Stat 441/505 Winter 2019
Assignment #3
Due date: April 10, 2019

Q1. ##----------##

Please refer to the textbook pages 395 and 396, or Lecture notes 22.

(a) Write down the forward and backward propagation equations for the sum-of-square error loss function and describe the back-propagation algorithm. Note, you shall be able to find all the necessary equations in the textbook or lecture notes.

(b) Derive the forward and backward propagation equations for the cross-entropy loss function and describe the back-propagation algorithm.

Q2. ##----------##

Generate 100 observations from the following model

$$Y = \sigma(a_1^T X) + (a_2^T X)^2 + 0.30 \cdot Z,$$

where σ is the sigmoid function, Z is standard normal, $X^T = (X_1, X_2)$, each X_j being independent standard normal, and $a_1 = (3, 3), a_2 = (3, -3)$.

(a) Using a single layer perceptron neural network to fit the data with varying number of hidden units in the network, from 1 to 10, and determine the minimum number needed to perform well for this task. There are two methods can be used for this task: one is the K-fold cross validation, choosing K =5 in this case; the other one is the validation method, where 1,000 new observations can be generated to test the fitted model as we know the ground truth.

(b) Use simple deep learning method on the validation method and compare it with the optimal results in term of testing error in (a).

Q3. ##----------##

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}.$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) Repeat (a), this time using single linkage clustering.

(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?

(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

Q4. ##----------##

In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations.

(b) Randomly assign a cluster label to each observation. You can use the `sample()` command in `R` to do this. Report the cluster labels for each observation.

(c) Compute the centroid for each cluster.

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) In your plot from (a), color the observations according to the cluster labels obtained.

Q5. ##----------##

Suppose that for a particular data set, we perform hierarchical clustering using single linkage and using complete linkage. We obtain two dendrograms.

(a) At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ fuse. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

(b) At a certain point on the single linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ fuse. On the complete linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

Q6. ##----------##

Consider the `USArrests` data. We will now perform hierarchical clustering on the states.

10.9

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, *after scaling the variables to have standard deviation one.*

(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Q7. ##----------##

In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

10.10

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

Hint: There are a number of functions in `R` *that you can use to generate data. One example is the* `rnorm()` *function;* `runif()` *is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.*

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

(c) Perform K-means clustering of the observations with $K = 3$. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

Hint: You can use the `table()` *function in* `R` *to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.*

(d) Perform K-means clustering with $K = 2$. Describe your results.

(e) Now perform K-means clustering with $K = 4$, and describe your results.

(f) Now perform K-means clustering with $K = 3$ on the first two principal component score vectors, rather than on the raw data. That is, perform K-means clustering on the 60×2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

(g) Using the `scale()` function, perform K-means clustering with $K = 3$ on the data *after scaling each variable to have standard deviation one*. How do these results compare to those obtained in (b)? Explain.

Q8. ##----------##

A firm is attempting to evaluate the quality of its sales staff and is trying to find an examination or series of tests that may reveal the potential for good performance in sales. The firm has selected a random sample of 50 sales people and has evaluated each on 3 measures of performance: growth of sales, profitability of sales, and new-account sales. These measures have been converted to a scale, on which 100 indicates 'average' performance. Each of the 50 individuals took each of 4 tests, which purported to measure creativity, mechanical reasoning, abstract reasoning, and mathematical ability, respectively. The $n = 50$ observations on $p = 7$ variables are listed in T9-12.DAT.

(a) Assume an orthogonal factor model for the standardized variables. Obtain the (unrotated) principal component solution and the (unrotated) maximum likelihood solution for $m = 2$ and $m = 3$ common factors. Given these four solutions, obtain the rotated solutions. For all eight of these fits, list (i) the proportions of variance accounted for by each of the m factors, (ii) estimated communalities, (iii) specific variances, and (iv) the sum of squares of the residuals for the $m = 2$ and $m = 3$ solutions. When I run your R program I should see eight pieces of output displayed, each similar to the following.

```
rotated pc solution, m = 2:
prop.  variance = 0.45 0.41
communalities are 0.96 0.89 0.89 0.85 0.69 0.81 0.87
specific variances are 0.04 0.11 0.11 0.15 0.31 0.19 0.13
ss.resids = 0.24
```

(b) Choose what you feel is the 'best' of the eight fits (explain why), and interpret the factor solutions in this model. (As is almost always the case, there is no one 'right' answer here.)

(c) Suppose a new salesperson, selected at random, obtains the (unstandardized) test scores $\mathbf{x}^T = (110, 98, 105, 15, 18, 12, 35)$. Calculate the salesperson's factor scores using the rotated ml model with $m = 2$, together with each of the weighted least squares and regression methods. (As a check, you might let $\mathbf{x}$ be the first row of the original data matrix, and check that in this case your answer agrees with R's 'scores' output.)

Q9. ##----------##

The Database of Faces contains a set of face images taken between April 1992 and April 1994 at the AT&T Laboratories Cambridge. There are ten different images of each of 40 distinct subjects. The files are in PGM format, which can be read by `read.pnm` at the R package `pixmap` The images are organized in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the form Y.pgm, where Y is the image number for that subject (between 1 and 10). We will focus on the 1.pgm image for each subject.

(a) Read the face images using `read.pnm` and extract the data. Plot the faces of subject 1, 11, 21, and 31. You should make sure the faces are heading up. You may use `image` of `image.plot` in package `fields`.

(b) Find the mean and standard deviation, lower limit and upper limit faces and plot them. Comment on them. You may want to use `image.plot` in package `fields`.

(c) Find the first $m = 4$ factor faces using principal component analysis method, plot them and comment.

(d) Find the first $m = 4$ rotated factor faces (by varimax) using principal component analysis method, plot them and comment.

(e) Can we use maximum likelihood method to do the factor analysis in this case? Why?