Introduction to Machine Learning

Lecture 10 - Unsupervised Learning: EM in General

Guang Bing Yang, PhD

EM in General

- Introduction to EM in General
- EM for K-means algorithm
- General EM revisit Gaussian mixture
- EM for mixture of Bernoulli distributions

The General EM Algorithm

- In general, the expectation maximization algorithm, or EM algorithm, is a technique for finding maximum likelihood solutions for probabilistic models having latent variables (Dempster et al., 1977; McLachlan and Krishnan, 1997).
- The goal is to maximize the likelihood function $p(X|\theta)$ with respect to θ.
  - $p(X|\theta) = \sum_{z} p(X, Z|\theta)$ assume Z is discrete
  - Direct optimize $p(X|\theta)$ is difficult, but
  - optimize the complete-data likelihood function $p(X, Z|\theta)$ is easier

Reference: Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B 39(1), 1–38. McLachlan, G. J. and T. Krishnan (1997). The EM Algorithm and its Extensions. Wiley.

The General EM Algorithm

- Initialze parameters θ^{old}
- In the E-step: compute posterior: $p(Z|X, \theta^{old})$ w.r.t. the latent variable Z
- In the M-step: search the new estimate of parameters θ^{new}, given
- $Q(\theta, \theta^{old}) = \sum_{Z} p(Z|X, \theta^{old}) \ \ln p(X, Z|\theta)$
- evaluate the convergence of either log-likelihood or the parameter values:
  - $\theta^{new} \leftarrow \theta^{old}$,
  - iterate till converged or the difference of $\theta^{new} - \theta^{old} \leq$ threshold, or over the loop limit

The General EM Algorithm

- The expected complete data log likelihood, also called **auxiliary function**, can be described as:

$$Q(\theta, \theta^{old}) = \mathbb{E}\left[\sum_i \log p(x_i, z_i | \theta)\right] = \sum_i \mathbb{E}\left[\log\left[\prod_{k=1}^{K} (\pi_k p(x_i, z_i | \theta_k))^{I(z_i=k)}\right]\right]$$

- $$= \sum_i \sum_k \mathbb{E}[I(z_i = k)]\log[\pi_k p(x_i, z_i | \theta_k)] = \sum_i \sum_k p(z_i = k | x_i, \theta^{old})\log[\pi_k p(x_i, z_i | \theta_k)]$$
$$= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(x_i | \theta_k)$$

- where $r_{ik} = p(z_i = k | x_i, \theta^{old})$, the posterior or responsibility for cluster k takes for data x_i.
  - E-step: $r_{ik} = \frac{\pi_k p(x_i | \theta_k^{old})}{\sum_{k'} \pi_{k'} p(x_i | \theta_{k'}^{old})}$,
  - M-step: $\pi_k = \frac{1}{N} \sum_i r_{ik} = \frac{r_k}{N}$, where $r_k = \sum_i r_{ik}$
  - $\mu_k = \frac{\sum_i r_{ik} x_i}{r_k}$, $\Sigma_k = \frac{\sum_i r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{r_k} = \frac{\sum_i r_{ik} x_i x_i^t}{r_k} - \mu_k \mu_k^T$

General EM: Variational Bound

- Given a joint distribution $p(X, Z | \theta)$ over observed variables X and hidden variables Z, the goal is to
- to maximize the likelihood function $p(X | \theta)$ with respect to θ.
  - $p(X | \theta) = \sum_z p(X, Z | \theta)$
- assume Z is discrete (change the summation to integral if Z is continuous, others are the same)
- For any distribution q(Z), there is following variational lower bound:
  - $\ln p(X | \theta) = \ln \sum_Z p(X, Z | \theta) = \ln \sum_Z q(Z) \frac{p(X, Z | \theta)}{q(Z)}$
  - Logarithm is concave, so Jensen's inequality exists: so, $\ln p(X | \theta) \geq \sum_Z q(Z) \ln \frac{p(X, Z | \theta)}{q(Z)}$

General EM: Variational Bound

$$\ln p(X | \theta) = \ln \sum_Z p(X, Z | \theta) = \ln \sum_Z q(Z) \frac{p(X, Z | \theta)}{q(Z)}$$
$$\geq \sum_Z q(Z) \ln \frac{p(X, Z | \theta)}{q(Z)}$$

- $$= \sum_Z q(Z) \ln p(X, Z | \theta) + \sum_Z q(Z) \ln \frac{1}{q(Z)}$$
- $$= E_{q(Z)}[\ln p(X, Z | \theta)] + \mathbb{H}(q(Z)) = \mathbb{L}(q, \theta)$$

