CS 475 Machine Learning: Homework 4
The EM Algorithm (and more)

Due: Monday November 9, 2015, 11:59pm
100 Points Total Version 1.0

Li-Yi Lin / llin34@jhu.edu

1 Analytical Questions (40 points)

1. Overfitting in Clustering (10 points) Given the data set $x_1, ..., x_n$, we want to do clustering by the K-means algorithm. The K-means algorithm aims to partition the n observations into k sets $(k < n)$ $S = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares

$$\min_{S=\{S_1,...,S_k\}} \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_j - \mu_j\|_2^2$$

(a) Prove that the objective value does not increase in each iteration of the K-means algorithm.
Ans:
We first prove that, in each iteration, using μ_j as the center of each cluster will minimize the objective value. To prove that, we perform partial derivative on each μ_j and set it equal to 0. Then we find the answer of μ_j.

$$\frac{\partial \sum_{x_i \in S_j} \|x_i - \mu_j\|_2^2}{\partial \mu_j} = \sum_{x_i \in S_j} -2\|x_i - \mu_j\| = 0$$

$$\sum_{x_i \in S_j} -2\|x_i - \mu_j\| = \sum_{x_i \in S_j} -2x_i - \sum_{x_i \in S_j} -2\mu_j = 0$$

$$\sum_{x_i \in S_j} x_i = \sum_{x_i \in S_j} \mu_j$$

We assume there are n_j elements in S_j.

$$\sum_{x_i \in S_j} x_i = \sum_{x_i \in S_j} \mu_j = n_j \mu_j$$

$$\mu_j = \frac{\sum_{x_i \in S_j} x_i}{n_j}$$

So we know that when μ_j is the mean of points in S_j, the object value will be minimized since the sum of squares error of each cluster is minimized.
Once we have updated the center, μ_j of each cluster, we reassign $x_1, ..., x_n$ to the new

nearest cluster respectively. The new sum of square error will be less than or equal to that of using the old cluster assignments, namely,

$$\|x_i - \mu_{\text{new assigned cluster center}}\|_2^2 \leq \|x_i - \mu_{\text{original assigned cluster center}}\|_2^2$$

Therefore, by doing the two steps in every iteration, the objective value does not increase in each iteration.

(b) Let γ_k denote the global optimal objective value, prove γ_k is non-increasing in k.
Ans:
Assume we now have k clusters and it is at the optimal solution having object value γ_k. Then we add another cluster and perform the k-means algorithm to adjust the $k+1$ clusters. By the proof given in (1), we know that the object value doesn't increase in each iteration. Therefore, when we found the optimal solution for $k+1$ clusters, $\gamma_k \leq \gamma_{k+1}$. Thus, we have proved that γ_k is non-increase in k.

2. Curse-of-dimensionality (10 points) In this problem, we study why K-NN could fail in high dimensions by means of a very simple example. Consider a sphere of radius r in d-dimensions together with a concentric hypercube of side $2r$. The sphere touches the hypercube at the center of each of its sides.

(a) V_c is the volume of the cube and V_s is the volume of the sphere, where the volume of a d-dimensional sphere with radius r is given as

$$V_s = \frac{r^d \sqrt{\pi}^d}{\Gamma(d/2+1)},$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$. Note that $\Gamma(z) = (z-1)!$ (the factorial of $z-1$) if z is a positive integer. Please show that:

$$\lim_{d\to\infty} \frac{V_s}{V_c} = 0 \tag{1}$$

Note that this relies on algebra and will not require any complex calculus (just some basic facts on limits). You may find the following limit useful:

$$\lim_{z\to\infty} \frac{\Gamma(z+1)}{\sqrt{2\pi z} e^{-z} z^z} = 1$$

Ans:

$$\lim_{d\to\infty} \frac{V_s}{V_c} = \lim_{d\to\infty} \frac{r^d \sqrt{\pi}^d}{\Gamma(d/2+1)(2r)^d} = \lim_{d\to\infty} \frac{\sqrt{\pi}^d}{\Gamma(d/2+1)2^d}$$

