Problem Set 5 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Gaussian Graphical Model

1.1 Derive $\phi_{ij}(x_i, x_j)$ and $\phi_i(x_i)$

$$P(X \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \Sigma}} exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) \tag{1}$$

$$P(X \mid \mu, \Omega) \propto exp(-\frac{1}{2}X^T \Omega X + (\Omega\mu)^T X) \tag{2}$$

$$= \prod_{i \in V} exp(-\frac{1}{2}x_i^T \Omega_{ii} x_i + (\Omega\mu)_i^T x_i) \prod_{(i,j) \in E} exp(-\frac{1}{2}x_i^T \Omega_{ij} x_j) \tag{3}$$

$$\psi_{ij}(x_i, x_j) = exp(-\frac{1}{2}x_i^T \Omega_{ij} x_j), \quad \psi_i(x_i) = exp(-\frac{1}{2}x_i^T \Omega_{ii} x_i + (\Omega\mu)_i^T x_i) \tag{4}$$

1.2 Prove $(i, j) \notin E \iff X_i \perp X_j | X_{V \setminus \{i,j\}}$

If $(i, j) \notin E$, there is no edge between node i and node j, then we can get:

$$\Omega_{ij} = 0, \qquad \psi_{ij}(x_i, x_j) = 1 \tag{5}$$

$$P(x_i, x_j \mid V\setminus\{i,j\}, \mu, \Omega) = \psi'_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j) \cdot \psi_i(x_i)^{\frac{1}{n(i)}} \cdot \psi_j(x_j)^{\frac{1}{n(j)}} \tag{6}$$

$$= \psi_i(x_i)^{\frac{1}{n(i)}} \cdot \psi_j(x_j)^{\frac{1}{n(j)}} \tag{7}$$

$$= P(x_i \mid V\setminus\{i\}, \mu, \Omega) \cdot P(x_j \mid V\setminus\{j\}, \mu, \Omega) \tag{8}$$

$$= P(x_i \mid V\setminus\{i,j\}, \mu, \Omega) \cdot P(x_j \mid V\setminus\{i,j\}, \mu, \Omega) \tag{9}$$

We can notice the above process could be inverse, which means if we know that the two points are independent, we could get that there is no edge between these two nodes. Therefore, we could prove the conclusion.

2 Sampling

2.1 Inverse Sampling

2.1.1 Prove Inverse Sampling

First, for y' ∈ R, we have:

$$P(h^{-1}(z) <= y') = P(infy : h(y) = z < y') \tag{10}$$

$$= P(z <= h(y')) = h(y') \tag{11}$$

Second, for $0 < z' < 1$, we have:

$$P(h(y) < z') = P(y < h^{-1}(z')) = z' \tag{12}$$

Therefor, we could prove that inverse sampling is reasonable. The drawback could be:

(1) In real application, it would require a closed form expression for F(x), which means we would need do normalization sometimes.

(2) In real application, we need to konw the p(y) to do sampling.

2.1.2 Find the Cauchy distribution transfermation

Given the density function, we could derive:

$$h(y) = \int_{-\infty}^{y} p(y')dy' = \frac{1}{2} + \frac{1}{\pi}arctan(y) \tag{13}$$

$$y = g(z) = h^{-1}(z) = tan(\pi(z - \frac{1}{2})) \tag{14}$$

2.2 Rejection Sampling

2.3 Markov Chain Monte Carlo

From above questions, we could get:

$$\epsilon <= e^{-2\sum_t \gamma_t^2} \tag{15}$$

With the smallest margin of γ, we could transfer to

$$\epsilon <= e^{-2T\gamma^2} \tag{16}$$

$$T = O(\frac{-log(\epsilon)}{\gamma^2}) \tag{17}$$

2.4 implementation

For each h, we choose the classifier with smallest error in each iteration. The result is:

The classfication rule: first, if $x > 2.5$, pos; second, if $x > 3.5$, pos; third, if $x < 4.5$, pos.

t	ϵ_t	α_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$	$D_t(9)$	$err_S(H)$
1	0.222	0.626	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.222
2	0.143	0.896	0.071	0.071	0.071	0.071	0.071	0.071	0.071	0.25	0.25	0.222
3	0.125	0.973	0.042	0.042	0.042	0.042	0.25	0.25	0.042	0.146	0.146	0

Table 1: AdaBoost results

3 Gaussian Mixture Model

3.1 Show the expectation

$$E(x) = \int p(x)dx = \sum_k p(x_k)x_k \tag{18}$$

$$= \sum_k \pi_k E_k(x) \tag{19}$$

We know that for each k, the implict distribution of x is a Gaussian distribution, and $E_k(x) = \mu_k$, μ_k is the mean of the kth guassian distribution, so we have:

$$E(x) = \sum_k \pi_k \mu_k \tag{20}$$

3.2 Show the covariance

$$cov(x) = E(xx^T) - E(x)E(x)^T \tag{21}$$

$$= \sum_k \pi_k E_k(xx^T) - E(x)E(x)^T \tag{22}$$

$$= \sum_k \pi_k(\Sigma_k + \mu_k \mu_k^T) - E(X)E(x)^T \tag{23}$$

4 K-Means

4.1 Theory

4.1.1 Prove the lemma

$$\sum_x \|x-s\|^2 - \sum_x \|x-\bar{x}\|^2 = \sum_x (x^2 - 2xs + s^2) - \sum_x (x^2 - 2x\bar{x} + \bar{x}^2) \tag{24}$$

$$= \sum_x s^2 + 2x\bar{x} - \bar{x}^2 - 2\bar{x}s \tag{25}$$

Considering that $\bar{x}$ is the center of x points, we have $\sum_x x\bar{x} = |X|\,\bar{x}^2$. Therefore:

$$\sum_x s^2 + 2x\bar{x} - \bar{x}^2 - 2\bar{x}s = \sum_x s^2 + \bar{x}^2 - 2\bar{x}s = |\mathcal{X}|\,\|\bar{x}-s\|^2 \tag{26}$$

