THU-70250043, Pattern Recognition (Spring 2017) **Homework: 11**

Clustering Methods

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student: XXX xxx@mails.tsinghua.edu.cn

Problem 1

K-means and EM algorithm

As you have learned in class, we are always trying to minimize the *distortion measure* in K-means, given by:

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\mathbf{x}_n - \mu_k||^2 \tag{1}$$

Where, μ_k are the centers of the K clusters, $\mathbf{x}_n$ are D-dimensional data points and $r_{nk} \in \{0, 1\}$ are the indicator variables for the data points.

We try to find the appropriate values of $\{r_{nk}\}$ and $\{\mu_k\}$ through an iterative procedure in which each iteration involves two successive steps corresponding to successive optimizations with respect to the $\{r_{nk}\}$ and the $\{\mu_k\}$. We shall see that these two stages correspond respectively to the $\mathbf{E-step}$ and the $\mathbf{M-step}$ of the EM algorithm.

E-step

Optimize for each n separately by choosing r_{nk} to be 1 for whichever value of k gives the minimum value of $||\mathbf{x}_n - \mu_k||^2$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = argmin_j||\mathbf{x}_n - \mu_j||^2 \\ 0 & \text{otherwise.} \end{cases}$$

M-step

Minimize J by setting its derivative with respect to μ_k to zero giving:

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

1.1 Consider the K-means algorithm discussed above. Show that as a consequence of there being a finite number of possible assignments for the set of discrete indicator variables r_{nk}, and that for each such assignment there is a unique optimum for the $\{\mu_k\}$, the K-means algorithm must converge after a finite number of iterations.

Consider a Gaussian mixture model in which the covariance matrices of the mixture components are given by $\epsilon\mathbf{I}$, where ϵ is a variance parameter that is shared by all of the components, and $\mathbf{I}$ is the identity matrix, so that:

$$p(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{D/2}} exp\{-\frac{1}{2\epsilon}||\mathbf{x} - \mu_k||^2\} \tag{2}$$

Now consider the EM algorithm for a mixture of K Gaussians of this form in which we treat ϵ as a fixed constant.

1.2 Derive the posterior probability $\gamma(z_{nk})$ for a particular data point $\mathbf{x}_n$ and show that with $\epsilon \longrightarrow 0$, $\gamma(z_{nk}) \longrightarrow r_{nk}$.

1.3 Write down the expected value of the complete-data log likelihood function $\mathbb{E}[lnp(\mathbf{X}, \mathbf{Z}|\mu, \mathbf{\Sigma}, \pi)]$ for the above Gaussian mixture model and show that as $\epsilon \longrightarrow 0$, $\mathbb{E}[lnp(\mathbf{X}, \mathbf{Z}|\mu, \mathbf{\Sigma}, \pi)] \longrightarrow -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\mathbf{x}_n - \mu_k||^2 + const.$

Hint: Above derivation shows that there is a close similarity between K-means and EM algorithm for Gaussian mixtures. Whereas the K-means algorithm performs a hard assignment of data points to clusters, in which each data point is associated uniquely with one cluster, the EM algorithm makes a soft assignment based on the posterior probabilities.

Problem 2

Programming

Test the clustering algorithms $\mathbf{K-means}$, **hierarhical clustering** and **spectral clustering** with different parameters on MNIST dataset or subsets of it when the scale is too large for the algorithm involved.

To compare the effectiveness of different clustering methods, *Normalized mutual information*(NMI) are widely used as a measurement. NMI is defined as following:

$$NMI = \frac{\sum_{s=1}^{K}\sum_{t=1}^{K} n_{s,t} log(\frac{n n_{s,t}}{n_s n_t})}{\sqrt{(\sum_s n_s log\frac{n_s}{n})(\sum_t n_t log\frac{n_t}{n})}} \tag{3}$$

Where n is the number of data points, n_s and n_t denote the numbers of the data in class s and class t, $n_{s,t}$ denotes the number of data points in both class s and class t. For more details and other measurements, google "evaluation of clustering".

2.1 Give a brief analysis of time complexity of each algorithm mentioned above (of standard implementation). Estimate how many samples each algorithm can manage with a reasonable time cost.

(Optional) Can you verify your estimations with experiments? Can you speed it up further?

2.2 Consider each data set, and use the true number of classes as the number of clusters.

- With K-means, will the initial partition affect the clustering results? How can you solve this problem? And do J_e and NMI match? Show your experiment results.
- When hierarchical clustering is adopted, the choice of linkage method depends on the problem. Give an analysis of linkage method's effects with experiments, and which is better in the sense of NMI? For more linkage methods, refer to the linkage function's help in matlab.
- As above, give an experimental analysis of the choice of similarity graph and corresponding parameters. Which one is better?

2.3 In practice, we may not know the true number of clusters much. Can you give a strategy to identify the cluster number automatically for each algorithm? Show your results.

2.4 According to the above analysis, which method do you prefer? Why?