### Clustering Part 2

EECS 349 Spring 2015

### **Expectation Maximization**

- Learning parameters in Bayes Nets is easy if data is complete
  - Just counting

#### But what about missing data?

- We could use our standard "missing data" techniques (use mean, median, etc.)
- But when lots of data is missing, we want to infer missing data and parameters simultaneously
  - We can use **Expectation Maximization**

- K classes, each class  $\omega_i$  produces Gaussian observations with mean  $\mu_i$  with variance  $\sigma^2 I$
- Assume σ<sup>2</sup>I given (for now), and we have lots of observations
- Task: estimate  $\mu_i$
- But, none of the data points are labeled...

### Gaussian Mixtures



- There are k components. The i'th component is called ω<sub>i</sub>
- Component ω<sub>i</sub> has an associated mean vector μ<sub>i</sub>



- There are k components. The i'th component is called ω<sub>i</sub>
- Component ω<sub>i</sub> has an associated mean vector μ<sub>i</sub>
- Each component generates data from a Gaussian with mean μ<sub>i</sub> and covariance matrix σ<sup>2</sup>I

Assume that each datapoint is generated according to the following recipe:



- There are k components. The i'th component is called  $\omega_i$
- Component ω<sub>i</sub> has an associated mean vector μ<sub>i</sub>
- Each component generates data from a Gaussian with mean μ<sub>i</sub> and covariance matrix σ<sup>2</sup>I
- Assume that each datapoint is generated according to the following recipe:
- 1. Pick a component at random. Choose component i with probability  $P(\omega_i)$ .



- There are k components. The i'th component is called ω<sub>i</sub>
- Component ω<sub>i</sub> has an associated mean vector μ<sub>i</sub>
- Each component generates data from a Gaussian with mean μ<sub>i</sub> and covariance matrix σ<sup>2</sup>I
- Assume that each datapoint is generated according to the following recipe:
- 1. Pick a component at random. Choose component i with probability  $P(\omega_i)$ .
- 2. Datapoint ~ N( $\mu_{\mu} \sigma^2 \mathbf{I}$ )



#### The data generated



### Computing the likelihood

Remember:

D

We have unlabeled data  $x_1 x_2 \dots x_R$ We know there are k classes We know P(w<sub>1</sub>) P(w<sub>2</sub>) P(w<sub>3</sub>) ... P(w<sub>k</sub>) We <u>don't</u> know  $\mu_1 \mu_2 \dots \mu_k$ 

We can write P( data | 
$$\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_k$$
)  
=  $p(x_1 \dots x_R | \boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_k)$   
=  $\prod_{i=1}^R p(x_i | \boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_k)$   
=  $\prod_{i=1}^R \sum_{j=1}^k p(x_i | w_j, \boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_k) P(w_j)$   
=  $\prod_{i=1}^R \sum_{j=1}^k K \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right) P(w_j)$ 

Most slides from http://www.autonlab.org/tutorials/

#### EM for GMMs

For Max likelihood we know  $\frac{\partial}{\partial \mu_i} \log \Pr \operatorname{ob}(\operatorname{data}|\mu_1...\mu_k) = 0$ 

Some wild'n' crazy algebra turns this into : "For Max likelihood, for each j,

$$\mu_j = \frac{\sum_{i=1}^R P(w_j | x_i, \mu_1 \dots \mu_k) x_i}{\sum_{i=1}^R P(w_j | x_i, \mu_1 \dots \mu_k)}$$

This is n nonlinear equations in  $\mu_i$ 's."

#### EM for GMMs

For Max likelihood we know  $\frac{\partial}{\partial \mu_i} \log \Pr \operatorname{ob}(\operatorname{data}|\mu_1 \dots \mu_k) = 0$ 

Some wild'n' crazy algebra turns this into : "For Max likelihood, for each j,

$$\mu_{j} = \frac{\sum_{i=1}^{R} P(w_{j} | x_{i}, \mu_{1}...\mu_{k}) x_{i}}{\sum_{i=1}^{R} P(w_{j} | x_{i}, \mu_{1}...\mu_{k})}$$

This is n nonlinear equations in  $\mu_i$ 's."

If, for each  $\mathbf{x}_i$  we knew that for each  $w_j$  the prob that  $\mathbf{\mu}_j$  was in class  $w_j$  is  $P(w_j|x_i,\mu_1...\mu_k)$  Then... we would easily compute  $\mu_j$ .

If we knew each  $\mu_j$  then we could easily compute  $P(w_j|x_i, \mu_1...\mu_j)$  for each  $w_j$  and  $x_i$ .



Compute Max. like **µ** given our data's class membership distributions

$$\mu_i(t+1) = \frac{\sum_k P(w_i | x_k, \lambda_t) x_k}{\sum_k P(w_i | x_k, \lambda_t)}$$

Gaussian Mixture Example: Start



Advance apologies: in Black and White this example will be incomprehensible

# After first iteration



# After 2nd iteration

Þ



# After 3rd iteration

Þ



# After 4th iteration



# After 5th iteration



# After 6th iteration



# After 20th iteration



### EM at the 10,000 foot level

#### • Guess some parameters, then

- Use your parameters to get a distribution over hidden variables
- Re-estimate the parameters as if your distribution over hidden variables is correct
- Seems magical. When/why does this work?

