CSE 575: Statistical Machine Learning Assignment #3

Instructor: Prof. Jingrui He
Out: Oct. 21th, 2016; Due: Nov. 15th, 2016

Submit electronically, using the submission link on Blackboard for Assignment #3, a file named `yourFirstName-yourLastName.pdf` *containing your solution to this assignment (a .doc or .docx file is also acceptable, but .pdf is preferred).*

1 Kmeans [20 points]

Given N data points $x_i, (i = 1, ..., N)$, Kmeans will group them into K clusters by minimizing the distortion function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \|x_n - \mu_k\|^2$, where μ_k is the center of the k^{th} cluster; and $r_{n,k} = 1$ if x_n belongs to the k^{th} cluster and $r_{n,k} = 0$ otherwise. In this exercise, we will use the following iterative procedure

- Initialize the cluster center $\mu_k, (k = 1, ..., K)$;
- Iterate until convergence
  - Update the cluster assignments for every data point x_n: $r_{n,k} = 1$ if $k = \text{argmin}_j \|x_n - \mu_j\|^2$; $r_{n,k} = 0$ otherwise.
  - Update the center for each cluster k: $\mu_k = \frac{\sum_{n=1}^{N} r_{n,k} x_n}{\sum_{n=1}^{N} r_{n,k}}$

(1) **Convergence of Kmeans** [10 points]

Prove that the above procedure will converge in finite steps.

- **Hint: Consider whether or not the number of possible cluster assignments is finite.**
- **Solution:** Notice that for each cluster assignment, the corresponding cluster centers μ_k(k=1,...K) are unique. Therefore, in each iteration, we must try a new cluster assignment. On the other hand, notice that all possible cluster assignments are finite (K^N). Therefore, the algorithm must converge in finite iterations.

(2) **Kmeans and GMM** [10 points]

Remember in GMM, $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $\pi_k = p(z_k = 1)$ is the prior for the k^{th} component; and μ_k,Σ_k are the mean and covariance matrix for k^{th} component respectively. In the E-step, we will update $p(z_k = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$

Now suppose that

(1) $\Sigma_k = \epsilon \mathbf{I}$ where ϵ is some *given* positive number;

(2) $\pi_k \neq 0$ $(k = 1, ..., K)$;

(3) $\|x_n - \mu_i\| \neq \|x_n - \mu_j\|$ for any $i \neq j$.

Under the above assumptions, prove that when $\epsilon \to 0$, $p(z_k = 1|x_n) = r_{n,k}$, where $r_{n,k}$ is the cluster assignment used in Kmeans.

- **Solution:**

$$\begin{aligned} p(z_k = 1|x_n) &= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \\ &= \frac{\pi_k \text{exp}\{-\frac{1}{2\epsilon}\|x_n - \mu_k\|^2\}}{\sum_{i=1}^{K} \pi_i \text{exp}\{-\frac{1}{2\epsilon}\|x_n - \mu_i\|^2\}} \\ &= \frac{1}{1 + \sum_{i \neq k} (\frac{\pi_i}{\pi_k}) \text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2)\}} \end{aligned} \tag{1}$$

Therefore, if $\|x_n - \mu_k\| = \min_i \|x_n - \mu_i\|$, for each $i \neq k$, we have $\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2 < 0$. Thus as $\epsilon \to 0^+$, $\text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2)\} \to 0$. So, $p(z_k = 1|x_n) \to 1$.

