Assignment 6 (3 pages)

Statistics 32950-24620 (Spring 2024)

Due date: 9 am Tuesday, May 7.

Requirements: Same as before. Make sure to submit to the correct section 246Pset6 or 329Pset6 in Gradescope.

References: 12.1 - 12.5 and 5.7 in Johnson & Wichern, 14.1 - 14.3, 6.8, and 8.5 in Hastie, Tibshirani and Friedman. Chapter 9 in Bishop.

Problem assignments:

1. (*Hierarchical clustering for small data*)

   The vocabulary "richness" of a text can be quantitatively described by counting the words used once, the wards used twice, and so forth. Base on these counts, a linguist proposed the following distances between chapters of an ancient book (the Old Testament book Lamentations):

$$\begin{array}{c} \\ Ch1 \\ Ch2 \\ Ch3 \\ Ch4 \\ Ch5 \end{array} \begin{array}{c} \begin{array}{ccccc} Ch1 & Ch2 & Ch3 & Ch4 & Ch5 \end{array} \\ \left[\begin{array}{ccccc} 0 & & & & \\ 0.76 & 0 & & & \\ 2.97 & 0.80 & 0 & & \\ 4.88 & 4.17 & 0.21 & 0 & \\ 3.86 & 1.96 & 1.51 & 0.51 & 0 \end{array}\right] \end{array}$$

   Cluster the chapters of the book using the three linkage hierarchical methods we have discussed: Single linkage, Complete linkage, and Average linkage. Compare the dendrograms (do not hand in the plots). You may use R.

   (a) Are the results from three methods similar? Are the vertical scales the same?

   (b) If we decide to have $k = 3$ clusters, list the three clusters produced by the three linkage methods.

   (c) Compare and comment. What's your takeaway?

2. (*K-means for small data*)

   Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are

Item	x_1	x_2
A	5	-4
B	1	-2
C	-1	1
D	3	1

   Use the K-means clustering to divide the items into K = 2 clusters.

   (a) Start with the initial groups (AB) and (CD). Give the final clusters, cluster centroids, and squared distances to cluster centroids for each points.

   (b) Use the `kmeans` function in R (which tries several initializations). Give the final clusters, cluster centroids, and squared distances to cluster centroids for each points.

3. (*Hierarchical and K-mean cluster analysis*)

   The national track records data for women are used in previous assignments:
   ladyrun24.dat (automatic download when clicked, also available next to the link of this p-set in Canvas).

```
ladyrun = read.table("ladyrun22.dat")
colnames(ladyrun)=c("Country","100m","200m","400m","800m","1500m","3000m","Marathon")
```

(a) Use the data to calculate the Euclidean distances between pairs of countries. Do not print the output. Give the two counties (translate into true names) with the maximum distance and the two countries with the minimum distance, along with the distance values. (R command: The `which` function in R can be useful.)
(Note: the variables are in comparable scales; you may either use the original data or the normalized data.)

(b) Treating the distances in (a) as measures of dissimilarity, cluster the countries using the single linkage and complete linkage hierarchical procedures. Construct dendrograms and compare the results. When $k = 8$ (or 7), provide the three smallest clusters (using country abbreviations) according to the complete linkage.

(c) Input the data into a K-means clustering program. Cluster the countries into groups using several values of K. Compare the results with those in Part (b). Which K would you choose?
Plot your K clusters (by colors or symbols) on the plane with the first two principal components as axes.

4. (*Two component mixture model for small data*)

Consider observations from a mixture of two densities $f(x) = p_1 f_1(x) + p_2 f_2(x)$, where $p_1 + p_2 = 1$,

$$f_1(x) = \begin{cases} 2x, & 0 \le x \le 1, \\ 0, & otherwise, \end{cases} \qquad f_2(x) = \begin{cases} 2(1-x), & 0 \le x \le 1, \\ 0, & otherwise. \end{cases}$$

(a) If the observed data are $\{x_1, x_2, x_3, x_4, x_5\} = \{0.1, 0.2, 0.3, 0.4, 0.7\}$, write the likelihood function.

(b) If the observed data are $\{x_1, x_2, x_3, x_4, x_5\} = \{0.1, 0.2, 0.3, 0.4, 0.9\}$, write the likelihood function.

