法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 小象学院

Machine Learning

Part 5: Statistical Learning

Graphical Models

Subjective Probability

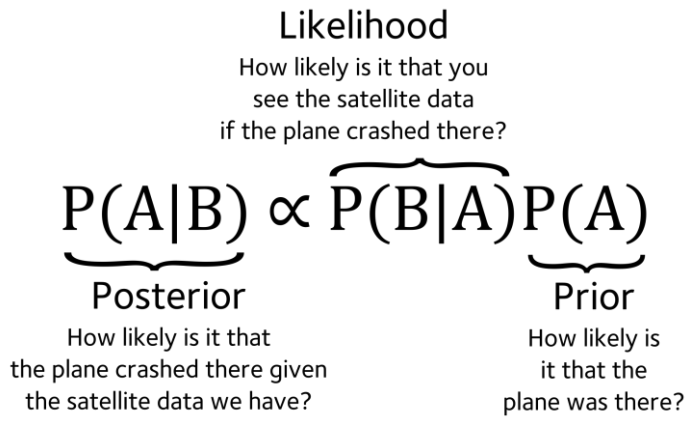
Frequentist: it defines an event's probability as the limit of its relative frequency in a large number of trials. This interpretation supports the statistical needs of experimental scientists and pollsters; probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion).

Bayesian: Instead of frequency of some phenomenon, probability is interpreted as reasonable expectation representing a state of knowledge or as quantification of a **personal belief.**

Example: (1) There is a high probability that Xiaoming's wife is beautiful! (2) The probability that I was married is high!

Bayesian Inference

(1) Malaysian Airline problem with Bayesian Inference.



(2) My ‘qipa’ roommate problem.

P(PlayGame|CinemaTicket) = P(CinemaTicket|PlayGame)P(PlayGame)

This intuition is already embedded in our mind!

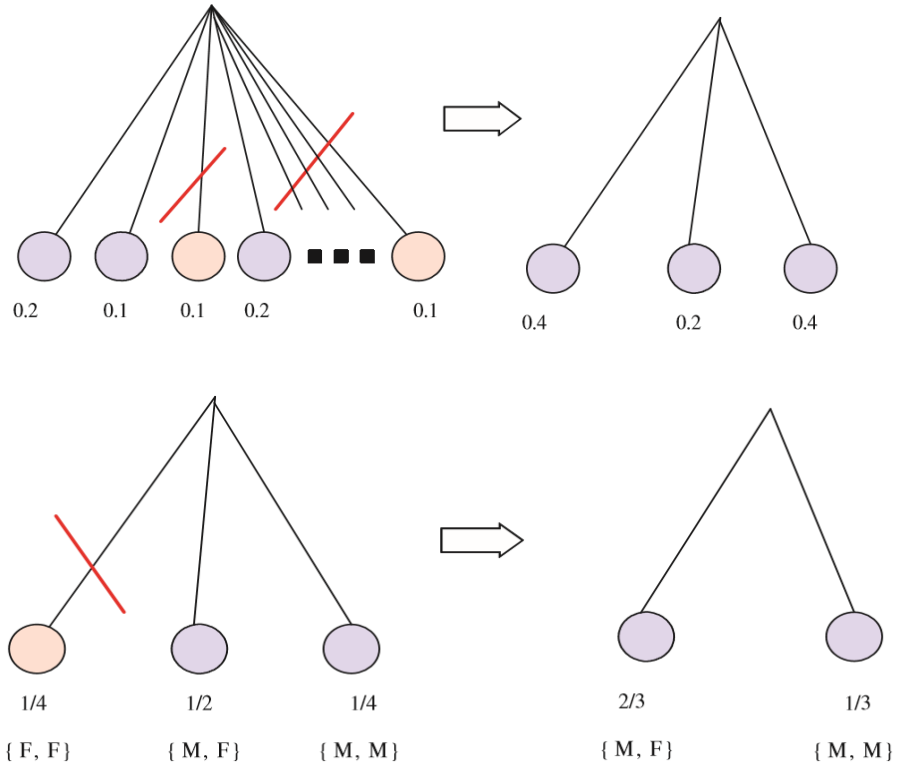
(3) Will sun rise tomorrow? (Alien on the Earth!)

P(SunRise) = P(SunRise|PastObservation)
P(SunRise|Observation) = P(Observation|SunRise) P(SunRise)

Probabilistic Intuition - 1

Problem 1. My neighbor has two children. It is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities - two boys or two girls - have probabilities 1/4 and 1/4 (this is obvious!).