Let $2z = d$

$$\lim_{d\to\infty} \frac{\sqrt{\pi}^d}{\Gamma(d/2+1)2^d} = \lim_{2z\to\infty} \frac{\pi^z}{\Gamma(z+1)2^{2z}}$$

We knew that when $z \to \infty$, $\Gamma(z+1) \to \sqrt{2\pi z} e^{-z} z^z$. We can substitute this into the above equation.

$$\lim_{2z\to\infty} \frac{\pi^z}{\Gamma(z+1)2^{2z}} = \lim_{2z\to\infty} \frac{\pi^z}{\sqrt{2\pi z} e^{-z} z^z 2^{2z}} = \lim_{2z\to\infty} \frac{\sqrt{\pi^{2z}}}{\sqrt{2\pi z}\sqrt{(2(2z)e^{-1})^{2z}}}$$

Since $\lim_{2z\to\infty} \frac{\pi^{2z}}{(2(2z)e^{-1})^{2z}} = 0$, we have

$$\lim_{d\to\infty} \frac{V_s}{V_c} = \lim_{2z\to\infty} \frac{\sqrt{\pi^{2z}}}{\sqrt{2\pi z}\sqrt{(2(2z)e^{-1})^{2z}}} = 0$$

(b) What is the connection between (1) and the curse of dimensionality?
Ans:
The ratio of the volume of sphere devided by the volume of cube is decreasing to zero as the dimentionality is getting larger and larger to infinity. That means if we randomly choose points in high dimentional space, they will likely be equidistant from each other, making it difficult to perform classification task.

3. Semi-supervised EM algorithm (10 points) Suppose that some of your observed data are labelled. You have $x_1, ..., x_n \in \mathbb{R}^d$ and $x_{n+1}, ..., x_{n+m}$. Meanwhile, you also know the labels corresponding to $x_{n+1}, ..., x_{n+m}$, i.e., $y_{n+1}, ..., y_{n+m} \in \{1, ..., K\}$, where K is the number of clusters. Please design a Gaussian Mixture Model-based EM algorithm to cluster the data. [Hint: For $x_{n+1}, ..., x_{n+m}$, the corresponding labels are no longer missing values]

(a) Write the new likelihood objective for this new algorithm.
Ans:

$$p(\mathbf{X},\ \mathbf{Z}|\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K} \pi_k^{z_{nk}}(N)(x_i|\mu_k, \Sigma_k)^{z_{ik}} \prod_{i=n+1}^{m} \mathcal{N}(x_i|\mu_{k=y_i}, \Sigma_{k=y_i})$$

(b) Write the new update rules in each iteration.
Ans:
In E-step, we will use both the labelled and unlabelled data for updating, but only update the label of the unlabelled data. In the M-step, we also use both the labelled and unlabelled data for updating the θ.

4. Modified EM (10 points) The EM algorithm we learned about in class is just one of several different general EM algorithms, all with similar goals and structures. In this problem we will consider an alternative EM algorithm which modifies the M-step. Instead of maximizing $\mathcal{L}(q, \theta)$ with respect to θ, the algorithm selects a single parameters $\theta_i \in \theta$ and modifies it to increase $\mathcal{L}(q, \theta)$.

(a) Will this new EM algorithm yield the same solution as the normal EM algorithm?
Ans:
The new EM algorithm will not necessarily yield the same solution as the normal EM algorithm. The resaon is that it might find another local optimal due to the difference in the updating of the parameters.

(b) Will this new EM algorithm converge (assuming lack of singularities)? If yes, then prove convergence. If no, then give a counterexample illustrating why not.
Ans:
The new EM algorithm will still converge because, in the E-step, it will try to make $q(Z)$ close to $p(Z|x, \theta)$ by adjusting the chosen θ_i. By doing so, $\mathcal{L}(q, \theta)$ will be increased. In the M-step, it will try to maximize the $\mathcal{L}$ with respect to the new θ (with

θ_i updated). So in each iteration, the log likelihood is a non-descresing function. In addition, if we assume there is a solution for the problem, then the likelihood will be bounded. Thus, because the log likelihood is non-descresing and bounded, it will converge.