- Bayesian Belief Networks

. Naive Bayes assumption of conditional independence is too restrictive.

. But it is intractable without some such assumptions

. Bayesian belief networks describe conditional independence among subsets of variables.

-> allow combining prior knowledge about (in)dependencies among variables with observed training data

. Bayes belief networks are also called Bayes nets.

- Conditional Independence

. definition: X is conditionally independent of Y given Z if the probability distribution governing X is independent of the value of Y given the value of Z, that is, if

$$(\forall x_i, y_j, z_k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

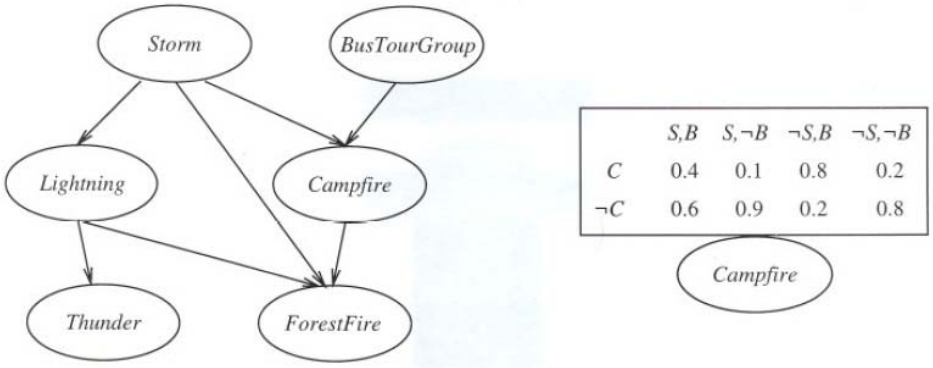
more compactly, we write

$$P(X|Y,Z) = P(X|Z)$$

. Example: Thunder is conditionally independent of Rain, given Lightning

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

- Bayesian Belief Network



. Each node represents a probability table of an attribute given its parents.

eg. $P(Campfire = T | Storm = T, BusTourGroup = T) = 0.4$

. Each branch in the graph represents the conditional dependence.

. Bayesian belief network represents the joint probability distribution over all variables.

eg. $P(Storm, BusTourGroup, \cdots, ForestFire)$

. In general,

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P(y_i | Parents(Y_i))$$

where $Parents(Y_i)$ denotes immediate predecessors of Y_i in graph.

. So, the joint distribution is fully defined by graph plus the conditional probability $P(y_i | Parents(Y_i))$.

- Inference in Bayesian Networks

. How can one infer the (probabilities of) values of one or more network variables, given observed values of others?

> Bayes net contains all information needed for this inference.

> If only one variable with unknown value, easy to infer it.

> In general case, problem is NP hard.

. In practice, can succeed in many cases

> Exact inference methods work well for some network structures.

> Monte Carlo methods simulate the network randomly to calculate approximate solutions.

- Learning of Bayesian Networks

. Several variants of this learning task

> Network structure might be known or unknown.

> Training examples might provide values of all network variables, or just some.

> If structure known and observe all variables, then it is easy as training a naive Bayes classifier.

> Suppose structure known, variables partially observable:

>> Similar to training neural network with hidden units.

>> In fact, can learn network conditional probability tables using gradient ascent.

>> Converge to network h that (loaclly) maximizes $P(D|h)$.

- Gradient Ascent for Bayes Nets

. Let w_{ijk} denote one entry in the conditional probability table for variable Y_i in the network, that is,

$$w_{ijk} = P(Y_i = y_{ij} | Parents(Y_i) = the\ list\ u_{ik}\ of\ values)$$

for example, if $Y_i = Campfire$, then u_{ik} might be

$\langle Storm = T, BusTourGroup = F \rangle$.

