Parameter Estimation

Learning

Professor Ajoodha

Lecture 7

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

ExplainableAI Lab
— MODELLING. DECISION MAKING. CAUSALITY —

Problem Statement

- What if we are **not given a model**?
- Then manually construct a graphical model with an expert.
- Knowledge acquisition from experts is a **nontrivial** task:
  1. Amount of knowledge is too large
  2. Experts time is too valuable
  3. Perhaps no expert has sufficient understanding of domain
  4. Properties of distribution changes over time
- We would we want to learn the model?
  1. Density Estimation
  2. Knowledge Discovery

Goals of Learning: Density Estimation

- We can learn the model for inference.
- We want $\tilde{\mathcal{M}}$ which models $\tilde{P}$ **as closely to** P^*.
- Relative entropy can measure "as closely to":

$$\mathbb{D}(P^* \parallel \tilde{P}) = \mathbb{E}_{\xi \sim P^*}[\log(\frac{P^*(\xi)}{\tilde{P}(\xi)})],$$

  which is 0 if $\tilde{P} = P^*$, and positive otherwise.
- **Intuition:** Measures the extent of the compression in bits of using $\tilde{P}$ instead of P^*.
- Usually, P^* is unknown, so we calculate the **negative of the empirical log-loss** instead:

$$\log P(\mathcal{D} : \mathcal{M}) = \sum_{m=1}^{M} \log P(\xi[m] : \mathcal{M}).$$

Goals of Learning: Knowledge Discovery

- A different goal is for **knowledge discovery**.
- Learning P^* to discovery knowledge about P^*.
- This can reveal **properties of the domain**.
- We want the model $\mathcal{M}^*$, rather than some other model $\tilde{\mathcal{M}}$ that induces a distribution similar to $\mathcal{M}^*$.
- Even with large amounts of data, the true model may not be **identifiable**.
- Assessing prediction confidence is critical, considering **available data and potential hypotheses**.

Does your learned model capture P^*

- **Compare** learned model with the ground-truth.
- We cannot access the generating distribution of real-life data sets.
- Synthetic studies aid learning procedure comprehension but lack representativeness of **actual data properties**.
- Lets look at 3 key experimental protocols:
  1. Evaluating Generalisation Performance
  2. Selecting a Learning Procedure
  3. Goodness of Fit

(1) Evaluating Generalization Performance

- How can we **evaluate the performance** of a given model?
- Holdout testing (provides empirical estimate of risk relative to P^*):
  1. Randomly divide our data set into two disjoint sets: the training set $\mathcal{D}_{\texttt{train}}$ and test set $\mathcal{D}_{\texttt{test}}$.
  2. Learn the model using $\mathcal{D}_{\texttt{train}}$ (**with objective function**)
  3. Measure the performance using $\mathcal{D}_{\texttt{test}}$ (**with appropriate loss function**)

     K-fold cross validation: In each iteration holding as test data one partition and training from all the remaining instances.

(2) Selecting a Learning Procedure

- How do we **select a learning procedure** for an application?
- Specifically, choosing learning algorithms or algorithmic parameters?
- We can use a **validation set**:
  1. Firstly, learn a choice of the learning procedure on $\mathcal{D}_{\texttt{train}}$.
  2. Then use a separate unseen set ($\mathcal{D}_{\texttt{validation}}$) to evaluate different variants of our learning procedure and select the best performing model
  3. Finally, evaluate the final performance on $\mathcal{D}_{\texttt{test}}$.
- For very few samples use nested cross-validation schemes.

(3) Goodness of Fit

- Does learned model **completely capture** P^*?
- A goodness of fit strategy is as follows:
  1. Consider some property f of data sets, and evaluate $f(\mathcal{D}_{\texttt{train}})$
  2. Generate a set of synthetic data samples $\mathcal{D}$ from the learned model $\mathcal{M}$.
  3. evaluate $f(\mathcal{D}_{\texttt{synthetic}})$
- If $f(\mathcal{D}_{\texttt{train}})$ deviates significantly from $f(\mathcal{D}_{\texttt{synthetic}})$ then we can reject the hypothesis that $f(\mathcal{D}_{\texttt{train}})$ was generated from $\mathcal{M}$.
- f can be the negative of the empirical log-loss:

$$\log P(\mathcal{D} : \mathcal{M}) = \sum_{m=1}^{M} \log P(\xi[m] : \mathcal{M}).$$

Choices for f: Mean or variance for features, autocorrelation function, histogram of pixel values, a degree distribution, pairwise correlations, Entropy, Distribution of class labels, Clustering coefficient.

