Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 22, 2011

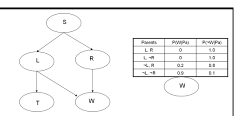
Today:

- Clustering
- Mixture model clustering
- Learning Bayes Net structure
  - Chow-Liu for trees

Readings:

Recommended:
- Jordan "Graphical Models"
- Muphy "Intro to Graphical Models"

Bayes Network Definition



A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of CPD's
- Each node denotes a random variable
- Edges denote dependencies
- CPD for each node X_i defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined as

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Pa(X) = immediate parents of X in the graph

Usupervised clustering

Just extreme case for EM with zero labeled examples…

Clustering

- Given set of data points, group them
- Unsupervised learning
- Which patients are similar? (or which earthquakes, customers, faces, web pages, …)

Mixture Distributions

Model joint $P(X_1 \ldots X_n)$ as mixture of multiple distributions.

Use discrete-valued random variable Z to indicate which distribution is being use for each random draw

So $P(X_1 \ldots X_n) = \sum_i P(Z=i) \, \boxed{P(X_1 \ldots X_n | Z)}$

Name of the cluster

Mixture of *Gaussians*:

- Assume each data point X=<X1, … Xn> is generated by one of several Gaussians, as follows:

1. randomly choose Gaussian i, according to P(Z=i)
2. randomly generate a data point <x1,x2 .. xn> according to $N(\mu_i, \Sigma_i)$

EM for Mixture of Gaussian Clustering

Let's simplify to make this easier:

1. assume X=<X_1 ... X_n>, and the X_i are conditionally independent given Z.

$$P(X|Z=j) = \prod_i N(X_i | \mu_{ji}, \sigma_{ji})$$

2. assume only 2 clusters (values of Z), and $\forall i, j, \sigma_{ji} = \sigma$

$$P(\mathbf{X}) = \sum_{j=1}^{2} P(Z=j|\pi) \prod_i N(x_i | \mu_{ji}, \sigma)$$

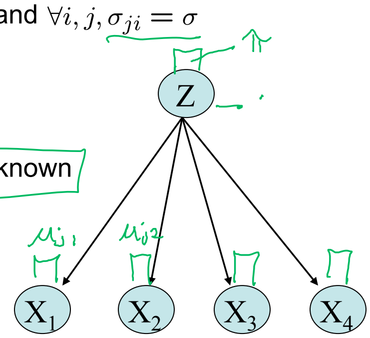
X_i

values of Z

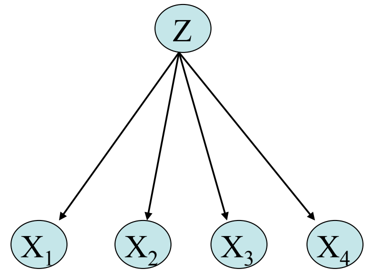
3. Assume σ known, $\boxed{\pi_1 \ldots \pi_K, \mu_{1i} \ldots \mu_{Ki} \text{ unknown}}$

Observed: X=<X_1 ... X_n>

Unobserved: Z



EM



Given observed variables X, unobserved Z

Define $Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')]$

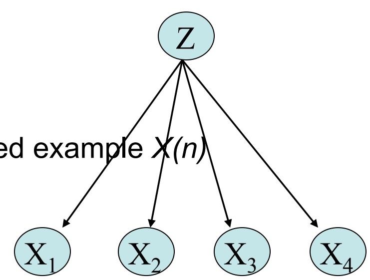
where $\theta = \langle \pi, \mu_{ji} \rangle$

Iterate until convergence:

- E Step: Calculate $P(Z(n)|X(n),\theta)$ for each example $X(n)$. Use this to construct $Q(\theta'|\theta)$
- M Step: Replace current θ by

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$$

EM – E Step



Calculate $P(Z(n)|X(n),\theta)$ for each observed example $X(n)$

$X(n)=<x_1(n), x_2(n), \ldots x_T(n)>$.

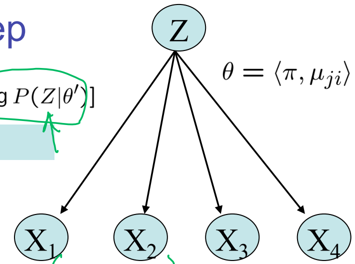
$$P(z(n)=k|x(n),\theta) = \frac{P(x(n)|z(n)=k,\theta)\ \ P(z(n)=k|\theta)}{\sum_{j=0}^{1} p(x(n)|z(n)=j,\theta)\ \ P(z(n)=j|\theta)}$$

$$P(z(n)=k|x(n),\theta) = \frac{[\prod_i P(x_i(n)|z(n)=k,\theta)]\ \ P(z(n)=k|\theta)}{\sum_{j=0}^{1} \prod_i P(x_i(n)|z(n)=j,\theta)\ \ P(z(n)=j|\theta)}$$

$$P(z(n)=k|x(n),\theta) = \frac{[\prod_i N(x_i(n)|\mu_{k,i},\sigma)]\ \ (\pi^k(1-\pi)^{(1-k)})}{\sum_{j=0}^{1}[\prod_i N(x_i(n)|\mu_{j,i},\sigma)]\ \ (\pi^j(1-\pi)^{(1-j)}))}$$

