Data Analysis in IC4.0 2021/2022

Final exam

---

1. (4 points) Explain whether each scenario is a classification, regression, clustering or dimensionality reduction problem. Identify the number of variables and the number of observation and sketch how you would address the data analysis in each case in terms of the methodology used and the assessment of the results.

(i) (1 point) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

(ii) (1 point) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(iii) (1 point) We are interested in obtaining an automatic procedure which recognizes plants from not high resolution photos (30×30 pixels) in a grayscale. For a grayscale images, the pixel value is a single number that represents the brightness of the pixel. To do so, a dataset containing 5342 images of flowers is collected. For these photos, five type of flowers have been identified: chamomile, tulip, rose, sunflower and dandelion.

(iv) (1 point) Assuming that the information about the flower types is not known for the dataset in (iii), our aim is to find groups of photos which resemble each other.

(v) (1 point) For the dataset in (iv), we are interested in obtaining (even) lower resolution images.

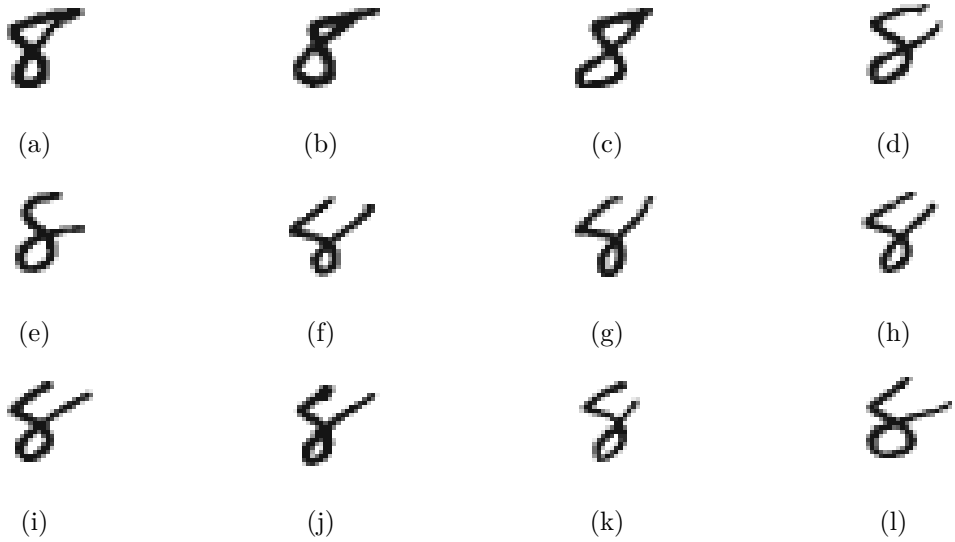
2. (1 point) Suppose that we take a dataset, divide it into equally-sized training and test sets, and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors ($K = 1$) and get an average error rate (average over both test and training datasets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

3. (3 points) We perform best subset, forward stepwise, and backward stepwise selection on a single dataset. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. True or False (**justify your answer**) assuming $k \in \{0, 1, 2, \ldots, p\}$:

(i) (1 point) The predictors in the $k-$variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)-$variable model identified by forward stepwise selection.

(ii) (1 point) The predictors in the $k-$variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)-$variable model identified by backward stepwise selection.

(iii) (1 point) The predictors in the $k-$variable model identified by best subset are a subset of the predictors in the $(k + 1)-$variable model identified by best subset selection.



(a) (b) (c) (d)

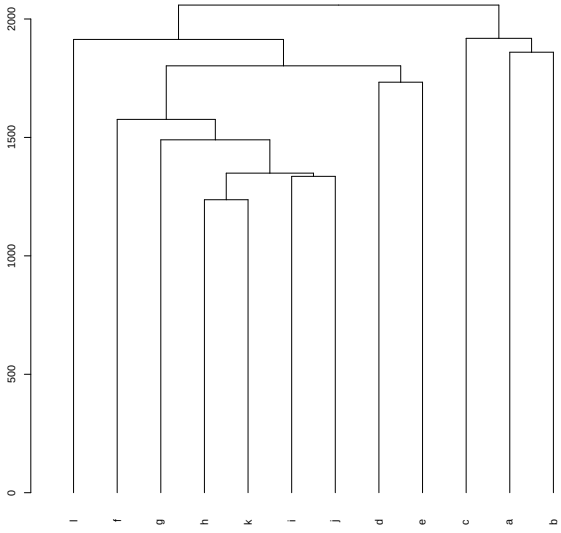
(e) (f) (g) (h)

(i) (j) (k) (l)

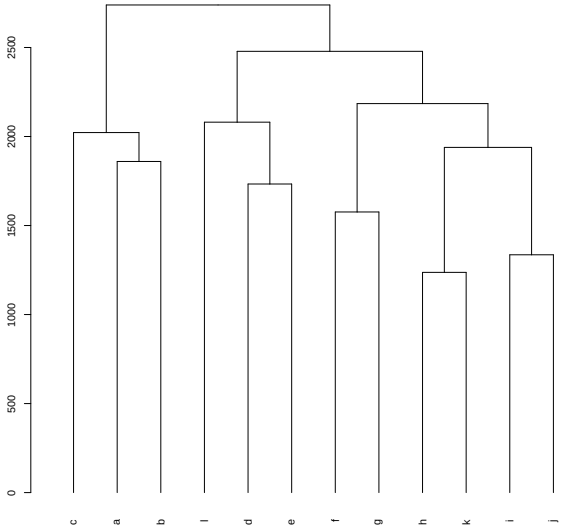
4. (2 points) The following 12 images belong to a dataset which records handwritten ZIP codes on envelopes from U.S. postal mail. Each image is a segment from a five digit ZIP code, isolating a single digit. The images are 16×16 grayscale maps, with each pixel ranging in intensity from 0 to 255. All the images in this sample correspond to the number 8.
With the aim of finding handwritten 8s that look similar, a hierarchical clustering approach has been carried out using the Single, Complete and Average linkages with the Manhattan distance. The dendrograms in the following page have been obtained.

(i) (1 point) Briefly comment the differences between the Single, Complete and Average linkage approaches in hierarchical clustering.

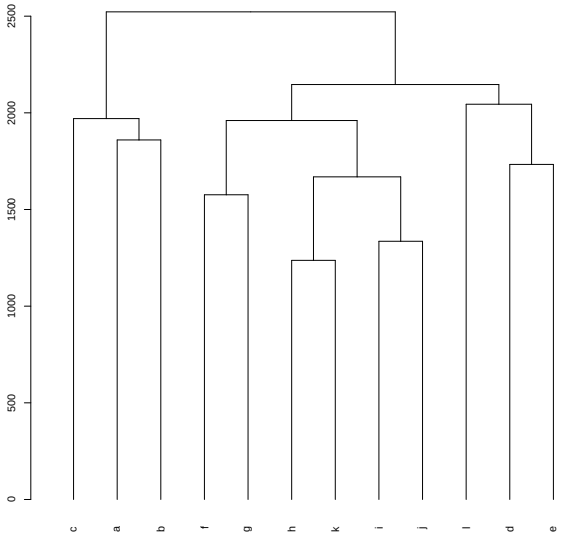
(ii) (1 point) Discuss the different results obtained for the dataset described above and decide which of the three choices better identifies the handwritten 8s that look similar. Based on that choice, how many groups of 8s would you consider?



Single Linkage



Complete Linkage



Average Linkage