Parameters Estimation

Assumptions:

- That $\mathcal{D} = \{\xi[1], \ldots, \xi[M]\}$ is sampled from P^*.
- Instances are *independent and identically distributed* (IID).

Problem

How do we estimate the parameters in a Bayesian network?

Two basic approaches have been used:

1. Maximum likelihood estimation (MLE)
2. Bayesian estimation

Maximum Likelihood Estimation (MLE) Example

Parameter Estimation

Professor Ajoodha

Problem Statement

Maximum Likelihood Estimation

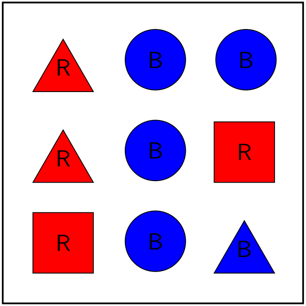
Bayesian Estimation

Partially Observed Data

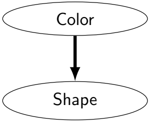
Expectation Maximisation

K-Means Clustering

Convergence



Color	
R	B
?	?



	Shape		
Color	T	S	C
R	?	?	?
B	?	?	?

$$L(\theta : \mathcal{D}) = \prod_i L_i(\theta_{X_i|Pa_{X_i}} : \mathcal{D})$$

$$L_i(\theta_{X_i|Pa_{X_i}} : \mathcal{D}) = \prod_m P(x_i[m] \mid pa_{X_i}[m] : \Theta_{X_i|Pa_{X_i}})$$

$$P(R) = \frac{M[R]}{M} = \frac{4}{9} = 0.4$$

$$P(B) = \frac{M[B]}{M} = \frac{5}{9} = 0.6$$

$$P(\blacktriangle \mid R) = \frac{M[\blacktriangle, R]}{M[R]} = \frac{2}{4} = 0.5$$

$$P(\blacktriangle \mid B) = \frac{M[\blacktriangle, B]}{M[B]} = \frac{1}{5} = 0.2$$

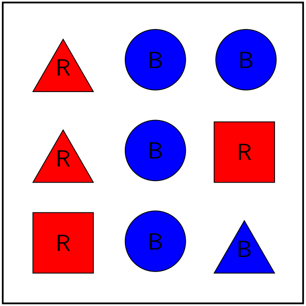
$$P(\blacksquare \mid R) = \frac{M[\blacksquare, R]}{M[R]} = \frac{2}{4} = 0.5$$

$$P(\blacksquare \mid B) = \frac{M[\blacksquare, B]}{M[B]} = \frac{0}{5} = 0$$

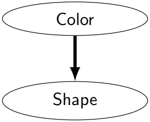
$$P(\bullet \mid R) = \frac{M[\bullet, R]}{M[R]} = \frac{0}{4} = 0$$

$$P(\bullet \mid B) = \frac{M[\bullet, B]}{M[B]} = \frac{4}{5} = 0.8$$

Maximum Likelihood Estimation (MLE): Example



Color	
R	B
0.4	0.6



	Shape		
Color	T	S	C
R	0.5	0.5	0
B	0.2	0	0.8

$$L(\theta : \mathcal{D}) = \prod_i L_i(\theta_{X_i|Pa_{X_i}} : \mathcal{D})$$

$$L_i(\theta_{X_i|Pa_{X_i}} : \mathcal{D}) = \prod_m P(x_i[m] \mid pa_{X_i}[m] : \Theta_{X_i|Pa_{X_i}})$$

$$\text{P(R)} = \frac{M[R]}{M} = \frac{4}{9} = 0.4$$

$$\text{P(B)} = \frac{M[B]}{M} = \frac{5}{9} = 0.6$$

$$\text{P}(\blacktriangle|\text{R}) = \frac{M[\blacktriangle, R]}{M[R]} = \frac{2}{4} = 0.5$$

$$\text{P}(\blacktriangle|\text{B}) = \frac{M[\blacktriangle, B]}{M[B]} = \frac{1}{5} = 0.2$$

$$\text{P}(\blacksquare|\text{R}) = \frac{M[\blacksquare, R]}{M[R]} = \frac{2}{4} = 0.5$$

$$\text{P}(\blacksquare|\text{B}) = \frac{M[\blacksquare, B]}{M[B]} = \frac{0}{5} = 0$$

$$\text{P}(\bullet|\text{R}) = \frac{M[\bullet, R]}{M[R]} = \frac{0}{4} = 0$$

$$\text{P}(\bullet|\text{B}) = \frac{M[\bullet, B]}{M[B]} = \frac{4}{5} = 0.8$$

Theoretical Overview of MLE

$$X \longrightarrow Y$$

$$L(\theta : \mathcal{D}) = \prod_{m=1}^{M} \Big(P(x[m] : \theta) P(y[m] \mid x[m] : \boldsymbol{\theta}) \Big)$$