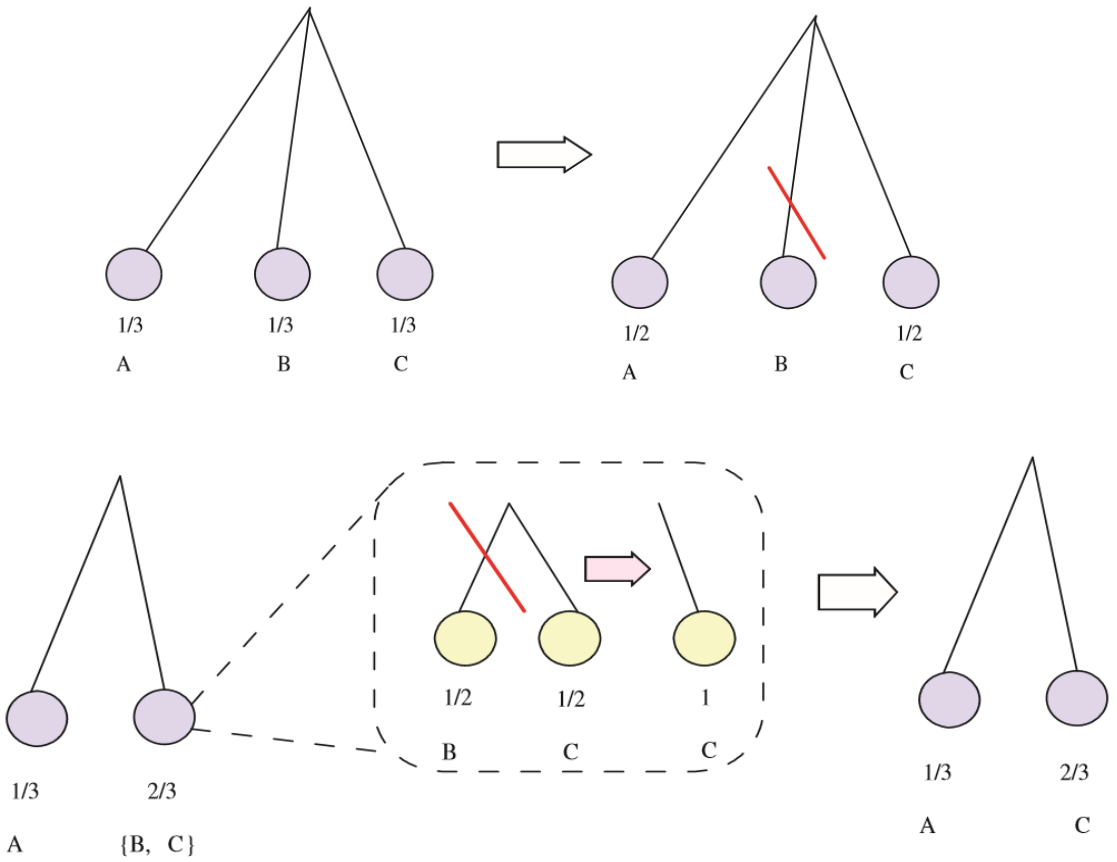
(1.1) If I ask him whether he/she has a boy, the answer is 'yes'. What is the probability of the other child is a girl?
(1.2) Suppose one day, I saw one of his/her children, it is a boy. What is the probability of the other child is a girl?



