The EM Algorithm

Preview

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

Learning with Missing Data

- **Goal:** Learn parameters of Bayes net with known structure
- For now: Maximum likelihood
- Suppose the values of some variables in some samples are missing
- If we knew all values, computing parameters would be easy
- If we knew the parameters, we could infer the missing values
- "Chicken and egg" problem

The EM Algorithm

Initialize parameters ignoring missing information

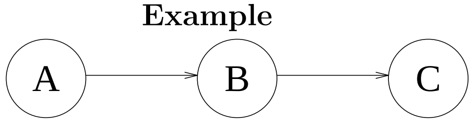
Repeat until convergence:

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)

Example



	A	B	C
Examples:	0	1	1
	1	0	0
	1	1	1
	1	?	0

Initialization: $P(B|A) =$ $P(C|B) =$
$P(A) =$ $P(B|\neg A) =$ $P(C|\neg B) =$

E-step: $P(? = 1) = P(B|A, \neg C) = \frac{P(A,B,\neg C)}{P(A,\neg C)} = \ldots = 0$

M-step: $P(B|A) =$ $P(C|B) =$
$P(A) =$ $P(B|\neg A) =$ $P(C|\neg B) =$

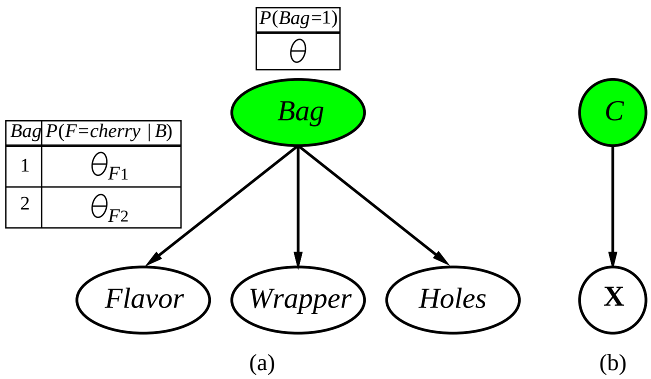
E-step: $P(? = 1) = 0$ **(converged)**

Hidden Variables

- What if some variables were always missing?
- In general, difficult problem
- Consider Naive Bayes structure, with class missing:

$$P(x) = \sum_{i=1}^{n_c} P(c_i) \prod_{j=1}^{d} P(x_j|c_i)$$

Naive Bayes Model



Clustering

- **Goal:** Group similar objects
- Example: Group Web pages with similar topics
- Clustering can be hard or soft
- What's the objective function?

Mixture Models

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

Objective function: Log likelihood of data

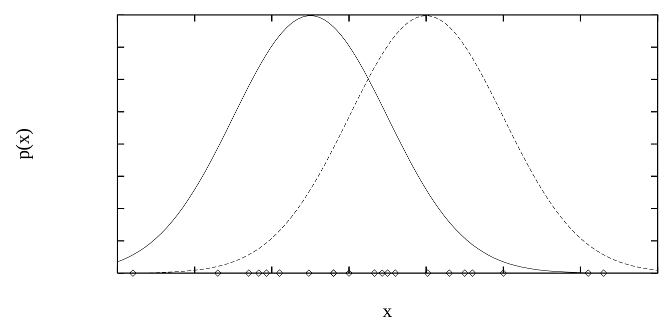
Naive Bayes: $P(x|c_i) = \prod_{j=1}^{n_d} P(x_j|c_i)$

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



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

EM for Mixtures of Gaussians

Simplest case: Assume known priors and covariances

Initialization: Choose means at random

E step: For all samples x_k:

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

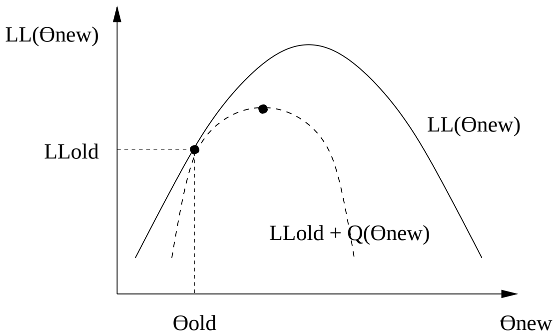
M step: For all means μ_i:

$$\mu_i = \frac{\sum_{x_k} x \, P(\mu_i|x_k)}{\sum_{x_k} P(\mu_i|x_k)}$$

Mixtures of Gaussians (cont.)

- K-means clustering $\prec$ EM for mixtures of Gaussians
- Mixtures of Gaussians $\prec$ Bayes nets
- Also good for estimating joint distribution of continuous variables

Why EM Works



$$\theta_{new} = \underset{\theta}{\operatorname{argmax}} E_{\theta_{old}}[\log P(X)]$$

EM Variants

MAP: Compute MAP estimates instead of ML in M step

GEM: Just increase likelihood in M step

MCMC: Approximate E step

Simulated annealing: Avoid local maxima

Early stopping: Faster, may reduce overfitting

Structural EM: Missing data and unknown structure

Summary

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