. The update rule for w_{ijk}:

$$w_{ijk}^{\neq w} = w_{ijk}^{old} + \eta \frac{\partial \ln P(D|h)}{\partial w_{ijk}}$$

Let $P_h(D) = P(D|h)$. Then,

$$\begin{aligned}
\frac{\partial \ln P_h(D)}{\partial w_{ijk}} &= \frac{\partial}{\partial w_{ijk}} \ln \prod_{d \in D} P_h(d) \\
&= \sum_{d \in D} \frac{\partial \ln P_h(d)}{\partial w_{ijk}} \\
&= \sum_{d \in D} \frac{1}{P_h(D)} \frac{\partial P_h(d)}{\partial w_{ijk}}
\end{aligned}$$

Let $u_i = Parents(Y_i)$. Then,

$$\begin{aligned}
\frac{\partial P_h(D)}{\partial w_{ijk}} &= \frac{\partial}{\partial w_{ijk}} \sum_{j',k'} P_h(d|y_{ij'}, u_{ik'}) P_h(y_{ij'}, u_{ik'}) \\
&= \frac{\partial}{\partial w_{ijk}} \sum_{j',k'} P_h(d|y_{ij'}, u_{ik'}) P_h(y_{ij'}|u_{ik'}) P_h(u_{ik'})
\end{aligned}$$

Since $P_h(y_{ij}|u_{ik}) = w_{ijk}$,

$$\begin{aligned}\frac{\partial \ln P_h(D)}{\partial w_{ijk}} &= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial}{\partial w_{ijk}} P_h(d|y_{ij},u_{ik}) P_h(y_{ij},u_{ik}) \\ &= \sum_{d \in D} \frac{1}{P_h(d)} \frac{P_h(y_{ij},u_{ik}|d) P_h(d) P_h(u_{ik})}{P_h(y_{ij},u_{ik})} \\ &= \sum_{d \in D} \frac{P_h(y_{ij},u_{ik}|d)}{P_h(y_{ij}|u_{ik})} \\ &= \sum_{d \in D} \frac{P_h(y_{ij},u_{ik}|d)}{w_{ijk}}\end{aligned}$$

Therefore,

$$w_{ijk}^{\neq w} = w_{ijk}^{old} + \eta \sum_{d \in D} \frac{P_h(y_{ij},u_{ik}|d)}{w_{ijk}}.$$

Algorithm:

Perform the following gradient ascent repeatedly:

Step 1. Update all w_{ijk} using training data D :

$$w_{ijk}^{\neq w} = w_{ijk}^{old} + \eta \sum_{d \in D} \frac{P_h(y_{ij},u_{ik}|d)}{w_{ijk}}$$

Step 2. Then, normalize the w_{ijk} to assure

(1) $0 \leqq w_{ijk} \leqq 1$ and

(2) $\sum_j w_{ijk} = 1$.

- Expectation and Maximization (EM)

. When to use

> Data are only partially observable

> Unsupervised clustering (target value unobservable)

> Supervised learning (some instance attribute unobservable)

. Applications:

> Training Bayesian belief networks

> Unsupervised clustering

> Learning Hidden Markov Models

- EM algorithm

. Converges to local maximum likelihood h and provides estimates of hidden variables z_{ij}

. In fact, local maximum in $E[\ln P(Y|h)]$

> Y is complete (observable plus unobservable variables) data.

> Expected value is taken over possible values of unobserved variables in Y.

- General EM Problem

. Given:

> Observed data $X = \{x_1, \cdots, x_m\}$

> Unobserved data $Z = \{z_1, \cdots, z_m\}$

> Parameterized probability distribution $P(Y|h)$

where

$Y = \{y_1, \cdots, y_m\}$ is the full data $y_i = x_i \cup z_i$ and

h are the parameters.

. Determine

> h that (locally) maximizes $E[\ln P(Y|h)]$.

- General EM Method

. Define likelihood function $Q(h'|h)$ which calculates $Y = X \cup Z$ using observed X and current parameters h to estimate Z :

$Q \leftarrow E[\ln P(Y|h')|h, X]$

. EM algorithm:

> Estimation (E) step: Calculate $Q(h'|h)$ using the current hypothesis h and the observed data X to estimate the probability distribution over Y, that is,

$Q \leftarrow E[\ln P(Y|h')|h, X]$.

> Maximization (M) step: Replace hypothesis h by the hypothesis h' that maximizes this Q function, that is,

