EM Algorithm

Likelihood, Mixture Models and Clustering

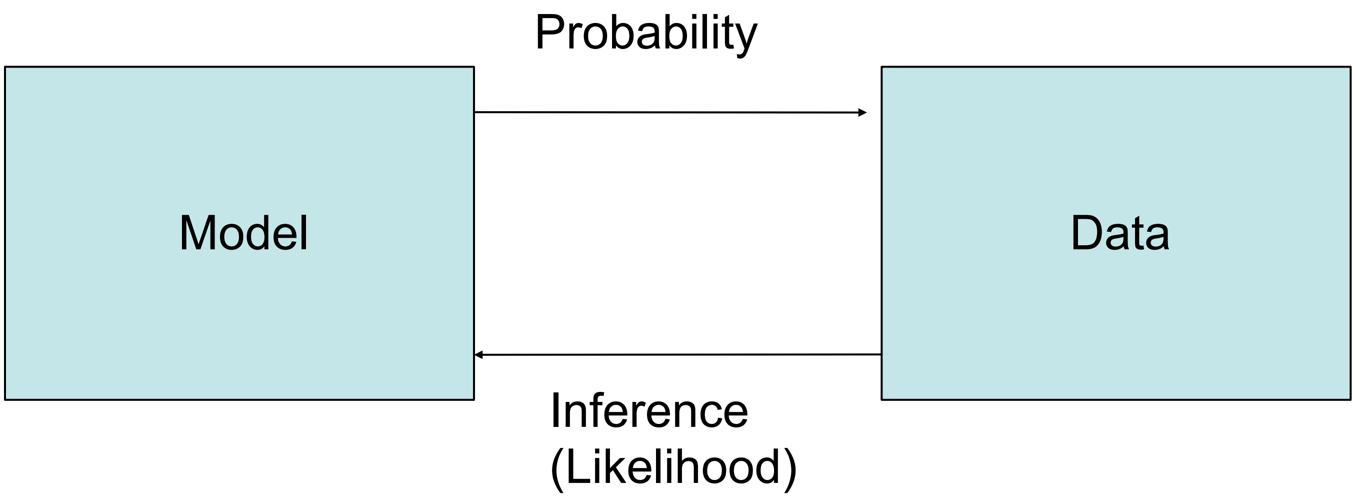
Introduction

- In the last class the K-means algorithm for clustering was introduced.
- The two steps of K-means: **assignment** and **update** appear frequently in data mining tasks.
- In fact a whole framework under the title "EM Algorithm" where EM stands for **Expectation** and **Maximization** is now a standard part of the data mining toolkit

Outline

- What is Likelihood?
- Examples of Likelihood estimation?
- Information Theory – Jensen Inequality
- The EM Algorithm and Derivation
- Example of Mixture Estimations
- Clustering as a special case of Mixture Modeling

Meta-Idea



A model of the data generating process gives rise to data.
Model estimation from data is most commonly through Likelihood estimation

From PDM by HMS

Likelihood Function

$$P(Model \mid Data) = \frac{P(Data \mid Model)P(Model)}{P(Data)}$$

Likelihood Function

Find the “best” model which has generated the data. In a likelihood function the data is considered fixed and one searches for the best model over the different choices available.

Model Space

- The choice of the model space is plentiful but not unlimited.
- There is a bit of “art” in selecting the appropriate model space.
- Typically the model space is assumed to be a linear combination of known probability distribution functions.

