1 Clustering

In the lectures, we discussed k-means. Given a set of data points $\{\mathbf{x}_n\}_{n=1}^N$, the method minimizes the following distortion measure (or objective or clustering cost):

$$D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$

where $\boldsymbol{\mu}_k$ is the prototype of the k-th cluster. r_{nk} is a binary indicator variable. If $\mathbf{x}_n$ is assigned to the cluster k, r_{nk} is 1 otherwise r_{nk} is 0. For each cluster, $\boldsymbol{\mu}_k$ is the representative for all the data points assigned to that cluster.

(a) In the lecture, we showed but did not prove that, $\boldsymbol{\mu}_k$ is the mean of all such data points. That is why the method has MEANS in its name and we keep referring to $\boldsymbol{\mu}_k$ as MEANS, CENTROIDS, etc. You will prove this rigorously next. Assuming all r_{nk} are known (that is, you know the assignments of all N data points), show that if $\boldsymbol{\mu}_k$ is the mean of all data points assigned to the cluster k, for any k, then the objective D is minimized. This justifies the iterative procedure of k-means[1].

(b) We now change the distortion measure to

$$D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|_1$$

In other words, the measurement of "closeness" to the cluster prototype is now using L_1 norm ($\|\mathbf{z}\|_1 = \sum_d |z_d|$).

Under this new cost function, show the $\boldsymbol{\mu}_k$ that minimizes D can be interpreted as the elementwise median of all data points assigned to the k-th cluster. (The elementwise median of a set of vectors is defined as a vector whose d-th element is the median of all vectors' d-th elements.)

(c) Now assume that we apply a mapping $\phi(\mathbf{x})$ to map data points into feature space. Then, we define the objective function of kernel K-means as

$$\tilde{D} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\phi(\mathbf{x}_n) - \tilde{\boldsymbol{\mu}}_k\|_2^2,$$

where $\tilde{\boldsymbol{\mu}}_k$ is the center of the cluster k in the feature space.

- Show that the $\tilde{D}$ can be represented in terms of only kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x_i})^{\mathrm{T}}\phi(\mathbf{x_j})$.
- Describe and write down the equation of assigning a data point to its cluster. You answer should only consist of kernel value $K(\mathbf{x}_i, \mathbf{x}_j)$.
- Write down the pseudo-code of the complete kernel K-means algorithm including initialization of cluster centers.

[1] More rigorously, one would also need to show that if all $\boldsymbol{\mu}_k$ are known, then r_{nk} can be computed by assigning $\mathbf{x}_n$ to the nearest $\boldsymbol{\mu}_k$. You are not required to do so.

2 Gaussian Mixture Model

Let our data be generated from a mixture of two univariate Gaussian distributions, where $f(x|\theta_1)$ is a Gaussian with mean $\mu_1 = 0$ and $\sigma_1^2 = 1$, and $f(x|\theta_2)$ is a Gaussian with mean $\mu_2 = 0$ and $\sigma_2^2 = 0.5$. The only unknown parameter is the mixing parameter α (which specifies the prior probability of θ_1.). Now we observe a single sample x_1, please write out the likelihood function of x_1 as a function of α, and determine the maximum likelihood estimation of α.

3 EM algorithm

Zero-inflated Poisson regression is used to model count data that has an excess of zero counts. For example, the number of insurance claims within a population for a certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim. The observed data probability of observation i is:

$$p(x_i) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } x_i = 0 \\ (1-\pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} & \text{if } x_i > 0. \end{cases}$$

Your task in this problem is to design an EM to estimate the parameter π and λ from observed data $\{x_i\}_{i=1}^N$.

(a) Define a proper hidden variable z_i for the observations (Hint: you only need hidden variables for some observations) and use them to write down the complete likelihood function.

(b) Write down the update equations for both the E-Step and the M-step.

4 Programming

In this problem, you will implement three different clustering methods, K-means, Kernel K-means and Gaussian Mixture Model. You will evaluate the performance of your method on two synthetic datasets.

4.1 Data

You are provided with two datasets, `hw5_blob.csv` and `hw5_circle.csv`. Both datasets have two dimensions.

4.2 Implement k-means

As we studied in the class, k-means tries to minimize the following distortion measure (or objective function):

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2$$

where r_{nk} is an indicator variable:

$$r_{nk} = 1 \quad \text{if and only if} \quad \mathbf{x}_n \quad \text{belongs to cluster} \quad k$$

and $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ are the cluster centers with the same dimension of data points.

(a) Implement k-means using random initialization for cluster centers. The algorithm should run until none of the cluster assignments are changed. Run the algorithm for different values of $K \in \{2, 3, 5\}$, and plot the clustering assignments by different colors and markers. (you need to report 6 plots, 3 for each dataset.)

(b) The k-means algorithm fails to separate the two circles in the `hw5_circle.csv` dataset. Please explain why this happens.

4.3 Implement kernel k-means

Implement kernel k-means you derived in Problem 2 and evaluate it on the `hw5_circle.csv` dataset. You should choose a kernel that can separate the two circles.

(a) Write down the choice of your kernel.

(b) Implement kernel k-means using random initialization for cluster centers (randomly pick data points as centers of the clusters). The algorithm should run until none of the cluster assignments are changed. Run the algorithm for $K = 2$, and plot the clustering assignments by different colors and markers. (you need to report 1 plot)

4.4 Implement Gaussian Mixture Model

In this problem, you need to implement the EM algorithm to fit a Gaussian Mixture model on the `hw5_blob.csv` dataset.

(a) Run 5 times of your EM algorithm with number of components $K = 3$, and plot the log likelihood of the data over iterations of EM for each of runs. (The x-axis is the number of iterations, and the y-axis is the log likelihood of the data given current model parameters. Please plot all five curves in the same figure)

(b) For the best run in terms of log likelihood, (1) Plot the most likely cluster assignment of each data point indicated by different colors and markers. (2) Report the mean and co-variance matrix of all the three Gaussian components.

5 Submission Instructions

Submission Instruction: You need to provide the followings:

- Provide your all your answers that are required for grading in hardcopy. The graders are not required to check your answers in the softcopy. The papers need to be stapled and submitted into the locker#19 on the first floor of PHE building.
- Provide your answers to problems in `*.pdf` file, named as `CSCI567_hw5_fall16.pdf`. You need to submit the homework in both hardcopy (at the collection locker #19 at the PHE building 1st floor by **5:00pm** of the deadline date) and electronic version as `*.pdf` file on Blackboard. If you choose handwriting instead of typing all the answers, you will get 40% points deducted.
- Submit ALL the code and report via Blackboard. The only acceptable language is MATLAB or Python2.7. For your program, you MUST include the main script called `CSCI567_hw5_fall16.m` or `CSCI567_hw5_fall16.py` in the root of your folder. After running this main file, your program should be able to generate all of the results needed for this programming assignment, either as plots or console outputs. You can have multiple files (i.e your sub-functions), however, the only requirement is that once we unzip your folder and execute your main file, your program should execute correctly. Please double-check your program before submitting. You should only submit one `*.zip` file. No other formats are allowed except `*.zip` file. Also, please name it as `[lastname]_[firstname]_hw2_fall16.zip`.

Collaboration You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solution and submit separately. You also need to list with whom you have discussed.