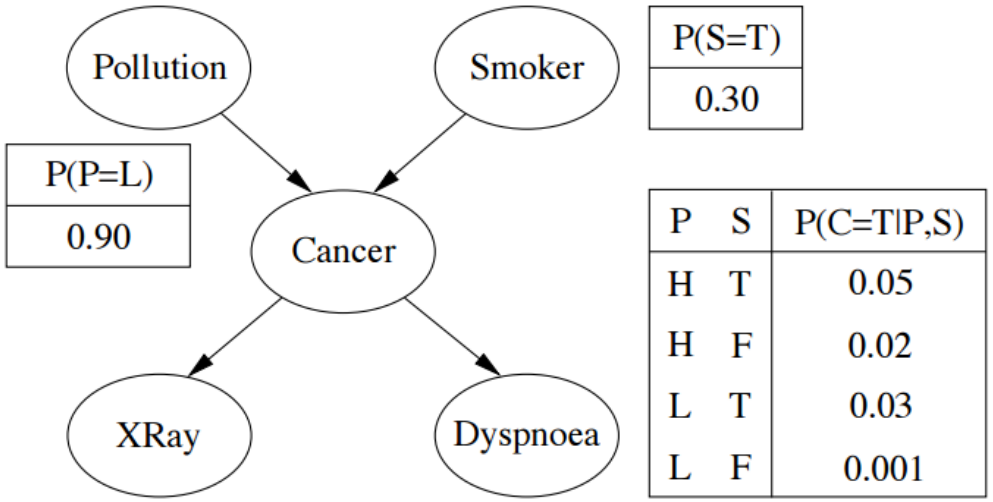
Probabilistic Intuition - 2

Problem 2. You are told that there is a million of dollars behind one of three doors A, B and C, and there is nothing behind the other two doors. Suppose you have chosen door A. You randomly select between door B and C and opened it. Say, door B is opened and there is nothing there. Will you stay with the same door A or change to the door C in order to get the money.

Problem 3. There is a million dollars behind one of three doors, we randomly select one door and open it. If we found it empty, what are the probabilities that the money is behind other two doors?



Bayesian Networks

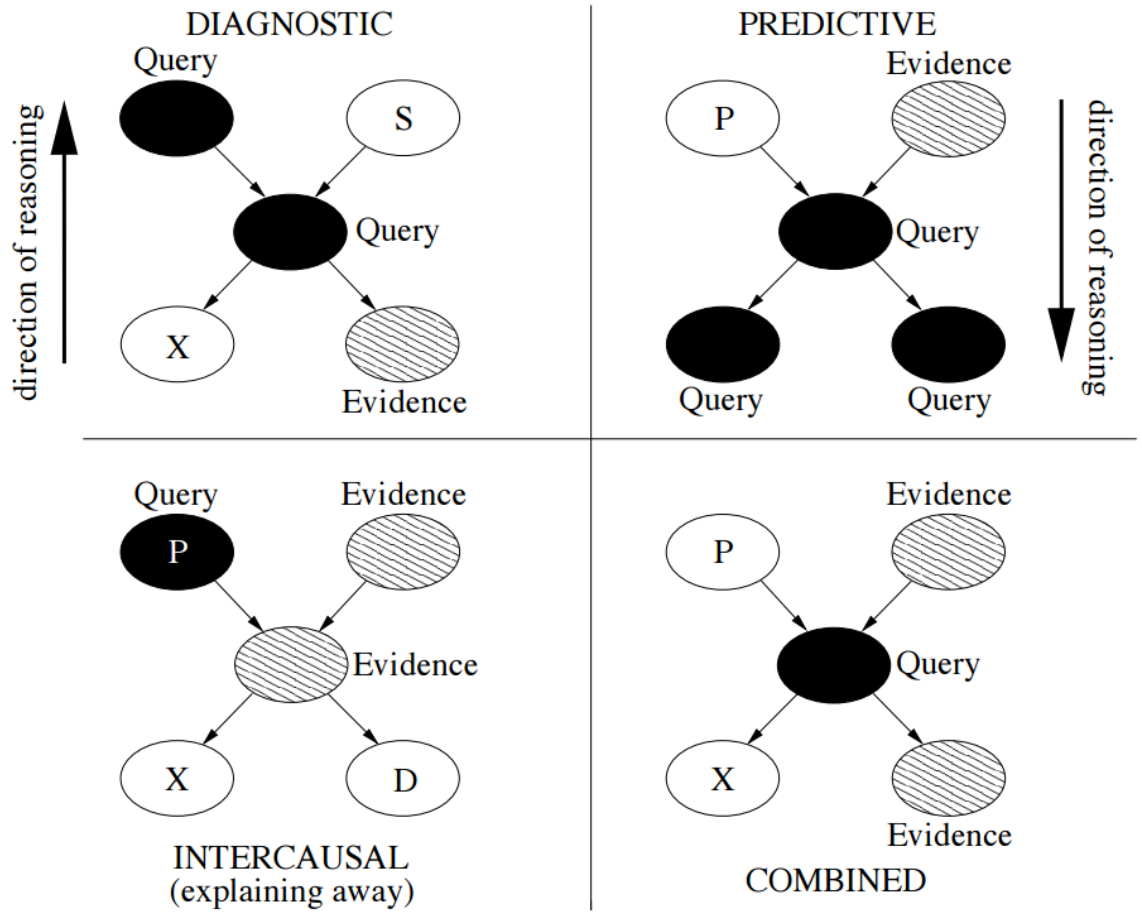


Given the Lung Cancer problem, the factors (or features) are shown in the table. The probabilistic relations are represented by a directed network. We may ask "what is the probability of this patient smokes if the Xray shows positive results"?

