Introduction to Machine Learning

Lecture 9 - Unsupervised Learning 2

Guang Bing Yang, PhD

Clustering: Gaussian Mixtures

- Introduction to Gaussian Mixture Models.
- Introduction to EM for Gaussian Mixtures.
- EM for K-means algorithm

What is Gaussian Mixture

- Gaussian mixture model is a simple linear superposition of Gaussian distributions.
- It arms to provide a richer class of density models than the single one.
- The The mixture of Gaussian:
  - $p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$
- It brings in a latent variable z, and gives a joint probability:
  - $p(x, z) = p(z)p(x \mid z)$, where z is a 1-to-K coding latent variable.

Gaussian Mixture

- $p(z_k = 1) = \pi_k$
- constraints: $0 \leq \pi_k \leq 1$, and $\sum_k \pi_k = 1$
- $p(x \mid z_k = 1) = \mathbb{N}(x \mid \mu_k, \Sigma_k)$
- $p(x \mid z) = \prod_k \mathbb{N}(x \mid \mu_k, \Sigma_k)^{z_k}$
- Marginal distribution:
  - $p(x) = \sum_z p(x, z) = \sum_z p(z)p(x \mid z) = \sum_k \pi_k N(x \mid \mu_k, \Sigma_k)$

Gaussian Mixture

- The use of joint probability p(x,z), leads to significant simplifications.
- Posterior or responsibility of component k to observations X,

$$\gamma(z_k) = p(z_k = 1 \mid x)$$

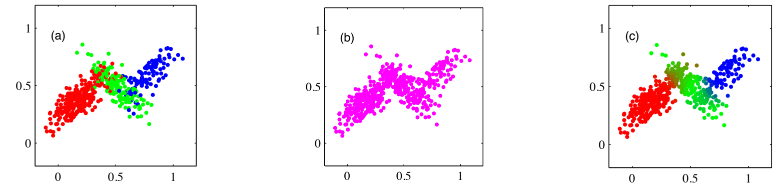
$$= \frac{p(z_k = 1)p(x \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x \mid z_j = 1)}$$

- $$= \frac{\pi_k N(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x \mid \mu_j, \Sigma_j)}$$
- π_k is the prior probability of z_k, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x.

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Gaussian Mixture

- Generate random samples with ancestral sampling:
- First generate zˆ from p(z)
- Second generate a value for x from p(x|zˆ)
- 500 points drawn from the mixture of 3 Gaussians shown on below

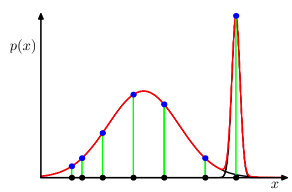


p(z)p(x|z) p(x) Responsibilities: $\gamma(z_{nk})$

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Gaussian Mixture: Maximum Likelihood

- Log likelihood:
- $\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$
- Singularity is a significant issue, when a mixture component collapse on a data point.
- Identifiability is another issue for a ML solution in a K-component mixture—there are K! equivalent solutions.



Singularity in the likelihood function

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Gaussian Mixture : EM for Gaussian mixtures

- EM stands for expectation-maximization
- It is a good approach for finding maximum likelihood solutions.
- Set the derivatives of $\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$ with respect to the mean μ_k of the Gaussian components to zero, we obtain:
- $$0 = -\sum_{n=1}^{N} \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} \Sigma_k^{-1}(x_n - \mu_k)$$

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Gaussian Mixture : EM for Gaussian mixtures

- For μ_k
  - $\mu_k = \dfrac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n,$
  - Where $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$
- For Σ_k:
  - $\Sigma_k = \dfrac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T$

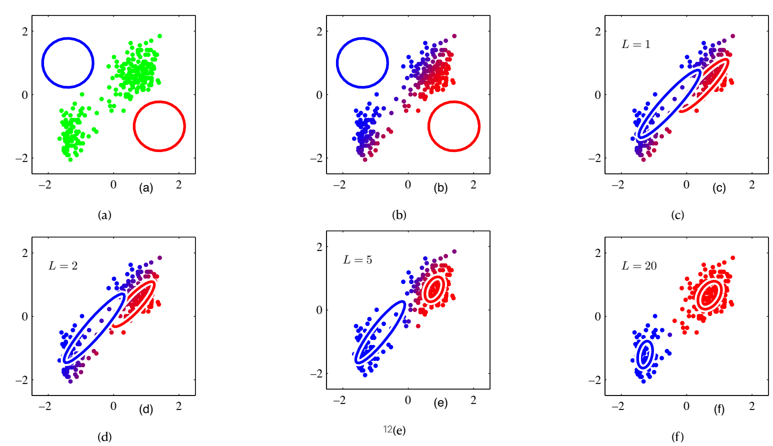
Gaussian Mixture : EM for Gaussian mixtures

- For the π_k
- Based on the constraint: $\sum_k \pi_k = 1$
- The Lagrange multiplier and maximizing the following quantity:
- $\ln p(X|\pi,\mu,\Sigma) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$
- with gives: $0 = \sum_{n=1}^{N}\dfrac{N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} + \lambda$
  - Then, $\pi_k = \dfrac{N_k}{N}$, and $N_k = \sum_k \gamma(z_k)$

Gaussian Mixture : Example of EM for Gaussian mixtures

- $\gamma(z_k)$ relies on parameters, there is no closed form solution for it.
- A simple iterative scheme can be applied for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_k)$ and updating the parameters $\{\mu_k, \Sigma_k, \pi_k\}$.
- For example, there is an instance of the EM algorithm for the particular case of the Gaussian mixture model.
  - First, choose the initial values for the means, covariances, and mixing coefficients
  - Then, alternate between the following two updates E step and M steps

Gaussian Mixture : Example of EM for Gaussian mixtures



Gaussian Mixture : Example of EM for Gaussian mixtures

- However, this approach needs more iterations to converge than the K-means algorithm, and each cycle requires more computation.
- Normally, use k-means to get initial parameters rather than starting from arbitrary values of the initial settings.

