Homework 2
Applied Machine Learning
Fall 2017
CSCI-P 556/INFO-I 526

Instructor: Hasan Kurban

September 18, 2017

Directions

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Monday Oct 2, 2017 11:59p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. Bring a hardcopy to Tuesdays class on the 3rd. If you do not bring a hardcopy with the statement of your own work, the homework will not be accepted. I will not make exceptions. All the work should be your own. Within a week, AIs can contact students to examine code; students must meet within three days. The session will last no longer than 5 minutes. If the code does not work, the grade for the program may be reduced. Lastly, source code cannot be modified post due date.

Problem 1 [20 points]

From textbook, Chapter 10 exercise 2 (Page 414).

Problem 2 [50 points]

Implement expectation-maximization algorithm for Gaussian mixture models (see the EM algorithm below) in R and call this program G_k. As you present your code explain your protocol for

3.1 initializing each Gaussian

3.2 maintaining k Gaussian

3.3 deciding ties

3.4 stopping criteria

Problem 3 [70 points]

In this questions, you are asked to run your program, G_k, against the Ringnorm and Ionosphere data sets and compare G_k with C_k (k-means algorithm from previous homework). Click on the below links to download the data sets.

- Ringnorm Data Set
- Ionosphere Data Set

Answer the following questions:

3.1 Initialize G_k and C_k with the same set of initial points (initial centroids for C_k and μ_i-s for G_k are identical) and run them for $k = 2, \ldots, 5$ for 20 runs each. Report error rates and iteration counts for each k using whisker plots that reveal comparison of C_k and G_k. An example of whisker plot is given below. A simple error rate can be calculated as follows:

- If $k = 2$: C_k and G_k will predict two clusters. Error calculation is trivial for two clusters.
- If $k > 2$: after C_k and G_k converge, combine the clusters as follows to ended up with two clusters: since the true clusters are known for a given arbitrary blocks number, final clusters are determined by measuring the Euclidean (this is the easiest choice) distances between true cluster centers and predicted cluster centers.

In other words, you will always calculate the error for $k = 2$ since there are only 2 clusters in the given data sets. Below is an example of error calculation for Ionosphere data set. You can similarly calculate an error rate for Ringnorm data set.

For each centroid C_i, and each Gaussian G_k form two counts (over Ionosphere Data Set) :

$$g_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = \text{“g”}], \quad \text{good}$$

$$b_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C == \text{“b”}], \quad \text{bad}$$

where $[x = y]$ returns 1 if True, 0 otherwise. For example, $[2 = 3] + [0 = 0] + [34 = 34] = 2$

The centroid C_i and Gaussian G_k is classified as good if $g_i > b_i$ and bad otherwise. We can now calculate a simple error rate. Assume C_i is good. Then the error is:

$$error(C_i) = \frac{b_i}{b_i + g_i} \quad \text{[same for error}(G_i)\text{]}$$

We can find the total error rate easily:

$$Error(\{C_1, C_2\}) = \sum_{i=1}^{2} error(C_i)$$

Discuss your results, i.e., which one performs better.

3.2 In this question, we will run your G_k with fixing the variances to ones and the priors to be uniform. Do not update the variances and priors throughout iterations. As explained in question 3.1, compare your new G_k and C_k using whisker plots. Discuss your results, i.e., which one performed better.

Problem 4 [50 points]

In this question, you will first perform principal component analysis (PCA) over Ionosphere and Rignorm data sets and then cluster the reduced data sets using G_k (from question 3.1) and C_k. You are allowed to use R packages for PCA. Ignore the class variables (35th and 1st variables for Ionosphere and Ringnorm data sets, respectively) while performing PCA. Answer the questions below:

4.1 Make a scatter plot of PC1 and PC2 for both data sets. Discuss principal components (The first and second principal components). What are PC1 and PC2?

4.2 Create scree plots after PCA and explain the plots.

4.3 Observe the loadings using prcomp() or princomp() functions in R and discuss loadings in PCA?i.e., how are principal components and original variables related?

4.4 Keep 90% of variance after PCA and reduce Ionosphere and Rignorm data sets. Run C_k and G_k with the reduced data sets and compare them using whisker plots as shown in question 3.1

4.5 Discuss that how PCA affects the performance of C_k and G_k.

Problem 5 [50 points]

Randomly choose 50 points from Ionosphere data set (call this data set I_{50}) and perform hierarchical clustering. You are allowed to use R packages for this question. (Ignore the class variable while performing hierarchical clustering.)

5.1 Using hierarchical clustering with complete linkage and Euclidean distance cluster I_{50}. Plot the dendrogram.

5.2 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate.

5.3 First, perform PCA on I_{50} (Keep 90% of variance). Then hierarchically cluster the reduced data using complete linkage and Euclidean distance. Plot the dendrogram

5.4 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate. How did PCA affect hierarchical clustering?

Extra credit [60 points]

This part is optional.