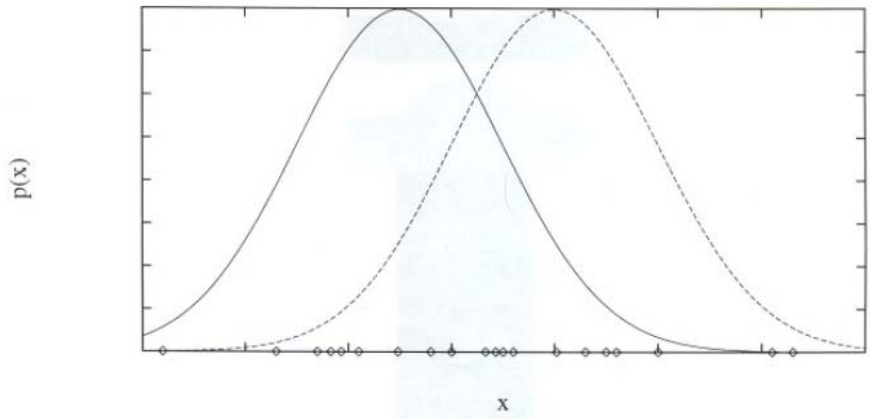
$h \leftarrow \arg\max_{h'} Q(h'|h)$.

- Generating Data from Mixture of k Gaussians

. Each instance x is generated by

> Choosing one of the k Gaussians with uniform probability

> Generating an instance at random according to that Gaussian



- EM for Estimating k Means

. Given:

> Instances from X generated by the mixture of k Gaussian distributions

> Unknown means $\langle \mu_1, \cdots, \mu_k \rangle$ of the k Gaussians

> Don't know which instance x_i was generated by which Gaussian

. Determine:

> Maximum likelihood estimates of $\langle \mu_1, \cdots, \mu_k \rangle$

- EM for Estimating k Means

. Think of full description of each instance as

$y_i = \langle x_i, z_{i1}, z_{i2} \rangle$

where z_{ij} is 1 if x_i is generated by the jth Gaussian

Here, note that x_i is observable and z_{ij} is unobservable.

. EM algorithm:

Pick the initial $h = \langle \mu_1, \mu_2 \rangle$ at random and then

iterate the E and M steps.

. E Step: Calculate the expected value $E[z_{ij}]$ of

each hidden variable z_{ij}, assuming that the current hypothesis

$h\langle \mu_1, \mu_2 \rangle$ holds, that is,

$$E[z_{ij}] = \frac{p(x = x_i | \mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i | \mu = \mu_n)} = \frac{\exp(-(x_i - \mu_j)^2/(2\sigma^2))}{\sum_{n=1}^{2} \exp(-(x_i - \mu_n)^2/(2\sigma_n^2))}.$$

. M Step: Calculate a new maximum likelihood hypothesis

$h' = \langle \mu_1', \mu_2' \rangle$ assuming that the value taken on by each

hidden variable z_{ij} is its expected value $E[z_{ij}]$ calculated above:

$$\mu_j' \leftarrow (\sum_{i=1}^{m} E[z_{ij}] x_j) / (\sum_{i=1}^{m} E[z_{ij}]), \quad j = 1, 2.$$

- EM for estimating k means

. Let $h' = \langle \mu_1', \cdots, \mu_k' \rangle$. Then,

$$P(y_i|h') = P(x_i, z_{i1}, \cdots, z_{ik}|h') = \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{k} z_{ij}(x_i - \mu_j')^2) \quad \text{and}$$

$$\begin{aligned} \ln P(Y|h') &= \ln \prod_{i=1}^{m} P(y_i|h') \\ &= \sum_{i=1}^{m} \ln P(y_i|h') \\ &= \sum_{i=1}^{m} (\ln \frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} z_{ij}(x_i - \mu_j')^2) \end{aligned}$$

Therefore,

$$Q(h'|h) = E[\ln P(Y|h')|h, X] = \sum_{i=1}^{m} (\ln \frac{1}{\sqrt{2\pi}\,\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} E[z_{ij}](x_i - \mu_j')^2).$$

Here, from definition,

$$E[z_{ij}] = \frac{p(x = x_i|\mu = \mu_j)}{\sum_{n=1}^{k} p(x = x_i|\mu = \mu_n)} = \frac{\exp(-(x_i - \mu_j)^2/(2\sigma^2))}{\sum_{n=1}^{k} \exp(-(x_i - \mu_n)^2/(2\sigma_n^2))}.$$

(Expectation Step)

. We need to select h' maximizing $Q(h'|h)$, that is,

$$\text{argmax}_{h'} Q(h'|h) = \text{argmax}_{h'} \sum_{i=1}^{m} (\ln \frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} E[z_{ij}](x_i - \mu_j')^2).$$

Alternatively,

$$\text{argmax}_{h'} Q(h'|h) = \text{argmin}_{h'} \sum_{i=1}^{m} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu_j')^2.$$

From the condition of

$$\frac{\partial}{\partial \mu_j'} \sum_{i=1}^{m} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu_j')^2 = 0, \quad j = 1, \cdots, k,$$

$Q(h'|h)$ is minimized when

$$\mu_j' = \frac{\sum_{i=1}^{m} E[z_{ij}] x_i}{\sum_{i=1}^{m} E[z_{ij}]}.$$

(Maximization Step)

- EM algorithm for the training of Bayes Nets

. EM (expectation and maximization) algortihm can be used repeatedly:

Step 1. Calculate probabilities of unobserved variables, assuming h.

