Machine Learning for Data Science (CS4786)
Lecture 16

EM Algorithm

Mar 24, 2015

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2015sp/

Examples

- Gaussian Mixture Model
  - Each θ consists of mixture distribution $\pi = (\pi_1, \ldots, \pi_K)$, means $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K$
  - At time t we generate a new tree as follows:

$$c_t \sim \pi, \qquad x_t \sim N(\mu_{c_t}, \Sigma_{c_t})$$

Probabilistic Models

More generally:

- Θ consists of set of possible parameters
- We have a distribution P_θ over the data induced by each $\theta \in \Theta$
- Data is generated by one of the $\theta \in \Theta$
- Learning: Estimate value or distribution for $\theta^* \in \Theta$ given data

Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

$$\theta_{MLE} = \text{argmax}_{\theta \in \Theta} \underbrace{\log P_\theta(x_1, \ldots, x_n)}_{\text{Likelihood}}$$

Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

$$\begin{aligned}\theta_{MAP} &= \text{argmax}_{\theta \in \Theta} \log P(\theta | x_1, \ldots, x_n) \\ &= \text{argmax}_{\theta \in \Theta} \underbrace{P(x_1, \ldots, x_n | \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}\end{aligned}$$

Latent Variables

- We only observe $x_1, \ldots, x_n$, cluster assignments $c_1, \ldots, c_n$ are not observed
- Finding $\theta \in \Theta$ (even for 1-d GMM) that directly maximizes Likelihood or A Posteriori given $x_1, \ldots, x_n$ is hard!
- Given latent variables $c_1, \ldots, c_n$, the problem of maximizing likelihood (or a posteriori) became easy

Can we use latent variables to device an algorithm?

Expectation Maximization Algorithm

- For demonstration we shall consider the problem of finding MLE (MAP version is very similar)

Expectation Maximization Algorithm

- For demonstration we shall consider the problem of finding MLE (MAP version is very similar)
- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

(M step)

$$\theta^{(i)} = \text{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta)$$

Example: EM for GMM

- E step: For every $k \in [K]$,

$$Q_t^{(i)}(c_t = k) = P\left(c_t = k | x_t, \theta^{(i-1)}\right) = P\left(x_t | c_t = k, \theta^{(i-1)}\right) \times P\left(c_t = k | \theta^{(i-1)}\right)$$
$$\propto \underbrace{\phi\left(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}\right)}_{\text{gaussian p.d.f.}} \times \pi_k^{(i-1)}$$

Example: EM for GMM

- E step: For every $k \in [K]$,

$$Q_t^{(i)}(c_t = k) = P\left(c_t = k | x_t, \theta^{(i-1)}\right) = P\left(x_t | c_t = k, \theta^{(i-1)}\right) \times P\left(c_t = k | \theta^{(i-1)}\right)$$
$$\propto \underbrace{\phi\left(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}\right)}_{\text{gaussian p.d.f.}} \times \pi_k^{(i-1)}$$

- M step: Given $Q_1, \ldots, Q_n$, we need to find

$$\theta^{(i)} = \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log P(x_t, c_t = k | \theta)$$
$$= \operatorname*{argmax}_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(\log P(x_t | c_t = k, \theta) + \log P(c_t = k | \theta)\right)$$
$$= \operatorname*{argmax}_{\pi, \mu_{1,\ldots,K}, \Sigma_{1,\ldots,K}} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(k) \left(\log \phi\left(x_t; \mu_k, \Sigma_k\right) + \log \pi_k\right)$$

Example: EM for GMM

For every $k \in [K]$, the maximization step yields,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) \left(x_t - \mu_k^{(i)}\right)\left(x_t - \mu_k^{(i)}\right)^\top}{\sum_{t=1}^n Q_t(k)}$$

$$\pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

Example: EM for GMM

For every $k \in [K]$, the maximization step yields,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) \left(x_t - \mu_k^{(i)}\right)\left(x_t - \mu_k^{(i)}\right)^\top}{\sum_{t=1}^n Q_t(k)}$$

$$\pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

Hard and Soft Assignments

- Hard GMM: make E-step a hard assignments with $Q_t^{(i)} = e_{k_t^*}$ where

$$k_t^* = \text{argmax}_{k \in [K]} \phi\left(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}\right) \times \pi_k^{(i-1)}$$

- Soft k-means:

$$Q_t^{(i)}(k) \propto \exp(-\left\|x_t - \mu_k^{(i-1)}\right\|_2^2 / \sigma^2) \text{ and } \mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

k-means can be seen as hard GMM with spherical covariance and

Why should EM work?

A very high level view:

- Performing E-step will never decrease log-likelihood (or log a posteriori)

Why should EM work?

A very high level view:

- Performing E-step will never decrease log-likelihood (or log a posteriori)
- Performing M-step will never decrease log-likelihood (or log a posteriori)

WHY SHOULD EM WORK?

Steps to show that $\log \mathrm{Lik}(\theta^{(i+1)}) \geq \log \mathrm{Lik}(\theta^{(i)})$:

- Insert latent variables
- Use Jensen's inequality (log is a concave function)
- Massage the terms!

Why should EM work?

- Likelihood never decreases
- So whenever we converge we converge to a local optima
- However problem is non-convex and can have many local optimal
- In general no guarantee on rate of convergence
- In practice, do multiple random initializations and pick the best one!

EM IN GENERAL

- There was nothing special about GMM or clustering problems
- EM can be used as a general strategy for any problem with latent/missing/unobserved variables
- The MAP version only involves an extra prior term over θ multiplied to the likelihood
- In general probabilistic models with observed and latent variables can be represented succinctly as graphical models.
  Next time …