Business Data Mining Homework Session 1

2.1 Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

a. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).

b. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.

c. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.

d. Identifying segments of similar customers.

e. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.

f. Estimating the repair time required for an aircraft based on a trouble ticket.

g. Automated sorting of mail by zip code scanning.

h. Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.

2.6 In fitting a model to classify prospects as purchasers or nonpurchasers, a certain company drew the training data from internal data that include demographic and purchase information. Future data to be classified will be lists purchased from other sources, with demographic (but not purchase) data included. It was found that "refund issued" was a useful predictor in the training data. Why is this not an appropriate variable to include in the model?

2.8 Normalize the data in Table 2.17, showing calculations.

TABLE 2.17

Age	Income ($)
25	49,000
56	156,000
65	99,000
32	192,000
41	39,000
49	57,000

2.9 Statistical distance between records can be measured in several ways. Consider Euclidean distance, measured as the square root of the sum of the squared differences. For the first two records in Table 2.17, it is

$$\sqrt{(25-56)^2+(49,000-156,000)^2}.$$

Can normalizing the data change which two records are farthest from each other in terms of Euclidean distance?

4.4 Chemical Features of Wine. Table 4.13 shows the PCA output on data (non-normalized) in which the variables represent chemical characteristics of wine, and each case is a different wine.

TABLE 4.13 PRINCIPAL COMPONENTS OF NON-NORMALIZED WINE DATA

code for running PCA on the wine data

```
wine.df <- read.csv("Wine.csv")
pcs.cor <- prcomp(wine.df[,-1])
summary(pcs.cor)
pcs.cor$rot[,1:4]
```

Output

```
> summary(pcs.cor)

importance of components:
                            PC1      PC2     PC3     PC4     PC5
Standard deviation     314.9632 13.13527 3.07215 2.23409 1.10853
Proportion of Variance   0.9981  0.00174 0.00009 0.00005 0.00001
Cumulative Proportion    0.9981  0.99983 0.99992 0.99997 0.99998
                           PC6    PC7    PC8    PC9   PC10
Standard deviation     0.91710 0.5282 0.3891 0.3348 0.2678
Proportion of Variance 0.00001 0.0000 0.0000 0.0000 0.0000
Cumulative Proportion  0.99999 1.0000 1.0000 1.0000 1.0000
                          PC11   PC12    PC13
Standard deviation     0.1938 0.1452 0.09057
Proportion of Variance 0.0000 0.0000 0.00000
Cumulative Proportion  1.0000 1.0000 1.00000

> pcs.cor$rot[,1:4]

                              PC1           PC2           PC3           PC4
Alcohol              -0.0016592647 -1.203406e-03 -0.016873809  0.141446778
Malic_Acid            0.0006810156 -2.154982e-03 -0.122003373  0.160389543
Ash                  -0.0001949057 -4.593693e-03 -0.051987430 -0.009772810
Ash_Alcalinity        0.0046713006 -2.645039e-02 -0.938593003 -0.330965260
Magnesium            -0.0178680075 -9.993442e-01  0.029780248 -0.005393756
Total_Phenols        -0.0009898297 -8.779622e-04  0.040484644 -0.074584656
Flavanoids           -0.0015672883  5.185073e-05  0.085443339 -0.169086724
Nonflavanoid_Phenols  0.0001230867  1.354479e-03 -0.013510780  0.010805561
Proanthocyanins      -0.0006006078 -5.004400e-03  0.024659382 -0.050120952
Color_Intensity      -0.0023271432 -1.510035e-02 -0.291398464  0.878893693
Hue                  -0.0001713800  7.626731e-04  0.025977662 -0.060034945
OD280_OD315          -0.0007049316  3.495364e-03  0.070323969 -0.178200254
Proline              -0.9998229365  1.777381e-02 -0.004528682 -0.003112916
```

a. The data are in the file *Wine.csv*. Consider the rows labeled "Proportion of Variance." Explain why the value for PC1 is so much greater than that of any other column.

b. Comment on the use of normalization (standardization) in part *(a)*.