Tutorial 12:

Introduction to Data Science

Unsupervised Learning

Ernst C. Wit

December 9, 2024

Exercises

1. **PCA: Principal Component Analysis**. We return to a dataset we encountered a few weeks ago. The Spanish Armada was a Spanish fleet of 130 ships that sailed from La Coruna in August 1588. During its preparation, several accounts of its formidable strength were circulated to reassure allied powers of Spain or to intimidate its enemies. One such account was given by Paz Salas et Alvarez (1588). The intent was bring the forces of Spain to invade England, overthrow Queen Elizabeth I, and reestablish Spanish control of the Netherlands. However the Armada was not as fortunate as hoped: it was all destroyed in one week's fighting.

   De Falguerolles (2008) reports the table given here as Armada as an early example of data to which multivariate methods might be applied. In fact, the aim is to discover the dimensionality of the information contained in the data. The data are available in the R-package `histData` as the object `Armada`. To load the data, type the following:

   ```
   > install.packages("histData")
   > library(histData)
   > ? Armada
   ```

   (a) **Information.** When you consider the various variables, can you imagine that some quantitative variables may contain almost the same information?

   (b) **Visualization.** Make a `pairs` plot of all the quantitative variables and revisit the question in (a).

   (c) **PCA analysis.** Use the function `prcomp` to fit a principal component analyis to the quantitative data.

      i. What is the variance explained by each of the principal components.

      ii. Consider the first two principal components and interpret their rotations.

      iii. Use the function `biplot` to replot the data onto the first two principal components and interpret the output.

      iv. What was the variance explained of the smalled principal component. Can you explain this?

2. **K-means clustering.** In this exercise we consider written characters and try to cluster them. Consider the file `digit10x10.csv` on iCorsi. It contains 366 images of 10 by 10 images of various characters.

   (a) Download the data and load the data in R as follows,

   ```
   dat<-read.csv("digit10x10.csv")
   dat<-as.matrix(dat)
   ```

   (b) Each row of the data contains the pixel values of one 10x10 image. You can visualize them as follows:

   ```
   image(z=matrix(dat[210,],ncol=10))
   ```

   You can also visualize 100 of these images at the same time, as follows,

   ```
   par(mfrow=c(10,10))
   apply(dat[101:200,],1,function(x){image(matrix(x,ncol=10))})
   ```

   Can you tell was these images are? Do you notice any missing character.

   (c) Use the function `kmeans` to perform k-means clustering. You have to provide the data and a number of clusters.

   (d) The output also provides the cluster label. How well did it identify the various groups of digits?