Tutorial 12 Unsupervised Learning

Yifan WANG

School of Data Science

2022.12.6

Final exam

- Date: Dec. 19 (Mon.)
- Time: 13:30-15:30pm(Beijing Time)
- No cheating paper!
- calculator is allowed (important!!)

Introduction

- K-means Clustering
- GMM (Gaussian Mixture Model)
- EM (Expectation Maximization)

K-means Clustering

K-means minimizes within-cluster variances

- First, choose K — the number of clusters. Then randomly put K feature vectors(centroids), to the feature space.
- Next, compute the distance from each example x to each centroid c using some metric, like the Euclidean distance. Then we assign the closest centroid to each example (like if we labeled each example with a centroid id as the label).
- For each centroid, we calculate the average feature vector of the examples labeled with it. These average feature vectors become the new locations of the centroids.
- We recompute the distance from each example to each centroid, modify the assignment and repeat the procedure until the assignments don't change after the centroid locations are recomputed
- Finally we conclude the clustering with a list of assignments of centroids IDs to the examples.

K-means Clustering

If fixing the covariance matrices Σ as the identity matrix I for all Gaussian components, EM for GMMs is reduced to a soft version of K-means

- Initialization: set K cluster centers c to random values
- Repeat until convergence (the assignments don't change):
  - Assignment(E-step): Given the cluster centers c, update the assignments r by solving the following sub-problem

    $min_r \sum_i^n \sum_k^K r_{ik}(x_i - c_k)^2, s.t. r \in 0, 1^{n \times K}, \sum_k^K r_{ik} = 1$
  - Refitting(M-step) Given the assignments r, update the centroid

    $min_c \sum_i^n \sum_k^K r_{ik}(x_i - c_k)^2$

Since the objective function J is non-convex, the coordinate descent on J is not guaranteed to converge to the global minimum

Exercise 1

A class has 10 students. They received marks for their mid-term exam as follows:

90, 86, 68, 59, 84, 80, 72, 67, 94, 79

To group the students into two tutorial groups according to their marks, we use k-means. We pick 68 as the initial centroid for Group A, and 80 for Group B, and assign the students to the two groups using Euclidean distance.

Which one is correct?
A. We will have 5 students in Group A.
B. We will have 5 students in Group B.
C. If we change the initial centroid, the clustering result by kmeans will not be changed.
D. The centroid for Group B is 85.5

Assignment (E-step)

Given the cluster centers c, update the assignments r

$$min_r \sum_i^n \sum_k^K r_{ik}(x_i - c_k)^2, \text{ s.t. } r \in 0, 1^{n \times K}, \sum_k^K r_{ik} = 1$$

the assignment for each data x_i can be solved independently, so we can ignore the $\sum_i^n$ term:

$$min_r \sum_k^K r_{ik}(x_i - c_k)^2, \text{ s.t. } r \in 0, 1^{n \times K}, \sum_k^K r_{ik} = 1$$

$$k* = argmin\{(x_i - c_k)^2\}_{k=1}^K, \text{ and } r_{ik*} = 1$$

Thus, we assign x_i to the closest cluster

Refitting (M-step)

Given the assignments r, update the cluster centers c

$$min_c \sum_i^n \sum_k^K r_{ik}(x_i - c_k)^2$$

$c_1 ... c_K$ can be optimized independently, so we can ignore the $\sum_k^K$ term:

$$min_{c_k} \sum_i^n r_{ik}(x_i - c_k)^2$$

take the second derivative, we will get:

$$\sum_i^n 2r_{ik}(x_i - c_k) = 0, \; c_k = \frac{\sum_i^n r_{ik}x_i}{\sum_i^n r_{ik}}$$

Thus, c_k is the center of the kth cluster, which is exactly same with the step of calculating the cluster center in basic K-means clustering

Gaussian Mixture Model

Before we talk about GMM, let's review Bayes Rule first

$$P(A, B) = P(A) * P(B|A)$$

$$P(A, B) = P(B) * P(A|B)$$

$$P(A|B) = P(B|A) * P(A)/P(B)$$

You may find some notations in GMM confusing, like $p(z^{(n)}|\pi)$, $p(x^{(n)}, z^{(n)}|\pi, \mu, \Sigma), p(x^{(n)}|z^{(n)}; \mu, \Sigma)$. Just recall the Bayes rule and clarify variables and parameters

- Anything following a semicolon denotes a parameter of the distribution
- We're not treating the parameters as random variables

Gaussian Mixture Model

Introduce a latent variable z, indicating which Gaussian component generates the observation x, with some probability.
Let z $\sim$ Categorical(π), where $\pi \geq 0$, $\sum_k \pi_k = 1$
Then:

$$
\begin{aligned}
p(x) = \sum_k^K p(x, z = k) &= \sum_k^K p(z = k) p(x|z = k) \\
&= \sum_k^K \pi_k N(x|\mu_k, \Sigma_k)
\end{aligned}
$$

Log-likelihood:

$$
\begin{aligned}
l(\pi, \mu, \Sigma) = lnp(X|\pi, \mu, \Sigma) &= \sum_{n=1}^N lnp(x^{(n)}|\pi, \mu, \Sigma) \\
&= \sum_{n=1}^N ln \sum_{z^{(n)}=1}^K p(x^{(n)}, z^{(n)}|\pi, \mu, \Sigma) \\
&= \sum_{n=1}^N ln \sum_{z^{(n)}=1}^K p(x^{(n)}|z^{(n)}; \mu, \Sigma) p(z^{(n)}|\pi)
\end{aligned}
$$

