EM ALGORITHM

Today

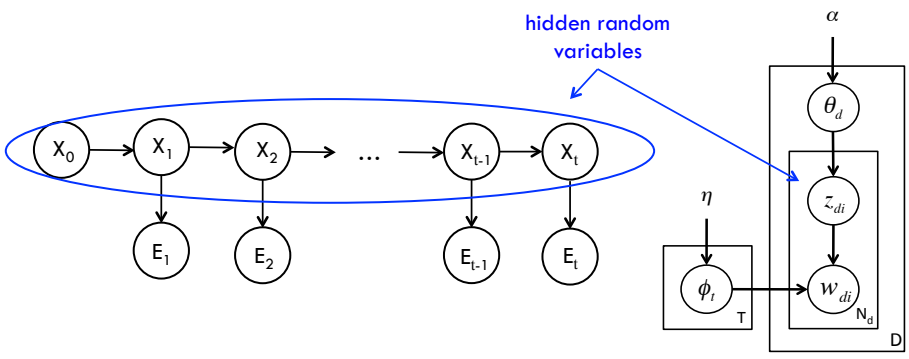
- Reading
  - IR Chapter 16 (AIMA Chapter 20)
- Goals
  - Maximum likelihood estimation
  - Expectation Maximization algorithm
- Announcements
  - Exam is on Friday. Still have TA hours this week
  - Transcendence Doodle – pick a time

Types of clustering algorithms

- Flat versus Hierarchical
  - Flat algorithms return an unstructured set of clusters
  - Hierarchical algorithms return a hierarchy of clusters
- Sequential (online) versus Batch
  - Sequential algorithms are typically fast
- Hard versus soft
  - Hard algorithms make a hard assignment of elements to clusters
  - Soft algorithms compute a distribution over clusters for each element

Expectation Maximization

- EM is an iterative algorithm for performing maximum likelihood estimation when there are hidden (i.e. unobserved or latent) random variables



Expectation Maximization

- EM is an iterative algorithm for performing maximum likelihood estimation when there are hidden (i.e. unobserved or latent) random variables

- Q1: What is maximum likelihood estimation?
- Q2: What does maximum likelihood estimation look like when there are hidden variables?
- Q3: How does EM help?
- Q4: What does EM have to do with clustering?

Maximum Likelihood Estimation

- Maximum likelihood estimation is a particular type of probabilistic inference

  X = observed data

  θ = model parameters

- The likelihood of X is p(X| θ) viewed as a function of θ
- We want to find the value of θ that maximizes the likelihood of X, i.e. the value of θ that makes the observed data the most likely

$$\theta^{\mathrm{MLE}} = \mathrm{argmax}_{\theta}\ p(X|\theta)$$

Called the maximum likelihood estimate (MLE)

Maximum Likelihood Estimation

- Example: Suppose we observe a set of test scores. We make the assumption that our data is Normally distributed.

$$X = \{63, 77, 85, 81, 92, 93, 86\}$$
$$x_i \sim \text{Normal}(\mu, \sigma^2)$$
$$\theta = \{\mu, \sigma^2\}$$

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$\theta^{\text{MLE}} = \text{argmax}_\theta \; p(X|\theta)$$

Maximum Likelihood Estimation

- We can find θ^{MLE} by taking the derivative of the log likelihood with respect to θ and setting equal to 0

$$\frac{\partial\, p(X|\theta)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \hat{\mu} = \frac{577}{7} = 82.43$$

$$\frac{\partial\, p(X|\theta)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2 \qquad \hat{\sigma}^2 = 90.24$$

variance

82.43 ± 9.5

standard deviation

ML Estimation with hidden variables

- Sometimes our probabilistic model includes hidden variables

  X = observed data

  Z = unobserved data

  θ = model parameters

- Computing the likelihood of X now requires summing over Z

$$p(X|\theta) = \sum_z p(X, Z = z|\theta)$$

Can be computationally intractable

- The complete-data likelihood is given by

$$p(X, Z|\theta)$$

If we knew Z, life would be simpler

ML Estimation with hidden variables

- Example: Clustering!

$$X = \{x_1, x_2, \ldots, x_N\} \quad \text{where} \quad x_i \in \mathbb{R}^M \qquad \text{// the data points}$$
$$Z = \{z_1, z_2, \ldots, z_N\} \quad \text{where} \quad z_i \in \{1, 2\} \qquad \text{// unknown cluster assignments}$$
$$\theta = \{\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2\} \qquad \text{// the model parameters}$$