EM – M Step

First consider update for $\pi = P(Z=1)$



$\theta = \langle \pi, \mu_{ji} \rangle$

$$Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')] = E[\log P(X|Z,\theta') + \log P(Z|\theta')]$$

obs unobs

π' has no influence

$$\pi \leftarrow \arg\max_{\pi'} E_{Z|X,\theta}[\log P(Z|\pi')]$$

$$E_{Z|X,\theta}\left[\log P(Z|\pi')\right] = E_{Z|X,\theta}\left[\log\left(\pi'^{\sum_n z(n)}(1-\pi')^{\sum_n (1-z(n))}\right)\right]$$

$$= E_{Z|X,\theta}\left[\left(\sum_n z(n)\right)\log \pi' + \left(\sum_n (1-z(n))\right)\log(1-\pi')\right]$$

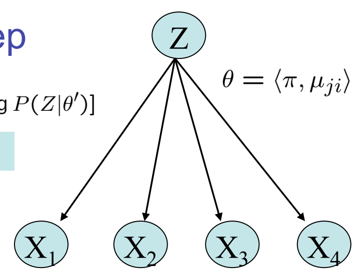
$$= \left(\sum_n E_{Z|X,\theta}[z(n)]\right)\log \pi' + \left(\sum_n E_{Z|X,\theta}[(1-z(n))]\right)\log(1-\pi')$$

$$\frac{\partial E_{Z|X,\theta}[\log P(Z|\pi')]}{\partial \pi'} = \left(\sum_n E_{Z|X,\theta}[z(n)]\right)\frac{1}{\pi'} + \left(\sum_n E_{Z|X,\theta}[(1-z(n))]\right)\frac{(-1)}{1-\pi'}$$

$$\pi \leftarrow \frac{\sum_{n=1}^N E[z(n)]}{\left(\sum_{n=1}^N E[z(n)]\right) + \left(\sum_{n=1}^N (1-E[z(n)])\right)} = \frac{1}{N}\sum_{n=1}^N E[z(n)]$$

EM – M Step

Now consider update for μ_{ji}



$\theta = \langle \pi, \mu_{ji} \rangle$

$$Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')] = E[\log P(X|Z,\theta') + \log P(Z|\theta')]$$

μ_{ji}' has no influence

$$\mu_{ji} \leftarrow \arg\max_{\mu'_{ji}} E_{Z|X,\theta}[\log P(X|Z,\theta')]$$

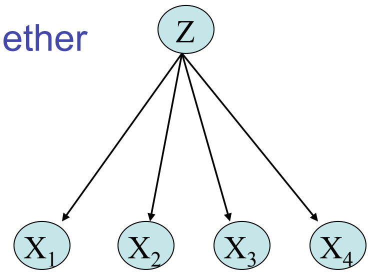
…
…
…

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^N P(z(n)=j|x(n),\theta) \quad x_i(n)}{\sum_{n=1}^N P(z(n)=j|x(n),\theta)}$$

Compare above to MLE if Z were observable:

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^N \delta(z(n)=j) \quad x_i(n)}{\sum_{n=1}^N \delta(z(n)=j)}$$

EM – putting it together



Given observed variables X, unobserved Z

Define $Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')]$

where $\theta = \langle \pi, \mu_{ji} \rangle$

Iterate until convergence:

- E Step: For each observed example X(n), calculate $P(Z(n)|X(n), \theta)$

$$P(z(n) = k \mid x(n), \theta) = \frac{[\prod_i N(x_i(n)|\mu_{k,i}, \sigma)] \quad (\pi^k (1-\pi)^{(1-k)})}{\sum_{j=0}^{1} [\prod_i N(x_i(n)|\mu_{j,i}, \sigma)] \quad (\pi^j (1-\pi)^{(1-j)}))}$$

- M Step: Update $\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$

$$\pi \leftarrow \frac{1}{N} \sum_{n=1}^{N} E[z(n)] \qquad \mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta) \quad x_i(n)}{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta)}$$

Mixture of Gaussians applet

Go to: http://www.socr.ucla.edu/htmls/SOCR_Charts.html
then go to Go to "Line Charts" → SOCR EM Mixture Chart

- try it with 2 Gaussian mixture components ("kernels")
- try it with 4

What you should know about EM

- For learning from partly unobserved data
- MLEst of θ = $\arg\max_\theta \log P(data|\theta)$
- EM estimate: θ = $\arg\max_\theta E_{Z|X,\theta}[\log P(X,Z|\theta)]$

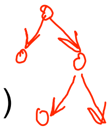
  Where X is observed part of data, Z is unobserved

  $E_{Z|X\theta} \log P(\theta|XZ)$ α

  α $P(XZ|\theta)P(\theta)$
- EM for training Bayes networks
- Can also develop MAP version of EM
- Can also derive your own EM algorithm for your own problem
  - write out expression for $E_{Z|X,\theta}[\log P(X,Z|\theta)]$
  - E step: for each training example X^k, calculate $P(Z^k \mid X^k, \theta)$
  - M step: chose new θ to maximize $E_{Z|X,\theta}[\log P(X,Z|\theta)]$

Learning Bayes Net Structure

How can we learn Bayes Net graph structure?