General EM: Variational Bound

- There are two components in the log likelihood function:
- $\ln p(X | \theta) \geq E_{q(Z)}[\ln p(X, Z | \theta)] + \mathbb{H}(q(Z)) = \mathbb{L}(q, \theta)$
- The first part is the Expected complete log-likelihood, the second part is the Entropy of the distribution q(Z).
- $\mathbb{L}(q, \theta)$ is the variational lower-bound.
- For a discrete random variable Z, the entropy is defined as:
  - $\mathbb{H}(p) = -\sum_i p(z_i) \log p(z_i)$, or $\mathbb{H}(p) = -\int p(z) \log p(z) dz$ for continuous random variables

General EM: Variational Bound

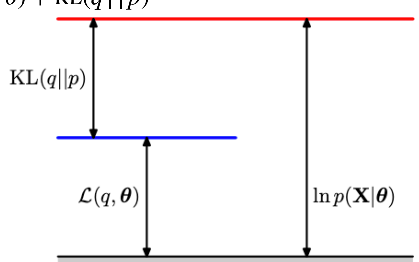
- Having:
  - $\ln p(X|\theta) \geq E_{q(Z)}[\ln p(X,Z|\theta)] + \mathbb{H}(q(Z)) = \mathbb{L}(q,\theta)$, and
  - $\ln p(X|\theta) = \mathbb{L}(q,\theta) + \mathrm{KL}(q||p)$, where
    - $\mathbb{L}(q,\theta) = \sum_Z q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}$, the lower bound
    - $\mathrm{KL}(q||p) = -\sum_Z q(Z) \ln \frac{p(Z|X,\theta)}{q(Z)}$, a relative entropy
- Since $\ln p(X,Z|\theta) = \ln p(Z|X,\theta) + \ln p(X|\theta)$, and substitute the $\mathbb{L}(q,\theta)$

General EM: Variational Bound

- Note that variational bound becomes tight iff $q(Z) = p(Z|X,\theta)$.
- In other words the distribution q(Z) is equal to the true posterior distribution over the latent variables, so that $\mathrm{KL}(q||p) = 0$.
- As $\mathrm{KL}(q||p) = 0$, it immediately follows that:
  - $\ln p(X|\theta) \geq \mathbb{L}(q,\theta)$,
  - which is also showed using Jensen's inequality,

General EM: Decomposition of q(Z)

- To illustrate the decomposition of the distribution q(Z):
- $\ln p(X|\theta) = \mathbb{L}(q,\theta) + \mathrm{KL}(q||p)$

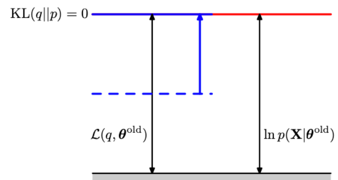


General EM: Summary

- uses the decomposition to define the EM algorithm, and
- shows that it maximizes the log-likelihood function:
  - $\ln p(X|\theta) = \mathbb{L}(q,\theta) + \mathrm{KL}(q||p)$
- In the E-step, the lower bound $\mathbb{L}(q,\theta)$ is maximized w.r.t. the distribution q with fixed parameters θ
- In the M-step, the lower bound $\mathbb{L}(q,\theta)$ is maximized w.r.t. the parameters θ with the distribution q fixed.
- These steps increase the corresponding log-likelihood

General EM: E-step

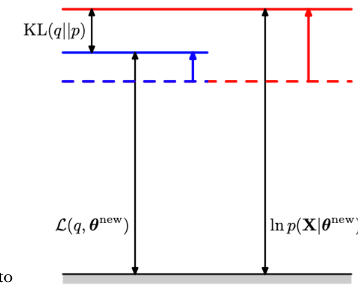
- Let the current value of parameters as θ^{old}
- In the E-step, maximize the lower bound w.r.t. q with θ^{old} fixed:
  - $\mathbb{L}(q, \theta^{old}) = \ln p(X \mid \theta^{old}) - \mathrm{KL}(q \mid\mid p)$
  - θ^{old} does not depend on q
- the lower bound $\mathbb{L}(q, \theta)$ is maximized when KL = 0
- In other words, $q(Z) = p(Z \mid X, \theta^{old})$
- The lower bound becomes equal to the log-likelihood



The E-step of the EM algorithm.

General EM: M-step

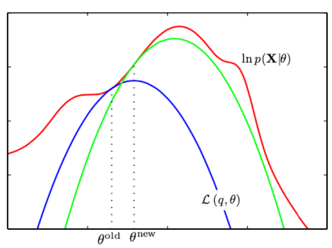
- the lower bound is maximized w.r.t. parameters θ with q fixed.
- In the E-step, maximize the lower bound w.r.t. q with θ^{old} fixed:
  $\mathbb{L}(q, \theta) = \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$
  - $+ \sum_Z p(Z \mid X, \theta^{old}) \ln \dfrac{1}{p(Z \mid X, \theta^{old})}$
  - $\mathbb{L}(q, \theta) = Q(\theta, \theta^{old}) + \text{const}$
  - the last part of the $\mathbb{L}(q, \theta)$ does not depend on q
- Hence the M-step maximizes the expected complete log-likelihood
  - $\theta^{new} = \mathrm{argmax}_{\theta} Q(\theta, \theta^{old})$
- Because the KL is non-negative, this makes the log-likelihood $p(X \mid \theta)$ to increase by at least as much as the lower bound does.