(c) If the observed data are $\{x_1, x_2, x_3, x_4, x_5\} = \{0.1, 0.2, 0.3, 0.6, 0.9\}$, write the likelihood function.

(d) Plot the log-likelihood functions in cases (a), (b) and (c) on the same graph, as functions of p_1.

(e) Based on your plots, what are the estimated $(\hat{p}_1, \hat{p}_2)$ in cases (a), (b) and (c)? Are the estimates reasonable in every case?

5. (*Mixture model analysis practice*)

The data heart.dat.txt (automatic download when clicked, also available next to the link of this p-set in Canvas) of 270 observations can be read in R by the following commands.

```
heart=read.table("heart.dat.txt")
colnames(heart)=c("age","sex","chest","bp","chl","sugar","ecg","rate","angina","peak","slope","vssl","thal","ill")
```

The 14 variables are:

```
-- 1. age
-- 2. sex
-- 3. chest pain type  (4 values)
-- 4. resting blood pressure
-- 5. serum cholestoral in mg/dl
-- 6. fasting blood sugar > 120 mg/dl
-- 7. resting electrocardiographic results  (values 0,1,2)
-- 8. maximum heart rate achieved
-- 9. exercise induced angina
-- 10. oldpeak = ST depression induced by exercise relative to rest
-- 11. the slope of the peak exercise ST segment
-- 12. number of major vessels (0-3) colored by flourosopy
-- 13.  thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
-- 14. illness: 1 = absence; 2 = presence of heart disease
```

The variable types are

```
Real: 1,4,5,8,10,12
Ordered:11
Binary: 2,6,9
Nominal:7,3,13,14
```

Conduct a mixture analysis on the data using the real variables or a subset of them. Show your main steps, plots, and results. Interpret (comment on the goodness of fit of the model, the clusters, etc.).

6. (*Hands-on EM for missing values in multivariate normal*)

The data consisting of 4 observations from trivariate normal $N_3(\boldsymbol{\mu}, \Sigma)$ are given with missing components:

$$\boldsymbol{X} = [x_{jk}] = \begin{bmatrix} 3 & 6 & 0 \\ 4 & 4 & 3 \\ - & 8 & 3 \\ 5 & - & - \end{bmatrix} \begin{matrix} obs1 \\ obs2 \\ obs3 \\ obs4 \end{matrix}$$

Use the Expectation (*a.k.a.* prediction) – Maximization (*a.k.a.* estimation) algorithm to estimate $\boldsymbol{\mu}$ and Σ.

(a) Impute values to obtain $\tilde{\boldsymbol{X}}$, the data matrix with missing values filled by initial estimates.

(b) Find $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ estimated from $\tilde{\boldsymbol{X}}$.

(c) Iterate to find the first revised estimates:

i. Find the revised estimate for the missing value x_{31} in observation 3, using the estimated $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ in step (b). What is the updated estimates of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$?

ii. Find the revised estimates for the missing observations x_{42} and x_{43} in observation 4, using the estimated $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ in step (i). (Note the slight difference from the update step used in the EM-impute demo, where parameter update happens after all records are imputed.)

Write out the updated estimates of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$, the revised estimates after the first iteration.

7. (*Global-equivalence of dissimilarity measures*) **Required for 32950. Optional for 24620.**

When we have two different similarity measures, it is important to know their relationship.

Two dissimilarity measures (also called *dissimilarity coefficients*) d_1 and d_2 defined in a space S (e.g. $S = \mathbb{R}^p$) are *global-order equivalent* if

$$d_1(x, y) \leq d_1(z, w) \quad \text{if and only if} \quad d_2(x, y) \leq d_2(z, w) \qquad \text{for all } x, y, z, w \text{ in } S.$$

(a) Let $d(a, b) = \sqrt{(a-b)^T(a-b)}$ be the Euclidean distance between vectors a and b in $\mathbb{R}^p$, and $D(a, b) = (d(a, b))^2$. Show that $d_1 = d$ and $d_2 = D$ are global-order equivalent.

(b) Show by a counter example that global-order equivalence does not hold for

$d_1 =$ *Euclidean distance* and $d_2 =$ *city-block metric.*

The city-block metric is defined as $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ (with the entry = 0 when $x_i - y_i = 0$).

(Note: This is a normalized or weighted version of the city-block metric, also called Canberra metric.)