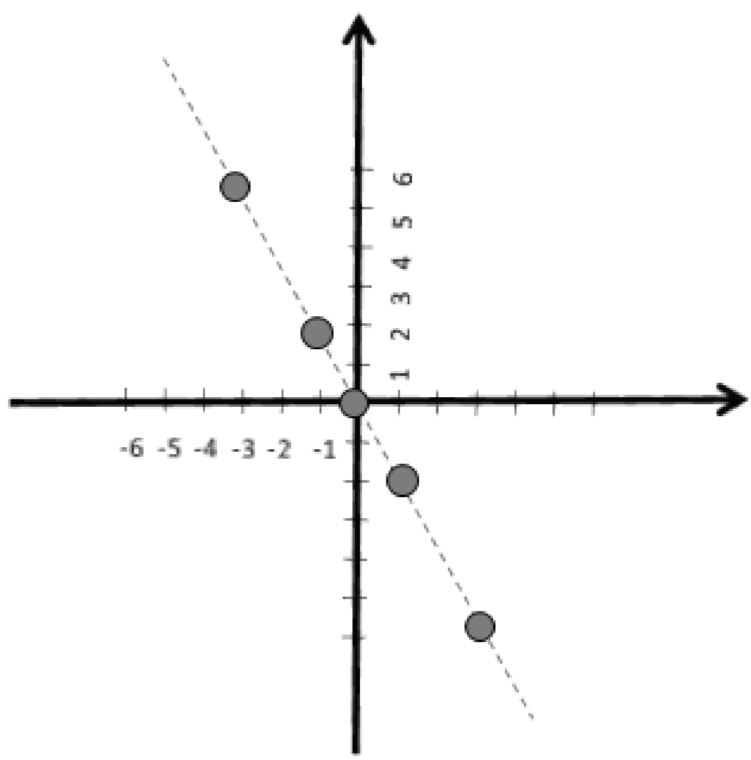
On the other hand, if $\|x_n - \mu_k\| \neq \min_i \|x_n - \mu_i\|$. Let $\|x_n - \mu_{\tilde{k}}\| \neq \min_i \|x_n - \mu_i\|$, we have $\|x_n - \mu_k\|^2 - \|x_n - \mu_{\tilde{k}}\|^2 > 0$. Thus as $\epsilon \to 0^+$, $\text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_{\tilde{k}}\|^2)\} \to +\infty$. So, $p(z_k = 1|x_n) \to \frac{1}{1+\infty} = 0$.

2 PCA [18 points]

Suppose we have the following data points in 2-d space $(0, 0)$, $(-1, 2)$, $(-3, 6)$, $(1, -2)$, $(3, -6)$.

1 [6 points] Draw them on a 2-d plot, each data point being a dot.

- **Solution:**



2 [6 points] What is the first principle component (3 points)? Give 1-2 sentences justification (3 points). (**Hint: You do not need to run Matlab to get the answer.**)

- **Solution:** $\frac{1}{\sqrt{5}}(-1, 2)$

3 [6 points] What is the second principle component (3 points)? Give 1-2 sentences justification (3 points). (**Hint: You do not need to run Matlab to get the answer.**)

- **Solution:** $\frac{1}{\sqrt{5}}(2, 1)$

3 HMM [24 points]

Suppose that we use four distinct words to write a paragraph with 100 segments, and we treat each word in the paragraph as a segment. We want to infer three posible class labels of all the segments in this paragraph, including (a) location (b) person name and (c) background by HMM (Hidden Markov Models).

1. [6 points] What is the size of the state transition probability matrix in our HMM model?

- **Solution:** 3×3

2. [6 points] What is the size of the state-observation probability matrix?

- **Solution:** 3×4

3. [6 points] In a particular trial, how many observations do you see [3pts]? What is the length of the path of states [3pts]?

- **Solution:** 100 and 100

4. [6 points] Suppose that the first state is about 'background', how many different possible state paths are there in total?

- **Solution:** 3^{99}

4 Kmeans Implementation [38 points]

Download the file hw3.zip and uppack it. The file seeds_dataset.txt contains 210 examples with 7 features. Implement the Kmeans algorithm with the number of clusters k changing from 2 to 10. For each number of clusters, compute the average within cluster distance, which is defined as $\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sum_{j=1}^{m}I(C(j)=i)}\sum_{j=1}^{m}I(C(j)=i)\|\mu_i - x_j\|^2$. Here $I(C(j)=i)$ is an indicator function. It is equal to 1 if $C(j)=i$, and 0 otherwise.

Plot the average within cluster distance vs. the number of clusters k. Can you pick the optimal number of clusters k to minimize the average within cluster distance? Why?
Hint: In Kmeans iterations, if a cluster does not have any data points in it, remove the cluster, and randomly split the largest cluster into 2 clusters. In this way, the total number of clusters remains unchanged. Solutions: We cannot pick the optimal number of clusters to minimize the

average within cluster distance as it will lead to the number of clusters equal to the number of examples.

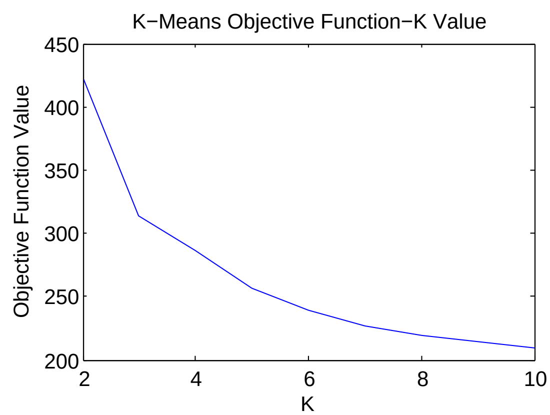


Figure 1: Average within cluster distance vs. number of clusters.