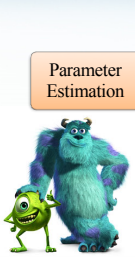
Lecture Notes on

Pattern Recognition

Session 03(B): Parameter Estimation (II)

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Road map



- Parameter Estimation
  - Expectation Maximization
    - Motivating Example
    - Intuition of EM
  - EM Algorithm
    - Expectation Step
    - Maximization Step
    - Illustration

Expectation Maximization

- Motivation Example
- Intuition of EM

Expectation Maximization

- The Expectation Maximization Algorithm is an approach to handle the problem of learning in the presence of unobserved variables.
  - When data is only partially observable
    - Unsupervised Clustering
      - Target Value is unobservable
    - Supervised Learning
      - Some instance attributes unobservable
  - Old idea (late 50's) but formalized by Dempster, Laird and Rubin in 1977

A Silly Example

- Let events be "grades in a class"
  - w_1 = Gets an A
  - w_2 = Gets a B
  - w_3 = Gets a C
  - w_4 = Gets a D

Assume
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

(Note $0 \leq \mu \leq 1/6$)

- Assume we want to estimate μ from data.
- In a given class there were

A	B	C	D
a	b	c	d

- What's the maximum likelihood estimate of μ given a,b,c,d ?

A Silly Example

- Let events be "grades in a class"
  - w_1 = Gets an A
  - w_2 = Gets a B
  - w_3 = Gets a C
  - w_4 = Gets a D

Assume
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

(Note $0 \leq \mu \leq 1/6$)

- Assume we want to estimate μ from data.
- In a given class there were

A	B	C	D
14	6	9	10

- What's the maximum likelihood estimate of μ given a,b,c,d ?

Trivial Statistics

P(A) = ½ P(B) = μ P(C) = 2μ P(D) = ½-3μ

P(*a,b,c,d* | μ) = K(½)^a(μ)^b(2μ)^c(½-3μ)^d

log P(*a,b,c,d* | μ) = log K + *a*log ½ + *b*log μ + *d*log 2μ + *d*log (½-3μ)

FOR MAX LIKE μ, SET $\frac{\partial \text{LogP}}{\partial \mu} = 0$

$\frac{\partial \text{LogP}}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{1/2 - 3\mu} = 0$

Gives max like $\mu = \frac{b+c}{6(b+c+d)}$

So if class got

A	B		D
14			10

Boring, but true!

Max like $\mu = \frac{1}{10}$

A Not-So-Silly Example

- Someone tells us that
- Number of High grades (A's + B's) = *h*
- Number of C's = *c*
- Number of D's = *d*
- What is the max. like estimate of μ now?

Assume
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

EXPECTATION

If we know the value of μ we could compute the expected value of *a* and *b*

Since the ratio a:b should be the same as the ratio ½ : μ

$a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu} h \qquad b = \frac{\mu}{\frac{1}{2}+\mu} h$

MAXIMIZATION

If we know the expected values of *a* and *b* we could compute the maximum likelihood value of μ

$\mu = \frac{b+c}{6(b+c+d)}$

A Taste of EM Algorithm

- We begin with a guess for μ
- We iterate between E Step and M Step to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t-th iteration
b(t) the estimate of *b* on t-th iteration

Assume
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

$\mu(0) = \text{initial guess}$

$b(t) = \frac{\mu(t)h}{\frac{1}{2}+\mu(t)} = \text{E}[b \mid \mu(t)]$ — E-step

$\mu(t+1) = \frac{b(t)+c}{6(b(t)+c+d)}$ = max like est of μ given $b(t)$ — M-step

A Taste of EM Algorithm

- We begin with a guess for μ
- We iterate between E Step and M Step to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t-th iteration
b(t) the estimate of *b* on t-th iteration

Assume
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

suppose we had
h = 20
c = 10
d = 10
μ(0) = 0

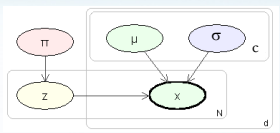
t	μ(t)	b(t)
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

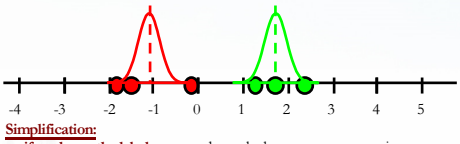
Convergence proof based on fact that Prob(data | μ) must increase or remain same between each iteration, and <=1