The M-step of EM algorithm

General EM: Bound Optimization

- The EM algorithm belongs to the general class of bound optimization methods:
- At each step, we compute:
  - E-step: a lower bound on the log-likelihood function for the current parameter values. The bound is concave with unique global optimum.
  - M-step: maximize the lower-bound to obtain the new parameter values.



The EM algorithm maximizes this bound to obtain the new parameter values

General EM: Extensions

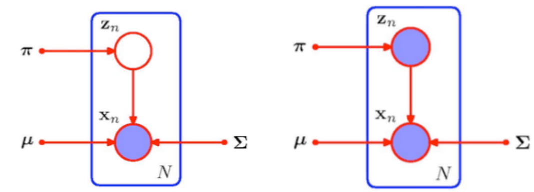
- For some complex cases, either the E-step or the M-step or both remain intractable
- Two possible approaches:
  - The generalized EM (GEM) — deals with the intractable in the M-step
  - generalized the E-step by performing a partial optimization of the lower-bound w.r.t. q
  - In GEM, using nonlinear optimization, conjugate gradient, etc to change parameters so as to increase its value.
  - use an incremental form of EM, in which at each EM step only one data point is processed at a time
  - In the E-step, instead of recomputing the responsibilities for all the data points, we just re-evaluate the responsibilities for one data point, and proceed with the M-step

Maximizing the Posterior using EM

- There is a way to use EM to maximize the posterior $p(\theta|X)$ for models having the prior defined as $p(\theta)$
- Because: $\ln p(\theta|X) = \ln p(X|\theta) + \ln p(\theta) - \ln p(X)$
- Decomposing the log-likelihood into lower-bound and KL terms:
  - $\ln p(X|\theta) = \mathbb{L}(q, \theta) + \mathrm{KL}(q||p)$
- $\ln p(\theta|X) = \mathbb{L}(q, \theta) + \mathrm{KL}(q||p) + \ln p(\theta) - \ln p(X)$
  - where $\ln p(X)$ is a constant.
- The E-step is the same as for the standard EM algorithm
- The M-step equations are modified through introduction of the prior term, which typically amounts to only a small modification to the standard ML M-step equations.

General EM: Gaussian Mixture Revisited

- Recall the maximize likelihood of the Gaussian mixture is given as:
- $\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}$



$\{\mathbf{X}\}$ -- incomplete dataset. $\{\mathbf{X}, \mathbf{Z}\}$ -- complete dataset.

General EM: Gaussian Mixture Revisited

- Use complete-data (log-)likelihood, and expectation given as:
- $p(X, Z|\pi, \mu, \Sigma) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k \mathbb{N}(x|\mu_k, \Sigma_k)]^{z_{nk}}$, taking the logarithm, obtain:
- $\ln p(X, Z|\pi, \mu, \Sigma) = \sum_{k=1}^{K}[\sum_{n=1}^{N} z_{nk}\{\ln \pi_k + \ln \mathbb{N}(x|\mu_k, \Sigma_k)\}$
- Maximizing w.r.t. mixing proportions given: $\pi_k = \frac{1}{N}\sum_{n=1}^{N} z_{nk}$
- Similarly for the means and covariances.

General EM: K-means Revisited

- Consider a Gaussian mixture model in which covariances are shared and are given by ϵ.
- $p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp[-\frac{1}{2\epsilon}||x - \mu_k||^2]$
- Consider EM algorithm for a mixture of K Gaussians, in which let ϵ as a fixed constant. The posterior responsibilities take form:
  - $\gamma(z_{nk}) = \frac{\pi_k \exp(-||x_n - \mu_k||^2/2\epsilon)}{\sum_{j=1}^{K} \pi_j \exp(-||x_n - \mu_j||^2/2\epsilon)}$
  - Consider the limit $\epsilon \to 0$
  - $\gamma(z_{nk}) \to r_{nk}$, while $r_{nk} = 1$ if $k = \mathrm{argmax}_j ||x_n - \mu_j||^2$ otherwise$r_{nk} = 0$.

