Mixture Models and The EM Algorithm
Preview

- Mixture models
- The EM algorithm
- Why EM works
- EM variants
Motivation

- “Standard” distributions (e.g., multivariate Gaussian) are too limited
- How do we represent and learn more complex ones?
- One answer: Mixtures of “standard” distributions
- In the limit, can approximate any distribution this way
- Also good (and widely used) as a clustering method
Mixture Models

\[ P(x) = \sum_{i=1}^{n_c} P(c_i)P(x|c_i) \]

**Objective function:** Log likelihood of data

**Naive Bayes:** \( P(x|c_i) = \prod_{j=1}^{n_d} P(x_j|c_i) \)

**AutoClass:** Naive Bayes with various \( x_j \) models

**Mixture of Gaussians:** \( P(x|c_i) = \) Multivariate Gaussian

**In general:** \( P(x|c_i) \) can be any distribution
Mixtures of Gaussians

\[ P(x|\mu_i) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_i}{\sigma} \right)^2 \right] \]
The EM Algorithm

Initialize parameters ignoring missing information

Repeat until convergence:

E step: Compute expected values of unobserved variables, assuming current parameter values

M step: Compute new parameter values to maximize probability of data (observed & estimated)

(Also: Initialize expected values ignoring missing info)
EM for Mixtures of Gaussians

Initialization: Choose means at random, etc.

E step: For all examples $x_k$:

$$P(\mu_i | x_k) = \frac{P(\mu_i)P(x_k | \mu_i)}{P(x_k)} = \frac{P(\mu_i)P(x_k | \mu_i)}{\sum_{i'} P(\mu_{i'})P(x_k | \mu_{i'})}$$

M step: For all components $c_i$:

$$P(c_i) = \frac{1}{n_e} \sum_{k=1}^{n_e} P(\mu_i | x_k)$$

$$\mu_i = \frac{\sum_{k=1}^{n_e} x_k P(\mu_i | x_k)}{\sum_{k=1}^{n_e} P(\mu_i | x_k)}$$

$$\sigma_i^2 = \frac{\sum_{k=1}^{n_e} (x_k - \mu_i)^2 P(\mu_i | x_k)}{\sum_{k=1}^{n_e} P(\mu_i | x_k)}$$
Why EM Works

\[ \theta_{new} = \operatorname{argmax}_{\theta} E_{\theta_{old}}[\log P(X)] \]
Other Instances of EM

• Learning Hidden Markov models
• Learning graphical models with missing data
EM Variants

**MAP:** Compute MAP estimates instead of ML in M step

**GEM:** Just increase likelihood in M step

**SEM/MCEM:** Approximate E step

**Simulated annealing:** Avoid local maxima

**Early stopping:** Faster, may reduce overfitting

**Structural EM:** EM with structure search
Summary

- The EM algorithm
- Mixture models
- Why EM works
- EM variants