Examples

- Suppose we have the following data
  - 0,1,1,0,0,1,1,0
- In this case it is sensible to choose the Bernoulli distribution (B(p)) as the model space.

$$P(X = x) = p^x (1 - p)^{1-x}$$

- Now we want to choose the best p, i.e.,

$$\text{argmax}_p P(Data|B(p))$$

Examples

Suppose the following are marks in a course

55.5, 67, 87, 48, 63

Marks typically follow a Normal distribution whose density function is

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma}(x-\mu)^2}$$

Now, we want to find the best μ,σ such that

$$argmax_{\mu,\sigma} p(Data|\mu, \sigma)$$

Examples

- Suppose we have data about heights of people (in cm)
  - 185,140,134,150,170
- Heights follow a normal (log normal) distribution but men on average are taller than women. This suggests a **mixture** of two distributions

$$\pi_1 N(\mu_1, \sigma_1) + \pi_2 N(\mu_2, \sigma_2)$$

Maximum Likelihood Estimation

- We have reduced the problem of selecting the best model to that of selecting the best parameter.
- We want to select a parameter p which will **maximize** the probability that the data was generated from the model with the parameter p plugged-in.
- The parameter **p** is called the maximum likelihood estimator.
- The maximum of the function can be obtained by setting the derivative of the function ==0 and solving for p.

Two Important Facts

- If $A_1,\ldots,A_n$ are independent then

$$P(A_1,\ldots,A_n)=\prod_{i=1}^{n}P(A_i)$$

- The log function is monotonically increasing. x · y ! Log(x) · Log(y)

- Therefore if a function f(x) >= 0, achieves a maximum at x1, then log(f(x)) also achieves a maximum at x1.

Example of MLE

$$\begin{aligned} L(p) &= P(0,1,1,0,0,1,0,1|p) \\ &= P(0|p)P(1|p)\ldots P(1|p) \\ &= (1-p)p\ldots p \\ &= p^4(1-p)^4 \end{aligned}$$

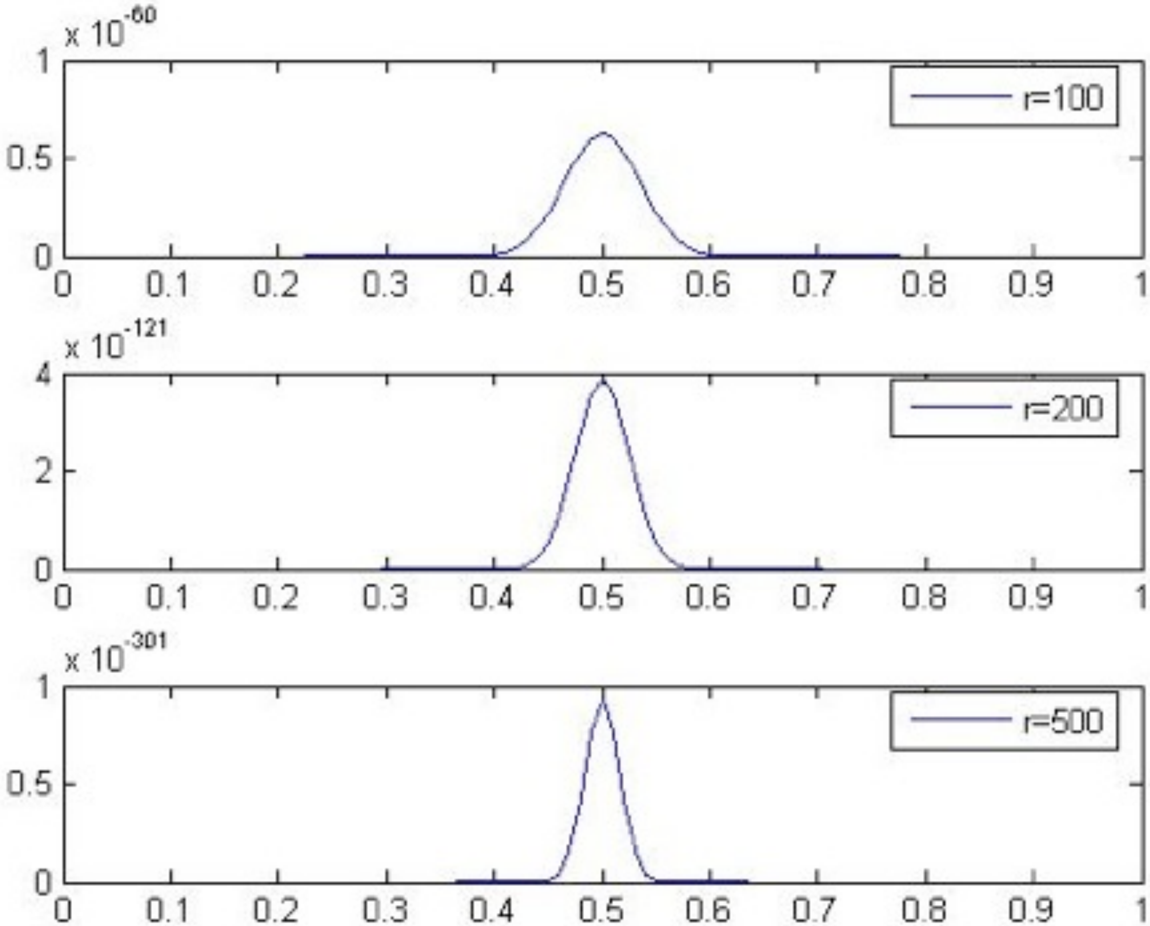
- Now, choose p which maximizes L(p). Instead we will maximize l(p)= LogL(p)