Mixture of Bernoulli Distributions

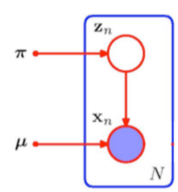
- Let's look at mixture of discrete binary variables described by Bernoulli distributions.
- Consider a set of binary random variables xi, i=1,...,D, each of which is governed by a Bernoulli distribution with μ_i:
  - $p(x|\mu) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$
- The mean and covariance of this distribution are:
  - $\mathbb{E}[x] = \mu$, $\mathrm{cov}[x] = \mathrm{diag}(\mu_i(1-\mu_i))$

Mixture of Bernoulli Distributions

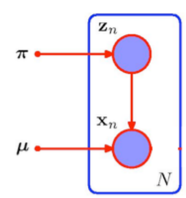
- Given a finite (K) mixture of Bernoulli distributions:
  - $p(x|\pi,\mu) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$
  - $p(x|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$
- The mean and covariance of this mixture distribution are:
  - $\mathbb{E}[x] = \sum_{k=1}^{K} \pi_k \mu_k$, and $\mathrm{cov}[x] = \sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T) - \mathbb{E}[x]\mathbb{E}[x]^T$,
  - where $\Sigma_k = \mathrm{diag}(\mu_{ki}(1-\mu_{ki}))$
- The covariance matrix is no longer diagonal, so the mixture distribution can capture correlations between the variables, unlike a single Bernoulli distribution.

Mixture of Bernoulli Distributions: Maximum Likelihood

- Given a dataset X, the log-likelihood:
  - $\ln p(x|\pi,\mu) = \sum_{n=1}^{N} \ln[\sum_{k=1}^{K} \pi_k p(x|\mu_k)]$
  - this is intractable, need to apply EM algorithm for maximizing this log-likelihood function



$\{\mathbf{X}\}$ -- incomplete dataset.



$\{\mathbf{X}, \mathbf{Z}\}$ -- complete dataset.

Mixture of Bernoulli Distributions: Maximum Likelihood

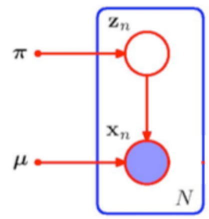
- Consider the complete log-likelihood:
  - $p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}$, and $p(x|z,\mu) = \prod_{k=1}^{K} p(x|\mu_k)^{z_k}$
  - the complete log-likelihood given as:
    - $\ln p(X,Z|\pi,\mu) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{nk}[\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln\, u_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]$
  - The expected complete-data log-likelihood:
    - $\mathbb{E}_Z[\ln p(X,Z|\pi,\mu)] = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})[\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln\, u_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]$,
    - where $\mathbb{E}[z_{nk}] = \gamma(z_{nk})$

Mixture of Bernoulli Distributions: E-step

- Similar to the mixture of Gaussians, in the E-step, using Bayes' rule to evaluate responsibilities:

$$\mathbb{E}[z_{nk}] = \frac{\sum_{Z_n} z_{nk}\prod_{k}[\pi_{k'}p(x_n|\mu_{k'})]^{z_{nk'}}}{\sum_{Z_n}\prod_{j}[\pi_j p(x_n|\mu_j)]^{z_{nj}}}$$

- $$= \frac{\pi_k p(x_n|\mu_k}{\sum_{j=1}^{K}\pi_j p(x_n|\mu_j)} = \gamma(z_{nk})$$

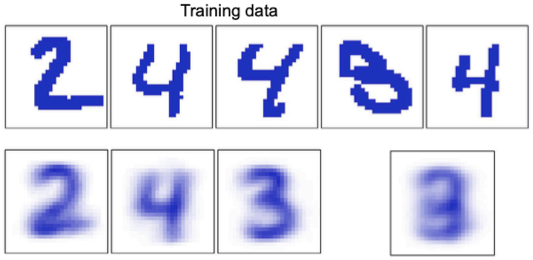


Mixture of Bernoulli Distributions: M-step

- The expected complete-data log-likelihood:
  - $\mathbb{E}_Z[\ln p(X,Z|\pi,\mu)] = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})[\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln u_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]$
- Maximizing the expected complete-data log-likelihood:
  - $\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$, $\pi_k = \frac{N_k}{N}$, and $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$,
  - where Nk is the effective number of data points associated with component k.
- Note that the mean of component k is equal to the weighted mean of the data, with weights given by the responsibilities that component k takes for explaining the data points.

Mixture of Bernoulli Distributions: Example

Training data



Learned μ_k for the first three components.

A single multinomial Bernoulli distribution fit to the full data.

Recap

- The general EM algorithm, is a technique for finding maximum likelihood solutions for probabilistic models having latent variables (Dempster et al., 1977; McLachlan and Krishnan, 1977)
- The goal is to maximize the likelihood function $p(X|\theta)$ with respect to θ.
- But easy to use complete data, complete log-likelihood by given a joint distribution $p(X,Z|\theta)$ over observed variables X and hidden variables Z
- By decompose the distribution q(Z), applying for the EM to maximize the log-likelihood over q and θ in a two-stage process.
- Iterate the E-step and M-step many times until converged, then the parameters are optimized values.

Questions?