Preliminary choices of nodes and values for the lung cancer example.

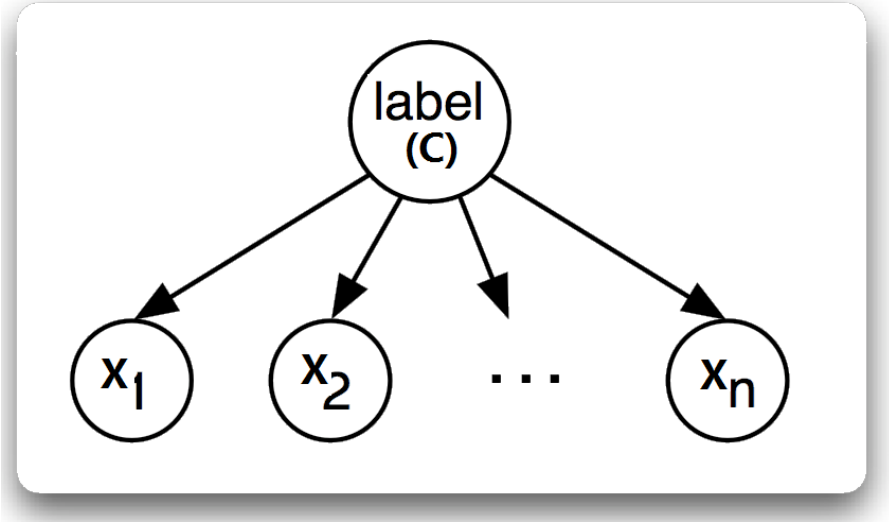
Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}

Bayesian Network Reasoning (Inference)

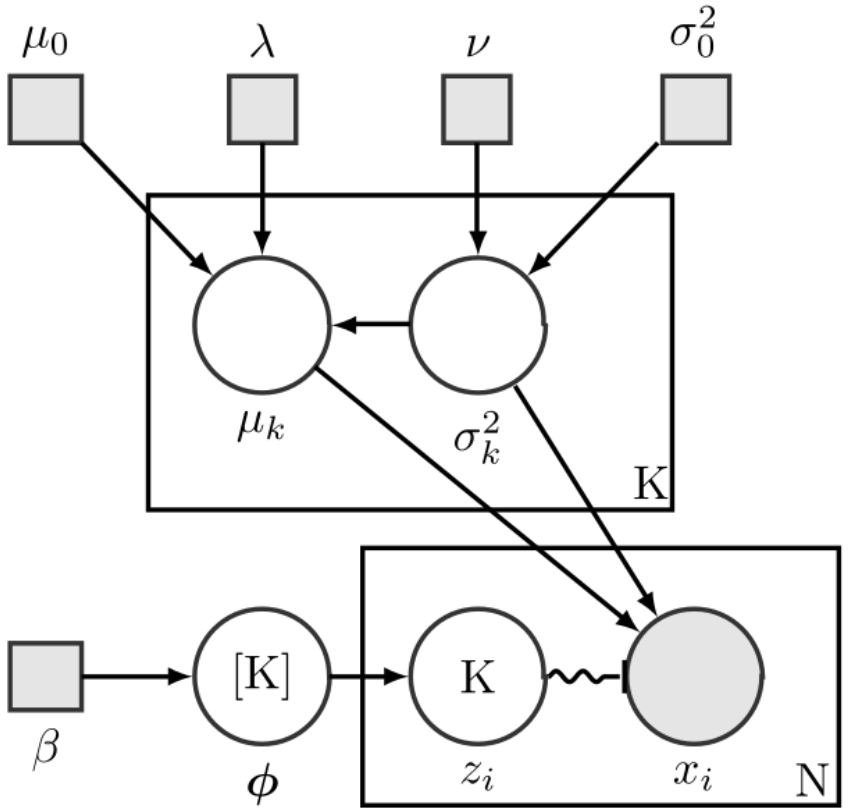


Four type of reasoning is introduced and it shows the internal probabilistic correlation of the random variables.

