Lecture 17 : 3/22/17

Expectation-Maximization

① EM is a framework for deriving local optimization algorithms

② Assumes a model of the form:

$$p(\vec{x}|\vec{\theta}) = \sum_z p(\vec{x}, z|\vec{\theta})$$

parameters

latent variable(s)

③ Locally optimizes the marginal likelihood:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(\vec{x}^{(i)}|\vec{\theta})$$

(b/c iid assumption)

EM: High Level

Alternating Optimization Alg.

$\boxed{\text{E-step}}$ Estimate some "unobserved" (latent) data from "observed" data and our current params.

z, x, θ

$\boxed{\text{M-step}}$ Find the MLE parameters, using the complete likelihood

θ, (x, z)

⇒ Each EM step increases $\ell(\theta)$

⇒ Converges to a local maximum of $\ell(\theta)$

EM: Detailed Version

Def: Complete log-likelihood

$$\ell_c(\theta) = \sum_{i=1}^{N} \log p(x^{(i)}, z^{(i)}|\theta)$$

Def: Expected Complete log-likelihood

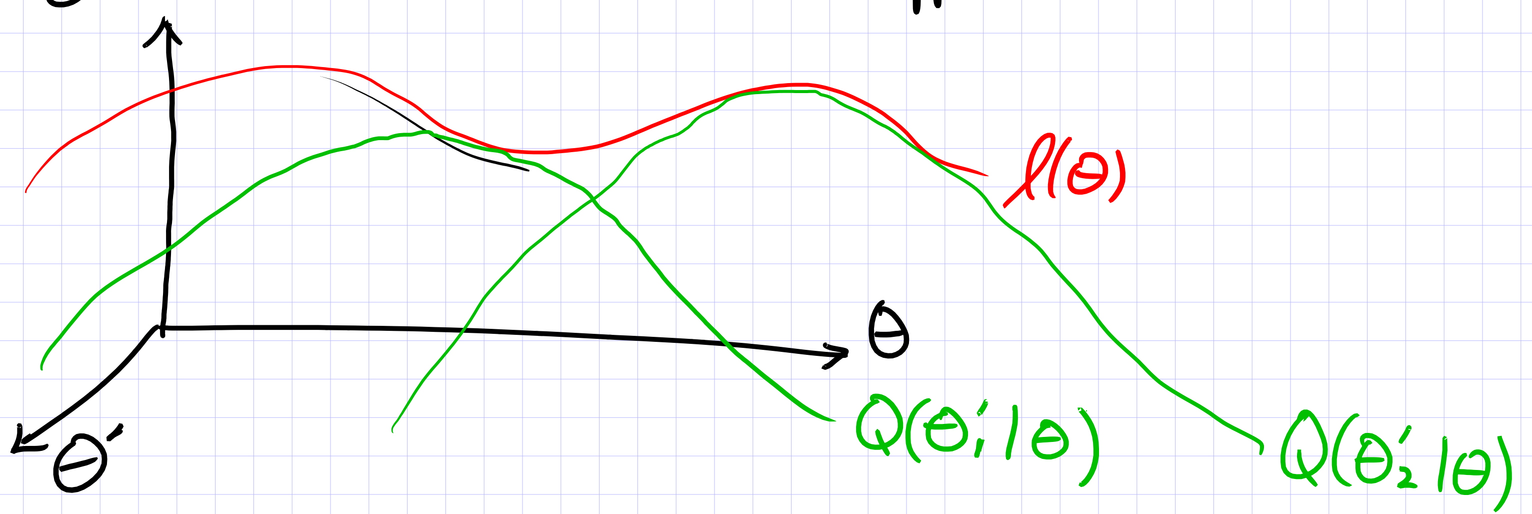
$$Q(\theta'|\theta) = \sum_{i=1}^{N} \mathbb{E}_{p(z|x^{(i)},\theta)}\left[\log p(x^{(i)}, z|\theta')\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \underbrace{p(z^{(i)}=k|x^{(i)},\theta)}_{\text{How much our model believes that latent var. has value } k.} \underbrace{\log p(x^{(i)}, z=k|\theta')}_{\text{the log-likelihood if it did equal } k}$$

Thm: $Q(\theta'|\theta) \leq \ell(\theta) \quad \forall \theta'$

↖ this is a lower bound for the marginal likelihood.

★EM is just Block Coordinate Descent applied to $Q(\theta'|\theta)$



EM Algorithm

① Randomly initialize θ

② Iterate until convergence:

$\boxed{\text{E-step}}$ Compute "expected" $p(z^{(i)}|x^{(i)},\theta)$ using current parameters θ

$\boxed{\text{M-step}}$ Update $\theta \leftarrow \underset{\theta'}{\operatorname{argmax}}\, Q(\theta'|\theta)$

new value ← current value

$$\left[\theta \leftarrow \underset{\theta}{\operatorname{argmax}}\, Q(\theta'|\theta)\right]$$

$$\left[\theta' \leftarrow \underset{\theta'}{\operatorname{argmax}}\, Q(\theta'|\theta)\right]$$

★ Each step increases $Q(\theta'|\theta)$ which in turn increases $\ell(\theta)$ marginal l.l.

