8 The Data Mining Process

Exercise 24.8 (The Data Mining Process)

24.9 Summary

Exercise 24.9 (Summary)