Step 2. Calculate new w_{ijk} to maximize $E[\ln P(D|h)]$ where D now includes both observed and (calculated probabilities of) unobserved variables.

. When structure unkown

> Algorithms use greedy search to add/subtract edges and nodes. (active research topic)

- Inference in Bayes Nets

. Suppose we have a Bayes net, complete with conditional probabilities, and know the values or probabilities of some of the states, we will be able to determine the maximum posterior value of the unknown variables in the net.

. The belief of a set of propositions $\underline{x}=(x_1, x_2, \cdots)$ on node X describe the relative probabilities of the variables given all the evidence $\underline{e}$ throughout the rest of the network, that is, $P(\underline{x}|\underline{e})$.

. Furthermore, we can divide the dependency of the belief upon the parents and the children in the following way:

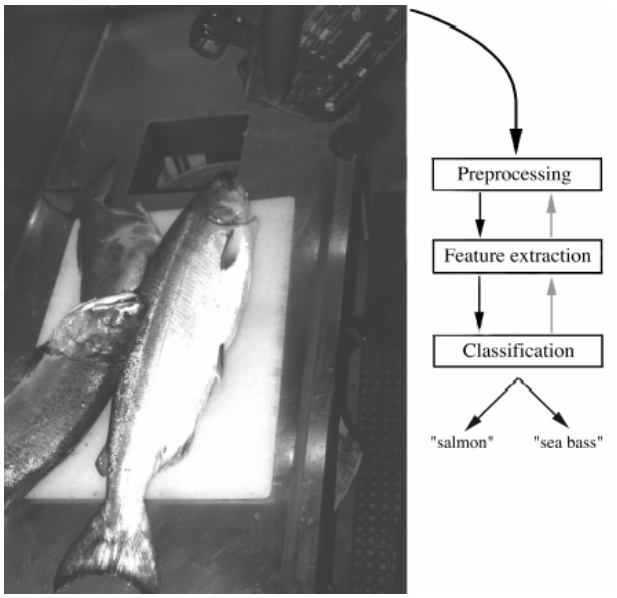
$$P(\underline{x}|\underline{e}) \propto P(\underline{e^C}|\underline{x})P(\underline{x}|\underline{e^P})$$

where $\underline{e}$ represents all evidence (i.e., values of variables on nodes other than X), $\underline{e^P}$ the evidence of parents nodes, and $\underline{e^C}$ the children nodes.

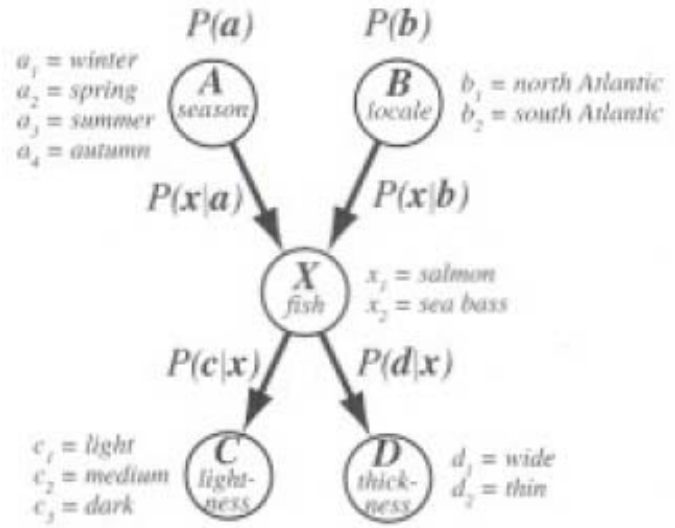
- Example: classifying fish (salmon, sea bass)

. We have the following picture (evidence).

Is this fish salmon or sea bass?



. Bayes Net:



A and B are parents of X while C and D are children of X.