$$\begin{aligned} \ell(p) &= logL(p) = 4log(p) + 4log(1-p) \\ \frac{d\ell(p)}{dp} &= \frac{4}{p} - \frac{4}{1-p} \equiv 0 \\ \rightarrow p &= \frac{1}{2} \end{aligned}$$

Properties of MLE

- There are several technical properties of the estimator but lets look at the most intuitive one:
  - As the number of data points increase we become more sure about the parameter p

Properties of MLE



r is the number of data points. As the number of data points increase the confidence of the estimator increases.

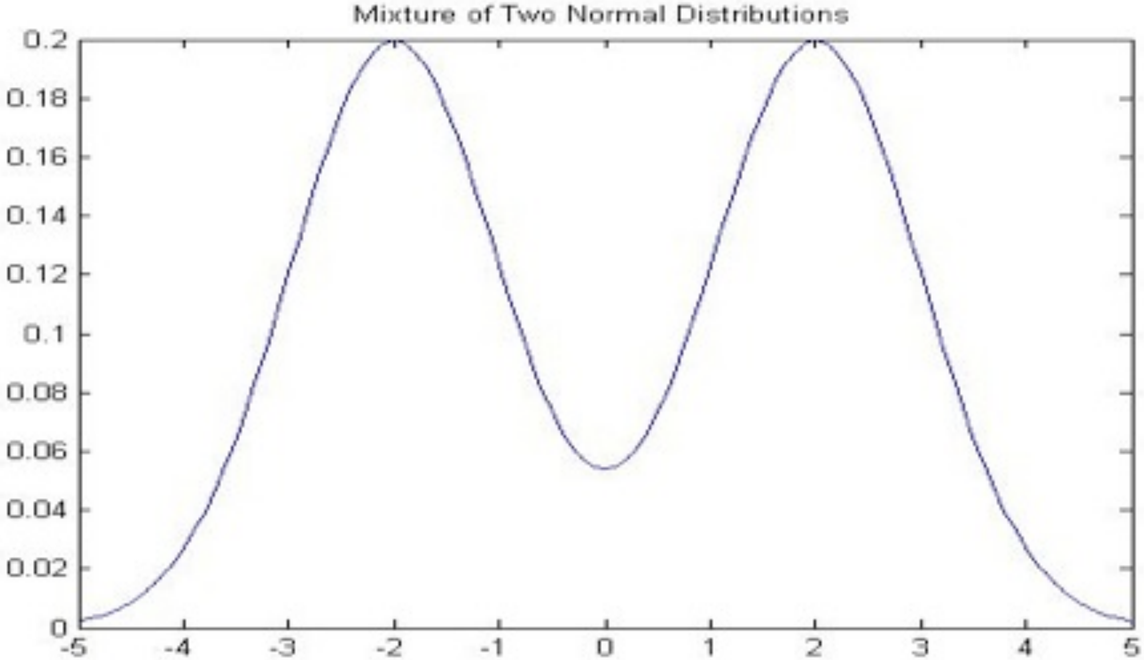
Matlab commands

- [phat,ci]=mle(Data,’distribution’,’Bernoulli’);
- [phi,ci]=mle(Data,’distribution’,’Normal’);