Probabilistic model

Imagine model generating data

- Need to introduce label, z, for each data point
- Label is called a *latent* variable, also called *hidden, unobserved, missing*



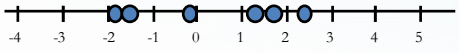


Simplification:
if we knew the labels, we can decouple the components as estimate parameters separately for each one

Motivating example

- Task: Fit mixture of Gaussian model with C=2 components $\theta^* = argmax_\theta \prod_{n=1}^{N} p(x_i|\theta)$

Data: $x = (x_1, x_2, \ldots, x_N)$



$p(x_i|\theta) = \sum_c p(c|\theta)\, p(x_i|c,\theta) = \sum_c w(c)\, N(x_i|\mu_c, \sigma_c)$

where $\sum_{c=1}^{C} w(c) = 1$

Motivating example

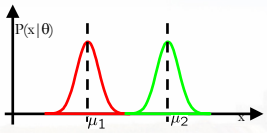
- Task: Fit mixture of Gaussian model with C=2 components $\theta^* = argmax_\theta \prod_{n=1}^{N} p(x_i|\theta)$

Data: $x = (x_1, x_2, \ldots, x_N)$

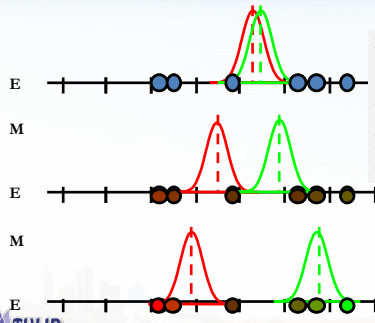
Parameters: $\theta = \{w, \mu, \sigma\}$

keep w, σ fixed

i.e. only estimate μ



Intuition of EM



E-step: Compute a *distribution* on the labels of the points, using current parameters

M-step: Update parameters using current guess of label distribution.

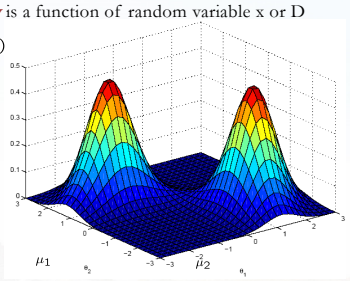
Expectation Maximization

- Consider a model with parameters **θ** and a number of observations (and hidden variables) D from which we want to estimate **θ**.
- MLE uses the probability distributions $P(D|\boldsymbol{\theta})$ in an unusual way.
  - D is treated as the constant and **θ** as the variable.
  - Using this perspective, the function value of p is viewed as a likelihood rather than a probability distribution value.

Likelihood function

- ***Likelihood*** is a function of parameters **θ**
  Probability is a function of random variable x or D

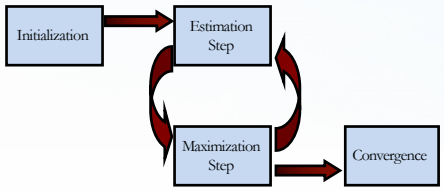
$l(\mu_1, \mu_2|x)$



Expectation Maximization

- Task: find **θ** that maximizes the likelihood of D.
  $\boldsymbol{\theta}^*$ = argmax $P(D|\boldsymbol{\theta})$ = argmax $ln\ P(D|\boldsymbol{\theta})$
  - $ln\ P(D|\boldsymbol{\theta})$ is used for simplicities sake.
  - Depending on the form of P(D|**θ**) this problem can be easy or hard.
    - The more complex the distribution, the more complex the technique of using the MLE.
  - While MLE is popular, other estimators may be used