## Underlying EM: The basic idea

- $\blacktriangleright$  EM: Given a guess  $\theta_{\rm old}$  for  $\theta$  , improve it
- Idea: construct lower bound that equals the true log likelihood at  $\theta_{old}$ :



## For exponential family

#### • E step:

• Use  $\theta_n$  to estimate **expected** sufficient statistics over **complete** data

#### M step

- Set  $\theta_{n+1} = ML$  parameters given sufficient statistics
  - (Or MAP parameters)

## EM in practice

#### Local maxima

- Random re-starts, simulated annealing...
- Variants
  - Hard EM: set Z to most likely value (e.g. k-means)
  - Generalized EM: increase (not nec. maximize) lower bound in each step
  - Approximate E-step (e.g. sampling)

#### **Seeking Life's Bare (Genetic) Necessities**

Haemophilus

genome 1703 genes

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms

required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a <u>computational</u> molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an





Simple intuition: Documents exhibit multiple topics.

From David Blei's 2012 ICML tutorial



#### Documents

#### Topic proportions and assignments



- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

#### Topics

#### Documents

#### Topic proportions and assignments



- In reality, we only observe the documents
- The other structure are hidden variables

LDA: Math Version

For each topic t
 Choose distribution φ<sub>t</sub> ~ Dirichlet(β)

For each doc

Choose  $\theta \sim \text{Dirichlet}(\alpha)$ 

For each token *i* 

choose topic  $z_i \sim Mult(\theta)$ choose word  $w_i \sim Mult(\phi_{z_i})$ 

Exact inference is intractable

> We will use a collapsed sampler that integrates out  $\phi$  and  $\theta$ [Griffiths and Steyvers, 2007]

#### Inference

Variational and sampling-based methods exist

- Simple collapsed Gibbs sampling approach:
  - Initialize all topic variables  $z_i$  randomly to one of K topics
  - For each sampling pass
    - For each token i

 $\Box$  Sample a new value for  $z_i$  given all other topic variable assignments

### Sampling Distribution

• P(topic z | word w, doc d) 
$$\propto \frac{n_z^d + \alpha}{n_z^d + \alpha K} \frac{n_z^w + \beta}{n_z^w + \beta V}$$

 $n_z^d = number of times topic t assigned in doc d$ 

- $h_{z}^{w}$  = number of times topic *t* assigned for word *w*
- K = number of topics
- V = number of unique words
- $\alpha, \beta$  : Dirichlet prior hyperparameters

#### Example Inference

human genome dna genetic genes sequence gene molecular sequencing map information genetics mapping project sequences

evolution evolutionary species organisms life origin biology groups phylogenetic living diversity group new two common

disease host bacteria diseases resistance bacterial new strains control infectious malaria parasite parasites united tuberculosis

computer models information data computers system network systems model parallel methods networks software new simulations

From David Blei's 2012 ICML tutorial

1	2	3	4	5
dna	protein	water	savs	mantle
dene	cell	climate	researchers	high
sequence	cells	atmospheric	new	earth
aenes	proteins	temperature	university	pressure
sequences	receptor	global	just	seismic
human	fig	surface	science	crust
genome	binding	ocean	like	temperature
genetic	activity	carbon	work	earths
analysis	activation	atmosphere	first	lower
two	kinase	changes	years	earthquakes
6	7	8	9	10
end	time	materials	dna	disease
article	data	surface	i rha i	cancer
start	two	high	transcription	patients
science	model	structure	protein	human
readers	fig	temperature	site	gene
service	system	molecules	binding	medical
news	number	chemical	sequence	studies
card	different	molecular	proteins	drug
circle	results	fig	specific	normal
letters		university	sequences	drugs
11	12	13	14	15
vears	Species	protein	Cells	space
million	evolution	structure	cell	solar
ago	population	proteins	virus	observations
age	evolutionary	two	hiv	earth
university	university	amino	infection	stars
north	populations	binding	immune	university
early	natural	acid	human	mass
fig	studies	residues	antigen	sun
evidence	genetic	molecular	infected	astronomers
record	blology	structural	viral	telescope
16				
	17	18	19	20
a fax	17 Cells	18 energy	19 research	20 neurons
tax manager	17 Cells cell	18 energy electron	19 research science	20 neurons brain
fax manager science	17 Cells cell gene	18 energy electron state	19 research science national	20 neurons brain cells
fax manager science aaas	17 Cells cell gene genes	18 energy electron state light	19 research science national scientific	20 Neurons brain cells activity
fax manager science aaas advertising	17 Cells cell gene genes expression	18 energy electron state light quantum	19 research science national scientific scientists	20 Neurons brain cells activity fig
fax manager science aaas advertising sales	17 Cells cell gene genes expression development	18 energy electron state light quantum physics	19 research science national scientific scientists new	20 Neurons brain cells activity fig channels
fax manager science aaas advertising sales member	17 Cells cell gene genes expression development mutant	18 energy electron state light quantum physics electrons	19 research science national scientific scientists new states	20 Neurons brain cells activity fig channels university
fax manager science aaas advertising sales member recruitment	17 Cells Cell gene genes expression development mutant mice	18 energy electron state light quantum physics electrons high	19 research science national scientific scientists new states university	20 Neurons brain cells activity fig channels university cortex
fax manager science aaas advertising sales member recruitment associate	17 Cells Cell gene genes expression development mutant mice fig	18 energy electron state light quantum physics electrons high laser	19 research science national scientists new states university unted	20 Neurons brain cells activity fig channels university cortex neuronal

From David Blei's 2012 ICML tutorial