Ex: EM for GMM

★ Simplifying assumptions:

① Identity covariance: $\Sigma_k = I \quad \forall k \in \{1,\dots,K\}$

② Equiprobable clusters: $\phi_k = \frac{1}{K} \quad \forall k$

$\Rightarrow$ only learning "cluster centers"/means $\vec{\mu}_k \quad \forall k$

$$\Rightarrow p(z=k|\vec{x},\mu) = \frac{\exp\{-\frac{1}{2}\|\vec{x}-\vec{\mu}_k\|^2\}}{\sum_{j=1}^{K}\exp\{-\frac{1}{2}\|\vec{x}-\vec{\mu}_j\|^2\}} \quad \text{for } \Sigma_k = I,\ \phi_k = \tfrac{1}{k}$$

EM for GMM

① Randomly initialize $\vec{\mu}_1,\dots,\vec{\mu}_K$

② Iterate until conv.

$\boxed{\text{E-step}}$ Compute $q_j^{(i)} \triangleq p(z^{(i)}=j|x^{(i)},\mu) \quad \forall i,j$

$\boxed{\text{M-step}}$ Update params. to maximize $Q(\mu'|\mu)$

$$\vec{\mu}_j \leftarrow \underbrace{\frac{\sum_{i=1}^{N} q_j^{(i)}\vec{x}^{(i)}}{\sum_{i=1}^{N} q_j^{(i)}}}_{\text{Average of all points weighted by how likely each point came from Gaussian } j}$$

Current model's estimate of prob. that $x^{(i)}$ came from Gaussian j

K-Means

new name for cluster centers

① Randomly initialize $\vec{\mu}_1,\dots,\vec{\mu}_K$

② Iterate until conv.

$\boxed{A}$ Compute

$$q_j^{(i)} \triangleq \begin{cases} 1 & \text{if } j = \underset{k}{\operatorname{argmin}}\|x^{(i)}-\mu_k\|^2 \\ 0 & \text{otherwise} \end{cases}$$

$\boxed{B}$ Update the parameters (cluster centers) to be the avg. of points in cluster

$$\vec{\mu}_j \leftarrow \frac{\sum_{i=1}^{N} q_j^{(i)}\vec{x}^{(i)}}{\sum_{i=1}^{N} q_j^{(i)}}$$

Equivalent Computation:

$$q_j^{(i)} = \begin{cases} 1 & \text{if } j = \underset{k}{\operatorname{argmax}}\, p(z^{(i)}=k|x^{(i)},\mu) \\ 0 & \text{otherwise} \end{cases}$$

where dist. is the GMM posterior

Connections btwn. EM for GMM and K-Means

With our simplifying assumptions

1. K-Means is EM for GMM where $\sigma^2 \to 0$ and $\Sigma = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix}$
2. K-Means is the result of Block Coordinate Descent applied to a different objective for GMM (Hard EM) $\leftarrow$ not covered

Data for PCA

$D = \{\vec{x}^{(i)}\}_{i=1}^{N} \quad \vec{x}^{(i)} \in \mathbb{R}^M$

$$X = \begin{bmatrix} — (x^{(1)})^T — \\ \vdots \\ — (x^{(N)})^T — \end{bmatrix}$$

<u>Assumption #1:</u> the data is "centered" (sample mean)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \vec{x}^{(i)} = 0 \text{ vector}$$

<u>Assumption #2</u>: the sample variance of each axis is 1

$$\sigma_m^2 = \frac{1}{N}\sum_{i=1}^{N} (x_m^{(i)})^2 = 1$$

<u>Q</u>: What if the data doesn't match these?

① subtract off μ

② divide each component by σ_m

<u>Def</u>: the sample covariance Σ is an $M \times M$ matrix

$$\Sigma_{jk} = \frac{1}{N}\sum_{i=1}^{N} (x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$

for centered data

$$\Sigma = \frac{1}{N} X^T X$$

Definition of PCA

- Given K vectors $\vec{v}_1, \dots, \vec{v}_K$ where $\vec{v}_k \in \mathbb{R}^M$, the projection of a vector a vector $\vec{x}^{(i)}$ into lower K-dimensional space is $\vec{u}^{(i)} \in \mathbb{R}^K$

$$\vec{u}^{(i)} \triangleq \begin{bmatrix} \vec{v}_1^T \vec{x}^{(i)} \\ \vec{v}_2^T \vec{x}^{(i)} \\ \vdots \\ \vec{v}_K^T \vec{x}^{(i)} \end{bmatrix}$$

- <u>Def</u>: PCA repeatedly chooses a next vector $\vec{v}_j$ that minimizes the reconstruction error s.t. $\vec{v}_1, \dots, \vec{v}_{j-1}$ are orthogonal to $\vec{v}_j$

<u>Recall</u>: two vectors $\vec{a}$ and $\vec{b}$ are orthogonal if $\vec{a}^T\vec{b} = 0$

$\Rightarrow$ dimensions in K-space are uncorrelated

- <u>Question</u>: How do we find the vectors?

Projection