In general case, open problem

- can require lots of data (else high risk of overfitting)
- can use Bayesian methods to constrain search

One key result:

- Chow-Liu algorithm: finds "best" tree-structured network
- What's best?
  - suppose $P(\mathbf{X})$ is true distribution, $T(\mathbf{X})$ is our tree-structured network, where $\mathbf{X} = \langle X_1, \ldots X_n \rangle$
  - Chou-Liu minimizes Kullback-Leibler divergence:

$$KL(P(\mathbf{X}) \mid\mid T(\mathbf{X})) \equiv \sum_k P(\mathbf{X}=k) \log \frac{P(\mathbf{X}=k)}{T(\mathbf{X}=k)}$$

Chow-Liu Algorithm

<u>Key result</u>: To minimize KL(P || T), it suffices to find the tree network T that maximizes the sum of mutual informations over its edges

Mutual information for an edge between variable A and B:

$$I(A,B) = \sum_a \sum_b P(a,b) \log \frac{P(a,b)}{P(a)P(b)}$$

This works because for tree networks with nodes $\mathbf{X} \equiv \langle X_1 \dots X_n \rangle$

$$\begin{aligned} KL(P(\mathbf{X}) \mid\mid T(\mathbf{X})) &\equiv \sum_k P(\mathbf{X}=k) \log \frac{P(\mathbf{X}=k)}{T(\mathbf{X}=k)} \\ &= -\sum_i I(X_i, Pa(X_i)) + \sum_i H(X_i) - H(X_1 \dots X_n) \end{aligned}$$

Chow-Liu Algorithm

1. for each pair of vars A,B, use data to estimate P(A,B), P(A), P(B)

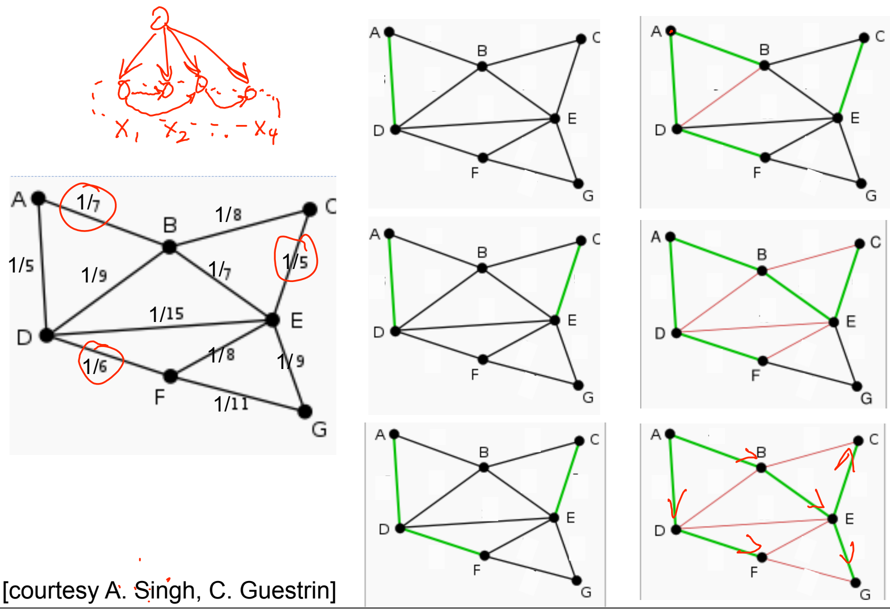
2. for each pair of vars A,B calculate mutual information
$$I(A,B) = \sum_a \sum_b P(a,b) \log \frac{P(a,b)}{P(a)P(b)}$$

3. calculate the maximum spanning tree over the set of variables, using edge weights *I(A,B)*
(given N vars, this costs only $O(N^2)$ time)

4. add arrows to edges to form a directed-acyclic graph

5. learn the CPD's for this graph

Chow-Liu algorithm example
Greedy Algorithm to find Max-Spanning Tree



[courtesy A. Singh, C. Guestrin]

Bayes Nets – What You Should Know

- Representation
  - Bayes nets represent joint distribution as a DAG + Conditional Distributions
  - D-separation lets us decode conditional independence assumptions
- Inference
  - NP-hard in general
  - For some graphs, closed form inference is feasible
  - Approximate methods too, e.g., Monte Carlo methods, …
- Learning
  - Easy for known graph, fully observed data (MLE's, MAP est.)
  - EM for partly observed data
  - Learning graph structure: Chow-Liu for tree-structured networks