Maximum Likelihood Estimation



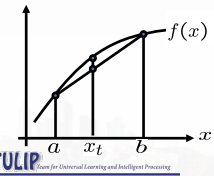
EM Algorithm

- Expectation Step
- Maximization Step
- Illustration

Jensen's Inequality

- **Jensen's Inequality**:
  - For a real continuous concave function $f()$and $\lambda_j \geq 0,\ \sum_j \lambda_j = 1,$ we have

$$f(\sum_j \lambda_j x_j) \geq \sum_j \lambda_j f(x_j)$$



Equality holds when all x are the same

Terminology

- **Observed data**:

$$x = (x_1, x_2, \ldots, x_N)$$

Continuous I.I.D

- **Latent variables**:

$$z = (z_1, z_2, \ldots, z_N)$$

Discrete 1 ... C

Terminology

- **Incomplete Log-Likelihood** (ILL):

$$l(\theta; x) = log\ p(x|\theta) = log\ \prod_x p(x|\theta) = \sum_x log\ \sum_z p(x, z|\theta)$$

- **Complete Log-Likelihood** (CLL):

$$l_c(\theta; x, z) \triangleq \sum_x \log p(x, z \mid \theta)$$

- **Expected Complete Log-Likelihood** (ECLL):

$$\langle l_c(\theta; x, z) \rangle_q \triangleq \sum_x \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta)$$

Lower bound on log-likelihood

$$\begin{aligned} l(\theta; x) &= \log p(x \mid \theta) \\ &= \log \sum_z p(x, z \mid \theta) \\ &= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\ &\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\ &\triangleq \mathcal{L}(q, \theta), \end{aligned}$$

Use Jensen's inequality

AUXILIARY FUNCTION

EM is alternating ascent

- Key: ILL is always larger than the Auxiliary function

$$\mathcal{L}(q, \theta) \leq l(\theta; x)$$

- EM Algorithm:
  - Alternately improving q then **θ**, is guaranteed to improve likelihood itself….

$$\begin{aligned} &(\textbf{E step}) && q^{(t+1)} &&= \arg\max_q \mathcal{L}(q, \theta^{(t)}) \\ &(\textbf{M step}) && \theta^{(t+1)} &&= \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta). \end{aligned}$$

Expectation-step: Choosing the optimal q(z | x, θ)

- Turns out that q(z|x,θ) = p(z|x,θ^t) is the best

$$\begin{aligned}\mathcal{L}(p(z \mid x,\theta^{(t)}),\theta^{(t)}) &= \sum_z p(z \mid x,\theta^{(t)}) \log \frac{p(x,z \mid \theta)}{p(z \mid x,\theta^{(t)})} \\ &= \sum_z p(z \mid x,\theta^{(t)}) \log p(x \mid \theta^{(t)}) \\ &= \log p(x \mid \theta^{(t)}) \\ &= l(\theta^{(t)};x).\end{aligned}$$

Expectation-step: Choosing the optimal q(z | x, θ)

- Turns out that q(z|x,θ) = p(z|x,θ^t) is the best

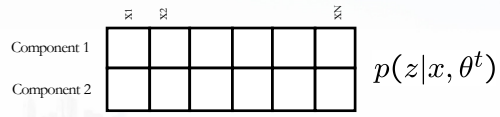
$$\begin{aligned}l(\theta;x) - \mathcal{L}(q,\theta) &= l(\theta;x) - \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)} \\ &= \sum_z q(z \mid x) \log p(x \mid \theta) - \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)} \\ &= \sum_z q(z \mid x) \log p(x \mid \theta) - \log \frac{q(z \mid x)}{p(z \mid x,\theta)} \\ &= D(q(z \mid x) \parallel p(z \mid x,\theta)).\end{aligned}$$

ALTERNATE DERIVATION

Expectation-step: Choosing the optimal q(z | x, θ)

- Turns out that q(z|x,θ) = p(z|x,θ^t) is the best

$$p(z_i = c|x_i,\theta^t) = \frac{p(x_i,z_i=c|\theta^t)}{\sum_c p(x_i,z_i=c|\theta^t)} = \frac{\pi(z_i=c)\, N(x_i|\mu_c,\sigma_c)}{\sum_d \pi(z_i=d)\, N(x_i|\mu_d,\sigma_d)}$$



$p(z|x,\theta^t)$

Maximization-step: Choosing the optimal θ

- Auxiliary function separates into ***ECLL*** and ***entropy*** term:

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)} \\ &= \sum_z q(z \mid x) \log p(x,z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) \\ &= \langle l_c(\theta;x,z)\rangle_q - \sum_z q(z \mid x) \log q(z \mid x),\end{aligned}$$