4.1.2 Prove the objective

We could first transfer $w(\mu_k, f; X)$ use n_k to represent all nodes in kth cluster:

$$w(\mu_k, f; X) = \sum_k \sum_i^n 1(f(x_i) = k) \left\| x_i - \mu_k \right\|^2 \tag{27}$$

$$= \sum_k \sum_i^{n_k} \left\| x_{ki} - \mu_k \right\|^2 \tag{28}$$

$$= \sum_k \sum_i^{n_k} \frac{1}{n_k} n_k \left\| x_{ki} - \mu_k \right\|^2 \tag{29}$$

Then we consider the form from ϕ, for the kth cluster and set the i, we have:

$$\sum_j^{n_k} \left\| x_{ki} - x_{kj} \right\|^2 = \sum_j^{n_k} x_{ki}^2 - 2x_{ki}x_{kj} + x_{ki}^2 \tag{30}$$

Using lemma1 and similar tricks in proving lemma1, then we could have:

$$n_k \left\| x_{ki} - \mu_k \right\|^2 = \sum_j \left\| x_{ki} - x_{kj} \right\|^2 \tag{31}$$

Seperately integrating into the ϕ and w, we have:

$$w(\mu_k, f; X) = \sum_k \sum_i^n 1(f(x_i) = k) \left\| x_i - \mu_k \right\|^2 \tag{32}$$

$$= \sum_k \sum_i^{n_k} \sum_j^{n_k} \frac{1}{n_k} \left\| x_{ki} - x_{kj} \right\|^2 = \phi \tag{33}$$

4.1.3 Prove the decrease of objective

Proving the decrease of objective during each iteration is quite intuive:

For step 1, suppose we have two cluster μ_1 and μ_2. Suppose μ_1 is closer. Then when we assign the point:

$$\left\| x - \mu_1 \right\|^2 < \left\| x - \mu_2 \right\|^2 \tag{34}$$

Therefore, to choose the minimize the total objective w, we should choose μ_1, which means that step 1 will decrease the objective.

For step 2, we have lemma1:

$$\sum_x \left\| x - s \right\|^2 - \sum_x \left\| x - \bar{x} \right\|^2 = |X| \left\| \bar{x} - s \right\|^2 > 0 \tag{35}$$

So we know that, chooseing any other point rather than the "average" as center would cause bigger error, which means that using the "average" will decrease the objective.

4.1.4 Prove the convergence with K

Proving the convergence of $\Omega(K)$ is quite intuitive. When we have the minimal objective, then in step 1, all the x_i would not change its assignment of cluster centain. Say we have the μ and any other μ', we would have since that Ω is the minimal objective.

$$\left\| x - \mu \right\|^2 < \left\| x - \mu' \right\|^2 \tag{36}$$

For step 2, since all the x_i would not change its assignment, then the $\mu_k = \frac{1}{n_k} \sum_i x_{ki}$ would not change. So the center would remain the same.

Therefore, in step 1 and step 2, there would be no change in the assignment and center, the Ω would not change with the increase of K

4.1.5 Prove the convergence is finite

The K-means would converge in finite numbers. We know that the center is the average of all x in a cluster. When we set the K, then the combination of K is a finite number. In the step 1, the assignment step, the choice of K is finite. Then in step 2, the change of center is set.

As mentioned from Wikipedia, the time is $O(n^{dk+1} logn)$, with k as the cluster, d as the dimension of data x and n as the data points number.

4.2 Implementation

4.2.1 and 4.2.2 Implement the K-means algorithms

The result is here:

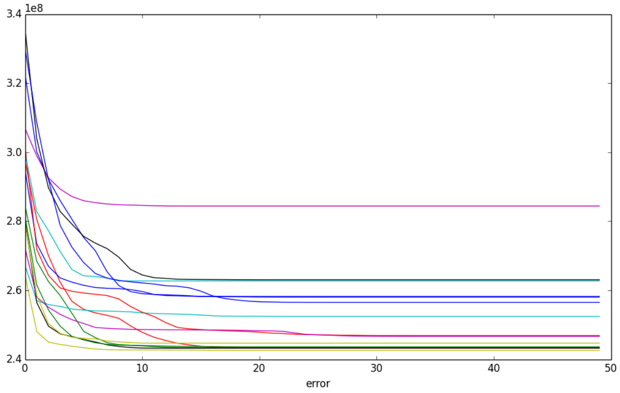

Figure 1: The error over iteration of K-means

As we could see from the figure, all the 15 iterations converge within 50 iteration. For most instances, the alogrithm would converge within 15 $\sim$ 20 iterations.

5 Collaboration

I did not collaborate with classmates in this assignment. However, I refered to some code online to finish my K-means algorithms. The link is: https://github.com/stuntgoat/kmeans.git

Basically, I used the general code structure, the update and assign function. I implemented my own data read and error plot function. Besides, I change the interface of the general kmeans function for stop conditions. I also implement the error function to get the error in each iteration for each cluster.