. Probability tables

> $P(x_i|a_j)$:

	salmon	sea bass
winter	0.9	0.1
spring	0.3	0.7
summer	0.4	0.6
autumn	0.8	0.2

> $P(x_i|b_j)$:

	salmon	sea bass
north	0.65	0.35
south	0.25	0.75

> $P(c_i|x_j)$:

	light	medium	dark
salmon	0.33	0.33	0.34
sea bass	0.8	0.1	0.1

> $P(d_i|x_j)$:

	wide	thin
salmon	0.4	0.6
sea bass	0.95	0.05

. We have the following evidence $\underline{e} = \{e_A, e_B, e_C, e_D\}$:

winter: $P(a_1|e_A) = 1$, $P(a_i|e_A) = 0$ for $i = 2,3,4$.

prefers to fish in the south atlantic: $P(b_1|e_B) = 0.2$, $P(b_2|e_B) = 0.8$.

fairly light: $P(e_C|c_1) = 1$, $P(e_C|c_2) = 0.5$, $P(e_C|c_3) = 0$.

no information on the width: $P(e_D|d_1) = P(e_D|d_2)$.

. What will you conclude about the fish in the picture?

That is, what is the probability of $P(x_1|\underline{e})$ and $P(x_2|\underline{e})$?

First. let us consider the expression of $P(\underline{x}|\underline{e})$.

$$
\begin{aligned}
P(\underline{x}|\underline{e}) &= P(\underline{x}|\underline{e^c}\,\underline{e^p}) \\
&= \frac{P(\underline{e^c}\,\underline{e^p}|\underline{x})P(\underline{x})}{P(\underline{e^c}\,\underline{e^p})} \\
&= \frac{P(\underline{e^c}|\underline{e^p}\,\underline{x})P(\underline{e^p}|\underline{x})P(\underline{x})}{P(\underline{e^c}\,\underline{e^p})} \\
&= \frac{P(\underline{e^c}|\underline{e^p}\,\underline{x})P(\underline{x}|\underline{e^p})P(\underline{e^p})}{P(\underline{e^c}\,\underline{e^p})}
\end{aligned}
$$

Thus, we can say $P(\underline{x}|\underline{e}) \propto P(\underline{e^c}|\underline{x})P(\underline{x}|\underline{e^p})$.

$$P(x_1|\underline{e^p}) = \sum_{i=1}^{4}\sum_{j=1}^{2} P(x_1|a_i b_j)P(a_i|e_A)P(b_j|e_B) = 0.82$$

$$P(x_2|\underline{e^p}) = \sum_{i=1}^{4}\sum_{j=1}^{2} P(x_2|a_i b_j)P(a_i|e_A)P(b_j|e_B) = 0.18$$

$$P(\underline{e}^c|x_1) = P(e_C|x_1)P(e_D|x_1)$$
$$= \left(\sum_{i=1}^{3} P(e_C|c_i)P(c_i|x_1)\right)\left(\sum_{j=1}^{2} P(e_D|d_i)P(d_i|x_1)\right)$$
$$= 0.495$$
$$P(\underline{e}^c|x_2) = P(e_C|x_2)P(e_D|x_2)$$
$$= \left(\sum_{i=1}^{3} P(e_C|c_i)P(c_i|x_2)\right)\left(\sum_{j=1}^{2} P(e_D|d_i)P(d_i|x_2)\right)$$
$$= 0.85$$

That is,

$$P(x_1|\underline{e}) \propto P(\underline{e}^c|x_1)P(x_1|\underline{e}^p) = 0.82 \cdot 0.495$$
$$P(x_2|\underline{e}) \propto P(\underline{e}^c|x_2)P(x_2|\underline{e}^p) = 0.18 \cdot 0.85$$

Therefore,

$$P(x_1|\underline{e}) = \frac{0.82 \cdot 0.495}{0.82 \cdot 0.495 + 0.18 \cdot 0.85} = 0.726$$
$$P(x_2|\underline{e}) = \frac{0.18 \cdot 0.85}{0.82 \cdot 0.495 + 0.18 \cdot 0.85} = 0.274$$

That is, the most probable outcome is

$x_1 = salmon$!

- Summary: Bayesian Belief Networks

. Combine prior knowledge with observed data

. Impact of prior knowledge (when correct!) is to lower the sample complexity

. Active research area

> Extend from boolean to real-valued variables

> Parameterized distributions instead of tables

> More effective inference methods

> Configuring the graphic model from data

> ...