ECLL

Entropy

Maximization-step: Choosing the optimal θ

- Take derivatives to maximize the ***ECLL***:

$$\langle l_c(\theta;x,z)\rangle_q \triangleq \sum_x \sum_z q(z \mid x,\theta) \log p(x,z \mid \theta)$$

$$\frac{\partial \mathcal{L}(q^{t+1},\theta)}{\partial \theta} = \frac{\partial \langle l_c(\theta;x,z)\rangle_{q^{t+1}}}{\partial \theta} = \frac{\partial \sum_x \sum_z p^{t+1}(z|x,\theta^t) \log p(x,z|\theta)}{\partial \theta} = 0$$

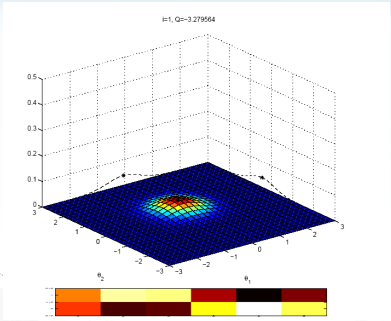
From E-step

Maximization-step: Choosing the optimal θ

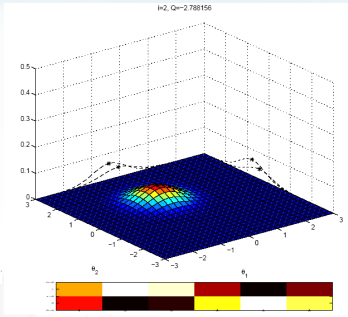
- Solution to the motivation example:
  - Calculate the new $\theta^{(t+1)}=(u_1^{(t+1)},u_2^{(t+1)})$, assuming the value taken by setting the hidden variable z as $p^{(t+1)}(z \mid x, \theta^t)$
  - Replace $\theta^t=(u_1^t,u_2^t)$ by $\theta^{(t+1)}=(u_1^{(t+1)},u_2^{(t+1)})$
  - Iterate…

$$\mu_c^{(t+1)} = \frac{\sum_x p(z_i = c|x_i,\theta^t)\, x_i}{\sum_x p(z_i = c|x_i,\theta^t)}$$

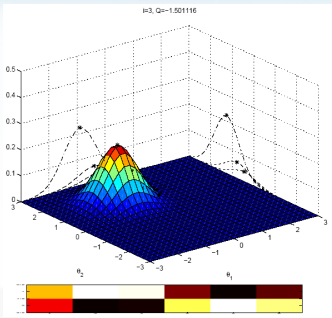
EM Illustration



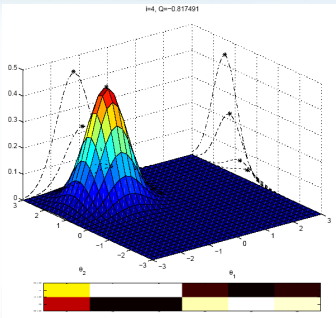
EM Illustration



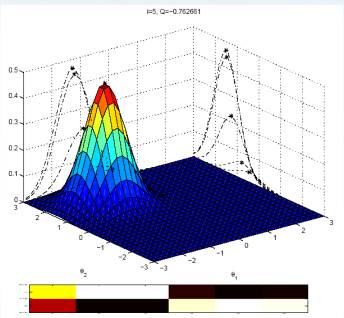
EM Illustration



EM Illustration



EM Illustration



EM Convergence

- When $\mathcal{L}(q, \theta)$ is continuous, EM algorithm converges to a maximum likelihood hypothesis
  - Guaranteed only to converge to a local maximum
  - Shares some of the same limitations as other methods such as gradient descent, line search, and conjugate gradient.

Seminar S04

- **Topics**
  - Present one specific example of the EM algorithm used to solve a real world problem.
  - Connection between EM and K-Means
- **Requirements**
  - Prepare a **15 minutes** talk on your chosen topic
  - Make **ppt** to assist your talk
  - Prepare **at least 3 questions** to ask the audience after your talk
  - Get ready to **take questions** from the audience
- **Hints**
  - You can search for research articles from Google Scholar
  - Or illustrate a solution step by step using EM

Questions?