Gaussian Mixture Model

Because z is a random variable, so we could assume that we know $z^{(n)}$ for every $x^{(n)}$

max $l(\pi, \mu, \Sigma) = \sum_{n=1}^{N} lnp(x^{(n)}, z^{(n)}|\pi, \mu, \Sigma)$

$= \sum_{n=1}^{N}[lnp(x^{(n)}|z^{(n)}; \mu, \Sigma) + lnp(z^{(n)}|\pi)]$

with the constraint 1 - $\sum_{k=1}^{K} \pi_k = 0$

For the above constrained optimization problem, we also resort to KKT conditions based on Lagrangian function, as follows

$$L(\pi, \mu, \Sigma, \lambda) = -l(\pi, \mu, \Sigma) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$$

Gaussian Mixture Model

Take the partial derivative to get μ_k, Σ_k, and π_k
Some important steps in Σ_k:

$$\frac{\partial L(\pi, \mu, \Sigma, \lambda)}{\partial \Sigma_k} = \frac{-\partial[\sum_{n=1}^{N} \mathbb{I}_{[z^{(n)}=k]} lnp(x^{(n)}; \mu_k, \Sigma_k)]}{\partial \Sigma_k} = 0$$

$$\frac{\partial[\frac{1}{2} \sum_{n=1}^{N} \mathbb{I}_{[z^{(n)}=k]} (ln|\Sigma_k^{-1}| - (x^{(n)} - \mu_k)^T \Sigma_k^{-1} (x^{(n)} - \mu_k))]}{\partial \Sigma_k} = 0$$

Why we can define $\Lambda = \Sigma_k^{-1}$ and take the derivative of Λ?

Gaussian Mixture Model

If we know $z^{(n)}$ for every $x^{(n)}$

$$\mu_k = \frac{\sum_{n=1}^N \mathbb{I}_{[z^{(n)}=k]} x^{(n)}}{\sum_{n=1}^N \mathbb{I}_{[z^{(n)}=k]}}$$

$$\Sigma_k = \frac{\sum_{n=1}^N \mathbb{I}_{[z^{(n)}=k]} (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T}{\sum_{n=1}^N \mathbb{I}_{[z^{(n)}=k]}}$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^N \mathbb{I}_{[z^{(n)}=k]}$$

Try to derive these by yourself

Expectation Maximum

E-step: Compute probability each data point came from certain cluster, given model parameters (Assignment in K-means)
M-step: Adjust parameters of each cluster to maximize probability it would generate data it is currently responsible for (Refitting in K-means)

Expectation Maximum

Assume the observed dataset $D = {x^{(n)}}_{n=1}^{N}$, and would like to fit θ using maximum log likelihood:

$$logp(D;\theta) = \sum_{n=1}^{N} logp(x^{(n)};\theta) = \sum_{n=1}^{N} log(\sum_{z^{(n)}} p(z^{(n)}, x^{(n)};\theta))$$

How to move the summation outside the log?
introduce latent variable $z^{(n)}$, and assume that the distribution of different latent variables are independent, i.e.:

$$q(z) = \prod_{n=1}^{N} q_n(z^{(n)})$$

Expectation Maximum

We don't specify the parameter value of q(z) because we will start from one pair of x, z to utilize q(z). We have:

$$lnp(x;\theta) = E_{q(z)}[ln(\frac{p(x;\theta)\cdot q(z)}{q(z)}))]$$

Due to Bayes rule, $p(x;\theta) = p(x,z;\theta)/p(z|x;\theta)$

$$lnp(x;\theta) = E_{q(z)}[ln(\frac{p(x,z;\theta)}{q(z)})] + E_{q(z)}[ln(\frac{q(z)}{p(z|x;\theta)})]$$

Extend it to the whole dataset D:

$$lnp(D;\theta) = \sum_{n=1}^{N}[E_{q_n(z^{(n)})}[ln(\frac{p(x^{(n)},z^{(n)};\theta)}{q(z^{(n)})})]+E_{q(z^{(n)})}[ln(\frac{q(z^{(n)})}{p(z^{(n)}|x^{(n)};\theta)}]]$$

$$= L(q;\theta) + KL(q(z)||p(z|D;\theta))$$

$$\geq L(q;\theta)$$

Expectation Maximum

We try to get $lnp(D;\theta)$ by maximum the lower bound $L(q;\theta)$. The EM algorithm alternates between making the bound tight at the current parameter values and then optimizing the lower bound

E-step

E step: given θ, update q(z) - making the bound tight at the current parameter values:

$$max_{q(z)} L(q;\theta) \equiv \max_{q(z)} \sum_{n=1}^{N} E_{q_n(z^{(n)})}[ln(\frac{p(x^{(n)}, z^{(n)}; \theta)}{q_n(z^{(n)})})]$$

We can get the optimal solution is:

$$q_n^*(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta)$$

At that time, the gap is 0, which means that:

$$logp(D; \theta^{old}) = L(q; \theta^{old})$$

M-step

M step: given q(z), update θ - optimizing the lower bound:

$$\theta^{new} = argmax_{\theta} L(q;\theta)$$

substitute in $q_n^*(z^{(n)}) = p(z^{(n)}|x^{(n)};\theta^{old})$

$$\theta^{new} = argmax_{\theta} \sum_{n=1}^{N} E_{p(z^{(n)}|x^{(n)};\theta^{old})}[logp(z^{(n)}, x^{(n)};\theta)]$$

Try to derive all the materials in lecture slides by yourself

Good luck for your final exam!