1 Improve the EM algorithm through initialization. k-means ++ is an extended k-means clustering algorithm and induces non-uniform distributions over the data that serve as the initial centroids. Read the paper and implement this idea to improve your G_k program (from question 3.1). Run your new G_k and old one (question 3.1) for $k = 2, \ldots, 5$ and compare the results using whisker plots. [30 points]

2 Run the EM algorithm for different mixture models, i.e., Poisson, and against different data sets. [30 points]

The EM algorithm

This part is provided to help you implement the EM algorithm.

Let $\mathbf{D} = \{\mathbf{x_j} \mid j = 1, \ldots, n\}$ be the data set where each $\mathbf{x_j} \in \Re^d$ ($\Re$: Reals) and $\mathbf{D}$ is a mixture of a Gaussian. Given $\mathbf{D}$, the number of blocks k, and convergence threshold ϵ, the EM-T algorithm partition data into k clusters, $\mathsf{C} = \{C_1, C_2, \ldots, C_k\}$, where each $C_i \in \mathsf{C}$ can be characterized as a Gaussian distribution. If each cluster $C_i \in \mathsf{C}$ can be represented by a multivariate normal distribution (MVN):

$$f(\mathbf{x_j}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} * \exp\Big\{ -\frac{1}{2}(\mathbf{x_j} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x_j} - \mu_i) \Big\}$$

where $\mu_i \in \Re^d$ and $\Sigma_i \in \Re^{d \times d}$ denote cluster mean and covariance matrix for cluster C_i, respectively, and they are unknown parameters. $f(\mathbf{x_j}|\mu_i, \Sigma_i)$ represents the probability density at $\mathbf{x_j}$ for cluster C_i. Lastly, $\mathbf{D}$ is a mixture of $C_1, C_2, \ldots, C_k$.

The algorithm iteratively alternates between (1) computing log-likelihood of each data point being from each Gaussian (E-step) (2) recalculating the parameters (M-step). Iteration continues until a set of means is stable.

- **Initialization:**
  - μ_i is randomly selected from $\mathbf{D}$ for each cluster.
  - $\Sigma_i \leftarrow I$. For each cluster, the covariance matrix is a $d \times d$ identity matrix.
  - $P(C_i) = \frac{1}{k}$. The priors are uniformly initialized.

EXPECTATION-MAXIMIZATION $(\mathbf{D}, k, \epsilon)$:

$t \leftarrow 0$

// Initialization

Randomly initialize $\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_k^t$

$\boldsymbol{\Sigma}_i^t \leftarrow \mathbf{I},\ \forall i = 1, \ldots, k$

$P^t(C_i) \leftarrow \frac{1}{k},\ \forall i = 1, \ldots, k$

repeat

  $t \leftarrow t + 1$

  // Expectation Step

  for $i = 1, \ldots, k$ *and* $j = 1, \ldots, n$ **do**

    $w_{ij} \leftarrow \frac{f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot P(C_i)}{\sum_{a=1}^{k} f(\mathbf{x}_j | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) \cdot P(C_a)}$ // posterior probability $P^t(C_i | \mathbf{x}_j)$

  // Maximization Step

  for $i = 1, \ldots, k$ **do**

    $\boldsymbol{\mu}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$ // re-estimate mean

    $\boldsymbol{\Sigma}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$ // re-estimate covariance matrix

    $P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$ // re-estimate priors

until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

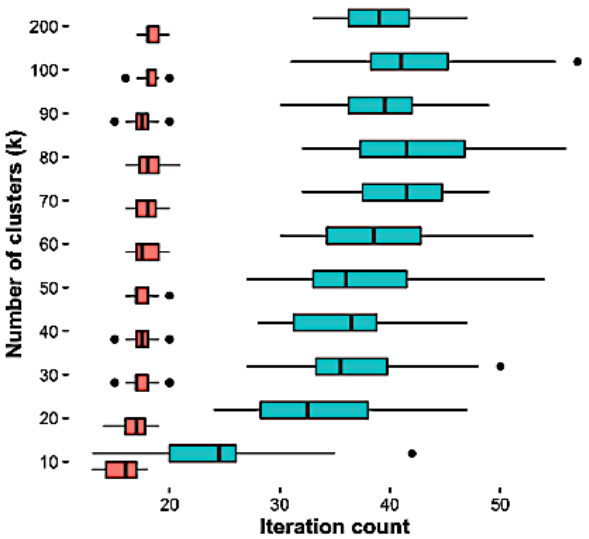

Figure 1: An example of whisker plot

What to Turn-in – Submission Instructions

- Zip the files requested below for your submission. The zipped folder should be named as "usename-section number", i.e, hakurban-P556
  - The *tex and *pdf of the written answers to this document.
  - *Rfiles for:
    - The EM implementation for problem 3.1 – call it "G_k3.1.R".
    - The EM implementation for problem 3.2 – call it "G_k3.2.R".
    - R code for problem 4. Include everything in a .R file and name it "pca4.R".
    - R code for problem 5. Include everything in a .R file and name it "hierarchical5.R".
    - extra credit questions – If you answer the extra credit question, include also the code for extra question in your submission folder. (call the file "extra.R".)
  - A README file that explains how to run your code and other files in the folder

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[2] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.