MLE for Mixture Distributions

- When we proceed to calculate the MLE for a mixture, the presence of the sum of the distributions prevents a “neat” factorization using the log function.
- A completely new rethink is required to estimate the parameter.
- The new rethink also provides a solution to the clustering problem.

A Mixture Distribution



$$f(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2) = \pi_1 N(\mu_1, \sigma_1) + \pi_2 N(\mu_2, \sigma_2)$$

Missing Data

- We think of clustering as a problem of estimating missing data.
- The missing data are the cluster labels.
- Clustering is only one example of a missing data problem. Several other problems can be formulated as missing data problems.

Missing Data Problem

- Let D = {x(1),x(2),…x(n)} be a set of n observations.
- Let H = {z(1),z(2),..z(n)} be a set of n values of a hidden variable Z.
  - z(i) corresponds to x(i)
- Assume Z is discrete.

EM Algorithm

- The log-likelihood of the observed data is

$$l(\theta) = \log p(D \mid \theta) = \log \sum_{H} p(D, H \mid \theta)$$

- Not only do we have to estimate θ but also H
- Let Q(H) be the probability distribution on the missing data.

EM Algorithm

$$
\begin{aligned}
\ell(\theta) &= log \sum_H p(D, H|\theta) \\
&= log \sum_H Q(H) \frac{P(D, H|\theta)}{Q(H)} \\
&\geq \sum_H Q(H) log \frac{P(D, H|\theta)}{Q(H)} \\
&= \sum_H Q(H) log p(D, H|\theta) + \sum_H Q(H) log \frac{1}{Q(H)} \\
&= F(Q, \theta)
\end{aligned}
$$

Inequality is because of Jensen's Inequality.
This means that the $F(Q,\theta)$ is a lower bound on $l(\theta)$

Notice that the log of sums is become a sum of logs

EM Algorithm

- The EM Algorithm alternates between maximizing F with respect to Q (theta fixed) and then maximizing F with respect to theta (Q fixed).

E-step $\quad Q^{k+1} = argmax_Q F(Q^k, \theta^k)$

M-step $\quad \theta^{k+1} = argmax_\theta F(Q^{k+1}, \theta^k)$

EM Algorithm

- It turns out that the E-step is just

$$Q^{k+1} = P(H|D, \theta^k)$$

- And, furthermore $\ell(\theta^k) = F(Q, \theta^k)$

- Just plug-in

EM Algorithm

- The M-step reduces to maximizing the first term with respect to θ as there is no θ in the second term.

$$\theta^{k+1} = argmax_{\theta} \sum_{H} p(H|D, \theta^{k}) logp(D, H|\theta^{k})$$

EM Algorithm for Mixture of Normals

$$
\begin{aligned}
f(x) &= \sum_{k=1}^{K} \pi_k f_k(x, \mu_k, \sigma_k) && \text{Mixture of Normals} \\
P(k|x) &= \frac{\pi_k f_k(x, \mu_k, \sigma_k)}{f(x)} && \text{E Step} \\
\pi_k &= \frac{1}{n} \sum_{i=1}^{n} P(k|x(i)) \\
\mu_k &= \frac{1}{n\pi_k} \sum_{i=1}^{n} P(k|x(i)) x(i) \\
\sigma_k &= \frac{1}{n\pi_k} \sum_{i=1}^{n} P(k|x(i)) (x(i) - \mu_k)^2
\end{aligned}
$$

M-Step

EM and K-means

- Notice the similarity between EM for Normal mixtures and K-means.

- The expectation step is the assignment.
- The maximization step is the update of centers.