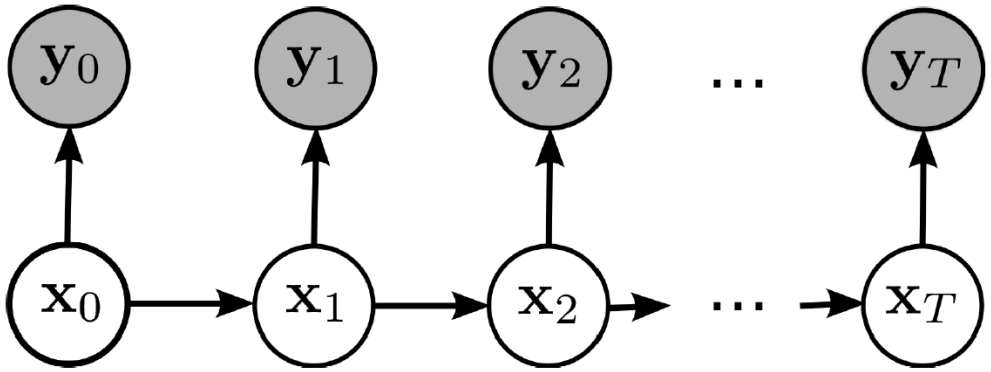
Hierarchical Graphical Model



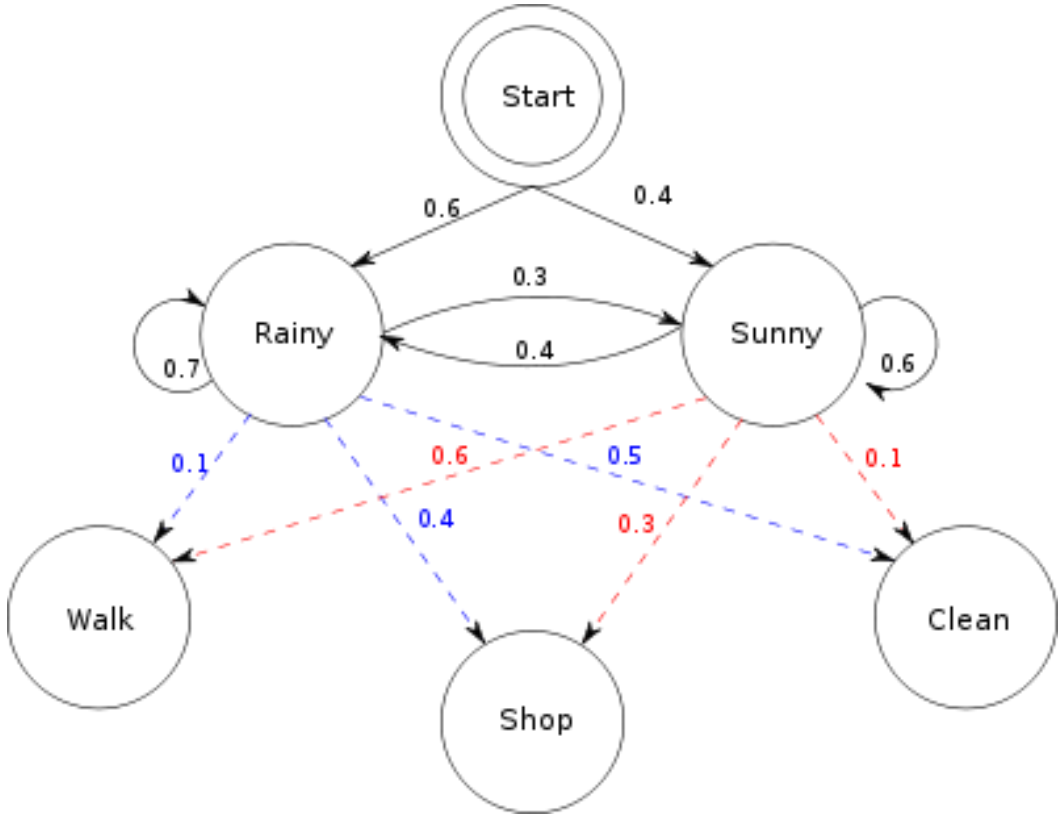
We can start from the simplest Naïve Bayes to complex models like mixture of Gaussians.



