Problem_Set_2

Pierpaolo De Blasi

deadline 2023-05-16 11.59pm

Exercise 1

The data set `psych` contains 24 psychological tests administered to 301 students (with ages ranging from 11 to 16) in a suburb of Chicago: a group of 156 students (74 boys, 82 girls) from the Pasteur School and a group of 145 students (72 boys, 73 girls) from the Grant-White School.

```
psych<-read.table("data/psych.txt",header=T)
dim(psych)
```

```
## [1] 301  28
```

```
head(psych)
```

```
##   Case Sex  Age V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
## 1    1   M 13.1 20 31 12  3 40  7 23 22  9  78  74 115 229 170  86  96   6   9
## 2    2   F 13.6 32 21 12 17 34  5 12 22  9  87  84 125 285 184  85 100  12  12
## 3    3   F 13.1 27 21 12 15 20  3  7 12  3  75  49  78 159 170  85  95   1   5
## 4    4   M 13.2 32 31 16 24 42  8 18 21 17  69  65 106 175 181  80  91   5   3
## 5    5   F 12.2 29 19 12  7 37  8 16 25 18  85  63 126 213 187  99 104  15  14
## 6    6   F 14.1 32 20 11 18 31  3 12 25  6 100  92 133 270 164  84 104   6   6
##   V19 V20 V21 V22 V23 V24   group
## 1  16   3  14  34   5  24 PASTEUR
## 2  10  -3  13  21   1  12 PASTEUR
## 3   6  -3   9  18   7  20 PASTEUR
## 4  10  -2  10  22   6  19 PASTEUR
## 5  14  29  15  19   4  20 PASTEUR
## 6  14   9   2  16  10  22 PASTEUR
```

`Sex` is a factor with levels `F` and `M`; `Age` is a numeric vector; `group` is a factor with levels `GRANT` and `PASTEUR`. The 24 psychological test scores are named `V1` to `V24`, see script file `problemset2.R` for further information.

1. Use the Grant–White students data. Obtain the maximum likelihood solution for $m = 5$ and $m = 6$ factors and compute the proportion of total sample variance due to each factor. List the specific variances, and assess the accuracy of the approximation of the correlation matrix. Compare the results. Which choice of m do you prefer? Why?
2. Give an interpretation to the common factors in the $m = 5$ solution with varimax rotation.
3. Make a scatterplot of the first two factor scores for the $m = 5$ solution obtained by the regression method. Is their correlation equal to zero? Should we expect so? Comment.
4. Obtain the maximum likelihood solution with marimax rotation for $m = 5$ factors by using the Pasteur students data. Is the interpretation to the common factors similar to that of Grant–White students?
5. Make a scatterplot of the first two factor scores from the rotated MLFA solution for each school. Comment.

Exercise 2

The `pendigits` data set was created by collecting 250 samples from 44 writers. These writers were asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution. The raw data on each of $n = 10992$ handwritten digits consisted of a sequence, (x_t, y_t), $t = 1, 2, ..., T$, of tablet coordinates of the pen at fixed time intervals of 100 milliseconds, where x_t and y_t were integers in the range 0-500. These data were then normalized to make the representations invariant to translation and scale distortions. The new coordinates were such that the coordinate that had the maximum range varied between 0 and 100. Usually x_t stays in this range, because most integers are taller than they are wide. Finally, from the normalized trajectory of each handwritten digit, 8 regularly spaced measurements, (x_t, y_t), were chosen by spatial resampling, which gave a total of $p = 16$ variables. The data includes a class attribute, column `digit`, coded $0, 1, \ldots, 9$, about the actual digit.

```
pendigits<-read.table("data/pendigits.txt", sep=",",head=F)
names(pendigits)<-c(paste0(rep(c("x","y"),8),rep(1:8,each=2)),"digit")
dim(pendigits)
```

```
## [1] 10992    17
```

```
head(pendigits)
```

```
##    x1  y1 x2  y2  x3  y3  x4  y4 x5 y5  x6 y6  x7 y7  x8 y8 digit
## 1  47 100 27  81  57  37  26   0  0 23  56 53 100 90  40 98     8
## 2   0  89 27 100  42  75  29  45 15 15  37  0  69  2 100  6     2
## 3   0  57 31  68  72  90 100 100 76 75  50 51  28 25  16  0     1
## 4   0 100  7  92   5  68  19  45 86 34 100 45  74 23  67  0     4
## 5   0  67 49  83 100 100  81  80 60 60  40 40  33 20  47  0     1
## 6 100 100 88  99  49  74  17  47  0 16  37  0  73 16  20 20     6
```

1. Use linear discriminant analysis (LDA). Display the first two LD variables in a scatterplot, color coding the observations according to variable `digit.col` below. How well do they discriminate the 10 digits? Refer also to theory.

```
lookup<-c("darkgreen",  "brown", "lightblue",  "magenta", "purple",
          "blue", "red", "lightgreen", "orange", "cyan")
names(lookup)<-as.character(0:9)
digit.col<-lookup[as.character(pendigits$digit)]
```

2. Compute the confusion matrix on the training data. What are the groups *more difficult* to discriminate from the others? Comment in view of the answer to point 1.

3. Use leave-one-out cross validation (CV). Compute the confusion matrix and the corresponding CV error. Is it larger than the training error? Why so?

4. Compute the 44-fold cross validation error for each reduced-rank LDA classifier, including full-rank LDA, by using the partition of the observations provided by the variable `groupCV` below. Plot the error curve against the number of discriminant variables. What classifier do you prefer? Comment.

```
groupCV<-rep(1:44, each=250)
groupCV<-groupCV[1:length(pendigits$digit)]
```

5. (*Optional*) Find a classification rule that improves on the CV error rate estimates found before. Feel free to use any classification method, even one not covered in class.