Gaussian Mixture : Summary of EM for Gaussian mixtures

- Initialize the means μ_k, covariance Σ_k and mixing coefficients π_k
- evaluate log-likelihood
- E-step: evaluate the responsibilities $\gamma(z_k)$,
- M-step:
  - $\mu_k^{new} = \dfrac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$

An Alternative View of EM

- Let X observed data, Z latent variables, θ parameters.
- Goal: maximize marginal log-likelihood of observed data
  - $\ln p(X|\theta) = \ln \left\{ \sum_{z} p(X, Z|\theta) \right\}$
  - Optimization problematic due to log-sum.
  - Assume straightforward maximization for complete data: $\ln p(X, Z|\theta)$

An Alternative View of EM

- Latent Z is known only through $p(Z|X, \theta)$.
- Let us consider expectation of complete data log-likelihood.
- Initialization: Choose initial set of parameters θ^{old}.
- E-step: use current parameters θ^{old} to compute $p(Z|X, \theta^{old})$
- to find expected complete-data log-likelihood for general θ:
  - $Q(\theta, \theta^{old}) = \sum_{z} p(Z|X, \theta^{old}) \ln p(X, Z|\theta)$

An Alternative View of EM

- In the M step, determine the revised parameter estimate θ^{new} by maximizing this function:
  - $\theta^{new} = \text{argmax}_{\theta} Q(\theta, \theta^{old})$
- Check convergence: if not converged, let $\theta^{old} \leftarrow \theta^{new}$, and return to step E, repeat.

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

An Alternative View of EM: Gaussian Mixture Revisited

- Recall the maximize likelihood of the Gaussian mixture is given as:
- $\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$
- which is computed using the observed data X.
- But it is more complex and difficult than a single Gaussian due to the presence of summation over k inside the logarithm.
- Use complete-data (log-)likelihood, and expectation given as:

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

An Alternative View of EM: Gaussian Mixture Revisited

- $p(X, Z \mid \theta) = \prod_{k=1}^{K} \pi_k^{z_k} \mathbb{N}(x_n \mid \mu_k, \Sigma_k)^{z_k}$, taking the logarithm, obtain:
- $\ln p(X, Z \mid \theta) = \sum_{k=1}^{K} z_k \{\ln \pi_k + \ln \mathbb{N}(x_n \mid \mu_k, \Sigma_k)\}$
- The logarithm now directly acts on the normal distribution, which is tractable.
- Since variable Z is unknown, so consider the expectation. Then, obtain:
- $Q(\theta) = \mathbb{E}_z[\ln p(x, z \mid \theta)] = \sum_{k=1}^{K} \gamma(z_k)\{\ln \pi_k + \ln \mathbb{N}(x; \mu_k, \Sigma_k)\}$

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Using Gaussian Mixture for Clustering

- Two main applications for mixture models-- as a black-box density model p(x) and clustering.
- As a black-box, a kind of mixture model can be applied for
  - data compression,
  - outlier detection,
  - creating generative classifiers.
- more common, used for clustering by:
  - first, fit the mixture model
  - second, compute $p(z_k \mid x, \theta)$ -- the probability for point x belongs to cluster k.
- This is called soft-clustering. K-means is a kind of hard-clustering.

yguangbing@gmail.com, Guang.B@chula.ac.th Mar 26, 2021

Mixture of Experts

- The goal of mixture of experts is to use clustering to create discriminative models for classification and regression.
- Each sub-model is considered to be an "expert" in a certain region of input space.
- Use responsibilities $p(z_i = k \mid x_i, \theta)$ as the gating function to decide which expert to use, which depends on the input data.
- Any model can be used as an "expert", for example a linear regression model can be an expert.

Recap

- Gaussian mixture model is a simple linear superposition of Gaussian distributions
- It brings in a latent variable z, and gives a joint probability.
- EM stands for expectation-maximization
- It is a good approach for finding maximum likelihood solutions
- In E-step: evaluate the responsibilities $\gamma(z_k)$.
- In M-step: update parameters by maximizing its corresponding function.
- Applications of mixture models include a black-box and clustering.

Assignment 4

- Assignment 4 worth 15%, and is about a clustering Python programming using Scikit-learn framework. It was also posted in MS Teams Assignments.
- Copy and download my Colab from Chula G drive to your Google drive (Important note: Don't modify my Colab notebook, otherwise other classmates will see your work.)
- Working on your copy of the Colab notebook. Don't forget to add your name and student id in it.
- After finishing it, share it with me (only me, do not share your work with others.)
- All programming exercises MUST be running correctly in Colab without any errors and exceptions. If your code cannot run at all, and I cannot see any kind of outputs, you receive no grade points for that part.
- Before you submit your Colab notebook, make sure to leave the outputs (results) of the functions in the notebook. I ONLY review the outputs of your functions or the final results.
- **The assignment due at Apr 9th@23:59 (your local time), 2021. It is an individual assignment. Please no late due. Any late due assignment will not be accepted.**
- **Make sure your share your Colab notebook having proper access permission to me to review your work.**
- **If I cannot view your work due to the permission issue, I will send you an email to remind you to re-assign me correct access permissions to your Colab notebook. After 12 hours start from the time that I sent you my reminding email, if I still cannot access your Colab notebook, no evaluation for this assignment will be given.**
- **I will start evaluating your work at Apr 10th, and try my best to give you feedback 1 week after.**

Questions?