Hidden Markov Model

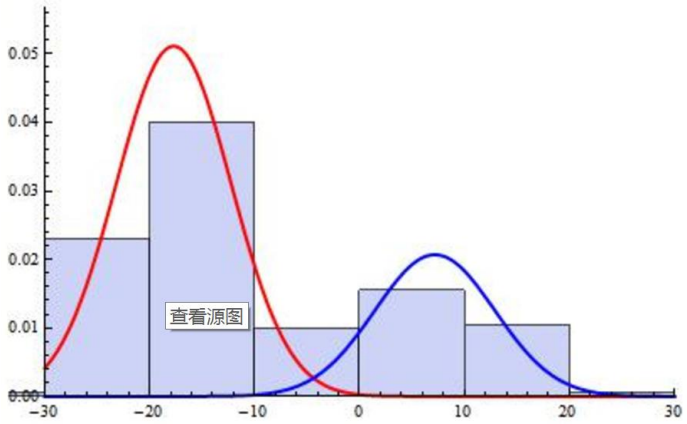


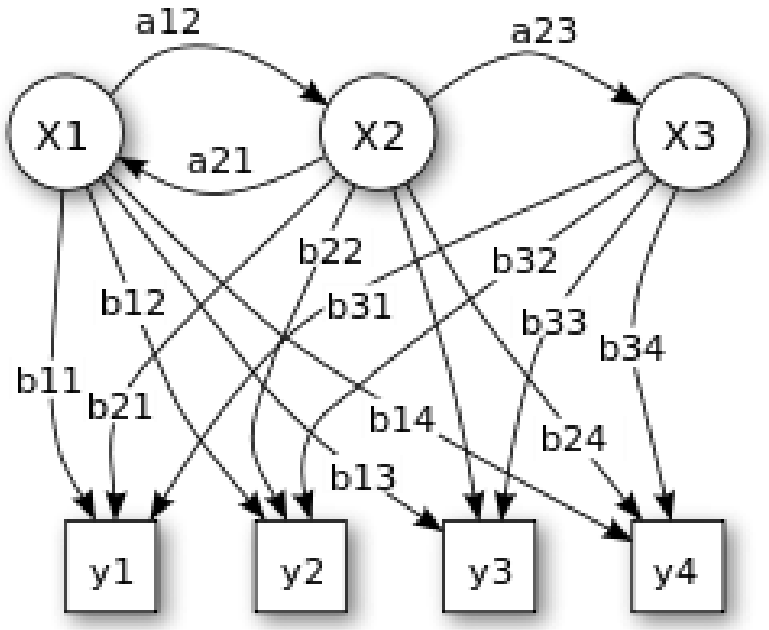
The state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible.



Mixture Model with Markov Property

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.



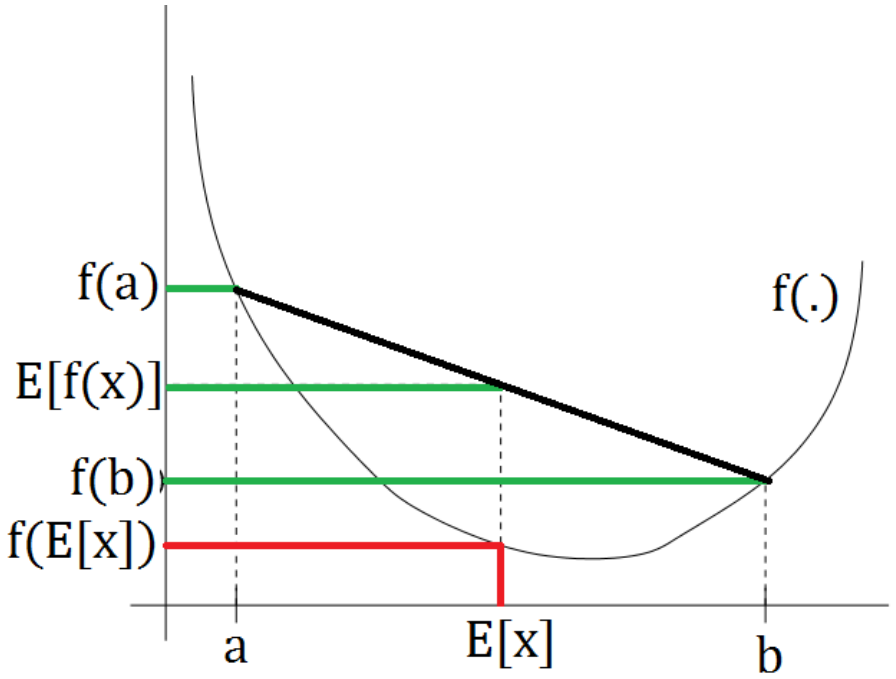


EM Algorithms

Jensen's Inequality

Jensen's inequality: Let f be a convex function, and let X be a random variable. Then: $E[f(x)] \geqslant f(E[x])$

Here, *f* is a **convex** function shown by the solid line. Incidentally, quite a lot of people have trouble remembering which way the inequality goes, and remembering a picture like this is a good way to quickly figure out the answer.