- The probabilistic model:

$$z_i \sim \text{Discrete}(\alpha, 1-\alpha)$$
$$x_i|z_i \sim \text{Normal}(\mu_{z_i}, \sigma_{z_i})$$

$$p(X, Z|\theta) = \prod_{i=1}^{N} p(x_i|\mu_{z_i}, \sigma_{z_i}) \cdot p(z_i|\alpha)$$
$$p(X|\theta) = \prod_{i=1}^{N} \left(\sum_k p(x_i|\mu_k, \sigma_k) \cdot p(z_i = k|\alpha) \right)$$

Expectation Maximization

1. Initialize θ
2. Expectation (E-Step): Compute the Q function

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}\left[\log p(X, Z|\theta)\right]$$
$$= \sum_z p(Z = z|X, \theta^t) \cdot \log p(X, Z = z|\theta)$$

3. Maximization (M-Step): Maximize the Q function

$$\theta^{t+1} = \text{argmax}_\theta \; Q(\theta|\theta^t)$$

4. Repeat E-step and M-step till convergence

E-step

- Compute the expected value of the complete data log likelihood function w.r.t the conditional distribution of Z given X and the current guess of θ^t
  - We don't know Z. If we did know Z, the likelihood would be easy to compute!
  - Let's use the expected value of Z given X and θ^t. This is as good a guess as any (and better than most)!

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}\left[\log p(X, Z|\theta)\right]$$
$$= \sum_z p(Z = z|X, \theta^t) \cdot \log p(X, Z = z|\theta)$$

M-step

- Optimize this new Q function

$$\theta^{t+1} = \text{argmax}_\theta \; Q(\theta|\theta^t)$$

- We wanted to optimize p(X| θ) but in fact we're optimizing the Q function
- In fact, it can be shown that

$$\text{argmax}_\theta \; Q(\theta|\theta^t) \leq \text{argmax}_\theta \; \log p(X|\theta)$$

EM applied to clustering

- E-Step: Compute the Q function

$$\begin{aligned} Q(\theta|\theta^t) &= E[\log p(X, Z|\theta)] \\ &= E[\log \prod_i p(x_i|\mu_{z_i}, \sigma_{z_i}) \; p(z_i|\alpha)] \\ &= \sum_i E[\log p(x_i|\mu_{z_i}, \sigma_{z_i}) + \log p(z_i|\alpha)] \\ &= \sum_i \sum_k p(z_i = k|x_i, \theta^t) \cdot [\log p(x_i|\mu_k, \sigma_k) + \log p(z_i = k|\alpha)] \end{aligned}$$

Ultimately, it comes down to computing this quantity

$$\begin{aligned} \omega_k^i &= p(z_i = k|x_i, \theta^t) \\ &\propto p(x_i|\mu_k^t, \sigma_k^t) p(z_i = k|\alpha^t) \end{aligned}$$

EM applied to clustering

- M-Step: Maximize Q function w.r.t. θ

$$\alpha_k^{t+1} = \frac{1}{N}\sum_i \omega_k^i$$
$$\mu_k^{t+1} = \frac{\sum_i \omega_k^i x_i}{\sum_i \omega_k^i}$$
$$\sigma_k^{t+1} = \frac{\sum_i (x_i - \mu_k^{t+1})^2}{\sum_i \omega_k^i}$$

- Given updated values, recompute the Q function

EM applied to clustering

- Now let's change our probablistic model. Assume we want to cluster text data

$X = \{x_1, x_2, \ldots, x_N\}$ where $x_i \in [0,1]^M$ // the documents
$Z = \{z_1, z_2, \ldots, z_N\}$ where $z_i \in \{1,2\}$ // unknown cluster assignments
$\theta = \{\alpha, q_1, q_2\}$ // the model parameters

- The probabilistic model:

$$z_i \sim \text{Discrete}(\alpha, 1-\alpha)$$
$$x_i | z_i \sim \text{Multivariate Bern}(q_{z_i})$$
$$p(x_i | z_i = k) = \prod_{j=1}^{M} q_{km}^{x_i} \cdot (1 - q_{km})^{(1-x_i)}$$

EM applied to clustering

- E-step: Ultimately it comes down to computing

$$\omega_k^i = p(z_i = k|x_i, \theta^t)$$
$$\propto p(x_i|q_k^t)p(z_i = k|\alpha^t)$$

reassign

- M-step: Optimize the Q function

$$\alpha_k^{t+1} = \frac{1}{N}\sum_i \omega_k^i$$
$$q_{km}^{t+1} = \frac{\sum_i \omega_k^i \, \mathrm{I}(x_{im} = 1)}{\sum_i \omega_k^i}$$

recompute