Maximum Likelihood

Suppose we have an estimation problem in which we have a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ consisting of m independent examples. We wish to fit the parameters of a model $p(x,z)$ to the data, where the likelihood is given by:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\ &= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta) \end{aligned}$$

Here, the $z^{(i)}$'s are the latent random variables; and it is often the case that if the $z^{(i)}$'s were observed, then maximum likelihood estimation would be easy.

In such a setting, the **EM Algorithm** gives an efficient method for maximum likelihood estimation.

Maximum Likelihood

Maximizing $\ell(\theta)$ explicitly might be difficult, and our strategy will be to instead repeatedly construct a lower-bound on ℓ (**E-step**), and then optimize that lower-bound (**M-step**).

$$\sum_i \log p(x^{(i)};\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) \quad (1)$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (2)$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (3)$$

$E[f(x)] \geqslant f(E[x])$ if f() is a convex function
$f(E[x]) \geqslant E[f(x)]$ if f() is a concave function

$\Rightarrow$ $log(E[x]) \geqslant E[log(x)]$

$$\Rightarrow f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right]$$

Tight Bound (Maximum Entropy)

There're many possible choices for the $Q^{(i)}$'s . Which should we choose to maximize the likelihood?

$$\mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right] = \mathrm{E}_{z^{(i)}\sim Q_i}\left[f(z^{(i)})\right]$$

To make the bound tight for a particular value of θ, we need for the step involving Jensen's inequality in our derivation above to hold with equality. For this to be true, we know it is sufficient that that the expectation be taken over a "constant"-valued random variable.

In order to obtain the maximum value, we need $f(z^{(1)}) = f(z^{(2)}) = \cdots = f(z^{(n)})$

$$\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = c \qquad \text{Which implies:} \qquad Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)};\theta)$$

Expectation Maximization

Since $\sum_z Q_i(z^{(i)}) = 1$ and $Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$ we can yield:

$$Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} = \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} = p(z^{(i)}|x^{(i)}; \theta)$$

From the above equation, we can find that the best Q is the posterior probability given current θ.

The **EM Algorithm**: Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta)$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \}$$

Expectation Maximization

At the time t, given $\theta^{(t)}$ we choose $Q^{(t)}$ based on the above equation.

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta^{(t)})$$

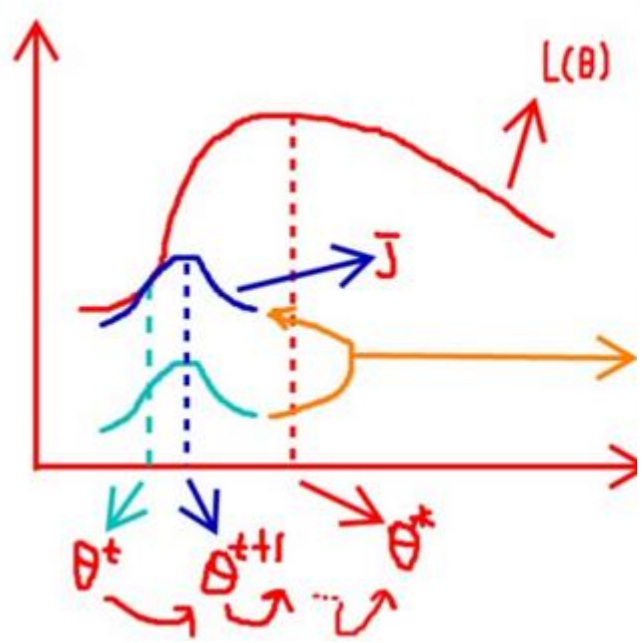
The current likelihood is:

$$\ell(\theta^{(t)}) = \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})}$$

The parameters $\theta^{(t+1)}$ are then obtained by maximizing the right hand side of the equation above. Thus,

$$\begin{aligned} \ell(\theta^{(t+1)}) &\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \\ &\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} = \ell(\theta^{(t)}) \end{aligned}$$

Illustration of EM Updating



Repeat until convergence: {

(E-step) For each i,j, set

$$w_j^{(i)} = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M-step) Update the parameters:

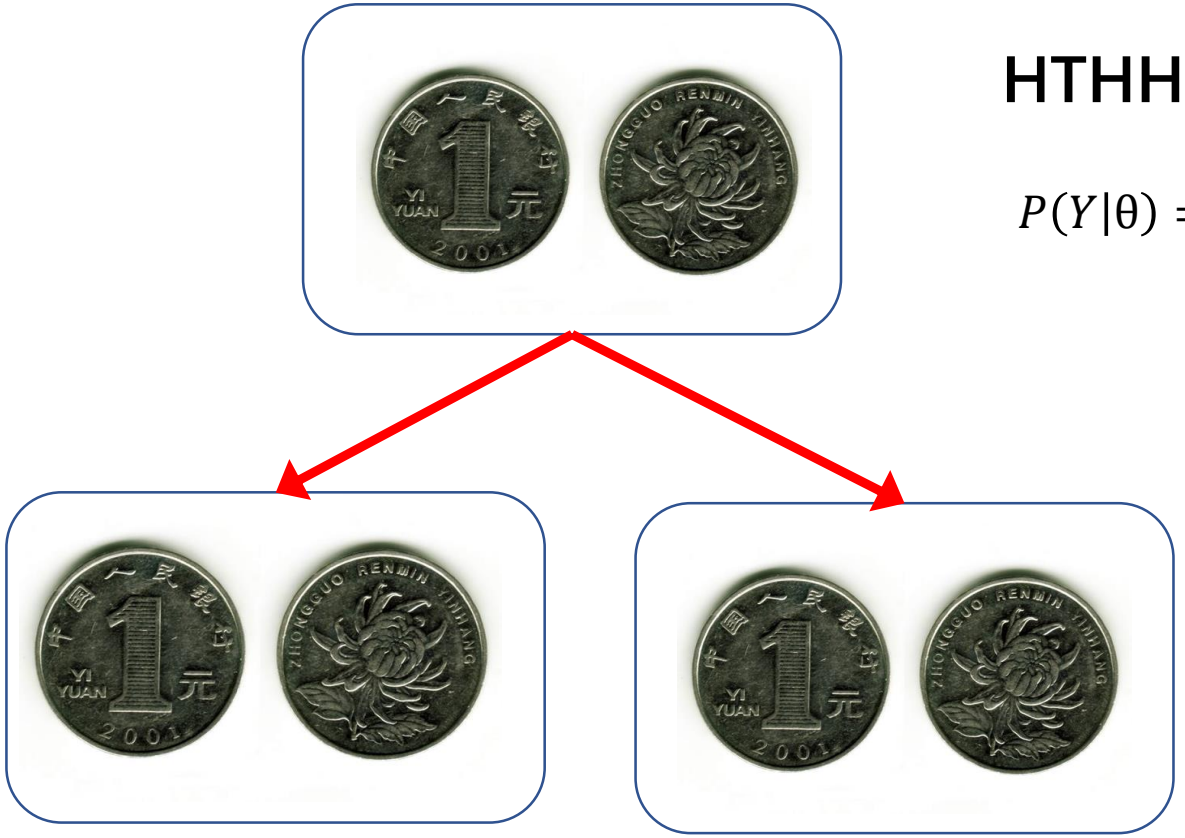
$$\phi_j = \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m}} w_j^{(i)}$$

}

3-Coin Problem



HTHHTTHHTHTHTTTHTHTTH……HT?

$$P(Y|\theta) = P(Y, Z|\theta)$$

Boys and Girls (Gaussian Mixture Model)

$$\begin{aligned}\ell(\phi,\mu,\Sigma) &= \sum_{i=1}^{m}\log p(x^{(i)};\phi,\mu,\Sigma)\\ &= \sum_{i=1}^{m}\log\sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)};\mu,\Sigma)p(z^{(i)};\phi)\end{aligned}$$

$$\ell(\phi,\mu,\Sigma)=\sum_{i=1}^{m}\log p(x^{(i)}|z^{(i)};\mu,\Sigma)+\log p(z^{(i)};\phi)$$

$$p(z^{(i)}=j|x^{(i)};\phi,\mu,\Sigma)=\frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{\sum_{l=1}^{k}p(x^{(i)}|z^{(i)}=l;\mu,\Sigma)p(z^{(i)}=l;\phi)}$$

Repeat until convergence: {

(E-step) For each i,j, set

$$w_j^{(i)}=p(z^{(i)}=j|x^{(i)};\phi,\mu,\Sigma)$$

(M-step) Update the parameters:

$$\phi_j=\frac{1}{m}\sum_{i=1}^{m}w_j^{(i)}$$

$$\mu_j=\frac{\sum_{i=1}^{m}w_j^{(i)}x^{(i)}}{\sum_{i=1}^{m}w_j^{(i)}}$$

$$\Sigma_j=\frac{\sum_{i=1}^{m}w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{m}}w_j^{(i)}$$

}