Can be further simlified on next line

$$L(\theta : \mathcal{D}) = \left(\prod_{m=1}^{M} P(x[m] : \boldsymbol{\theta}_X) \right) \left(\prod_{m=1}^{M} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|X} \right)$$

$$\prod_{m:x[m]=x^0}^{M} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^0}) \prod_{m:x[m]=x^1}^{M} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^1})$$

This is called **decomposability**.

Local Likelihood Decomposition

Consider only:

$$\prod_{m:x[m]=x^0}^{M} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^0}) = \theta_{y^1|x^0}^{M[x^0,y^1]} \theta_{y^0|x^0}^{M[x^0,y^0]}$$

$$\theta_{y^1|x^0}^{M[x^0,y^1]} = \frac{M[x^0,y^1]}{M[x^0,y^1] + M[x^0,y^0]} = \frac{M[x^0,y^1]}{M[x^0]}$$

These M-terms are called **sufficient statistics**.

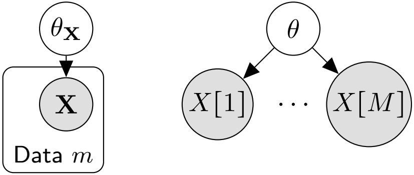
The **local likelihood** decomposes as:

$$L(\theta : \mathcal{D}) = \theta_{\mathbf{x^1}}^{\mathbf{M[x^1]}} \theta_{\mathbf{x^0}}^{\mathbf{M[x^0]}} \theta_{\mathbf{y^1|x^0}}^{\mathbf{M[x^0,y^1]}} \theta_{\mathbf{y^0|x^0}}^{\mathbf{M[x^0,y^0]}} \theta_{\mathbf{y^1|x^1}}^{\mathbf{M[x^1,y^1]}} \theta_{\mathbf{y^0|x^1}}^{\mathbf{M[x^1,y^0]}}$$

$$L_i(\boldsymbol{\theta}_{X_i|Pa_{X_i}} : \mathcal{D}) = \prod_m P(x_i[m] \mid Pa_{X_i}[m] : \boldsymbol{\theta}_{X_i|Pa_{X_i}})$$

Bayesian Estimation

- The biggest issue with MLE is its reliability in the parameter estimate. That is $\frac{1}{3} = \frac{1000000}{3000000}$.
- Instead we encode our knowledge (or lack of) as a **prior knowledge** about θ using a probability distribution.
- Here we assume that the outcome is **conditional independent given the parameter** θ.



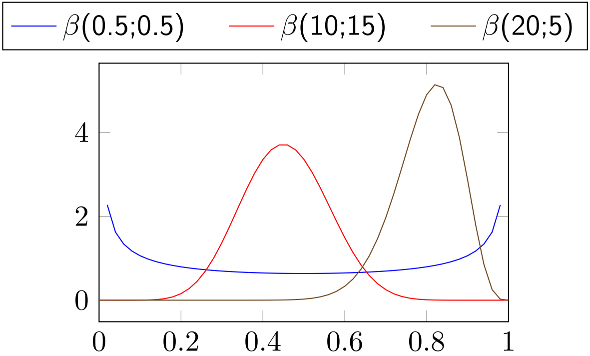
$$P(x[1], \ldots, x[M], \theta) = P(x[1], \ldots, x[M] | \theta) P(\theta)$$
$$= P(\theta) \prod_{m=1}^{M} P(x[m] \mid \theta)$$

Bayesian Estimation (BE)

$$P(\theta \mid x[1], \ldots, x[M]) = \frac{\overbrace{P(x[1], \ldots, x[M] \mid \theta)}^{\texttt{likelihood}} \overbrace{P(\theta)}^{\texttt{prior}}}{\underbrace{P(x[1], \ldots, x[M])}_{\texttt{constant}}}$$

- If we use a uniform prior then whats the difference between MLE and BE?
- If the prior is a Beta distribution, then the posterior distribution **is also a Beta distribution** (conjugate prior).
- As we obtain more data, the effect of the prior diminishes.
- Thus the **Bayesian framework** allows us to trade-off a diminishing prior as more data becomes available.

Choosing a Prior: Beta Distribution



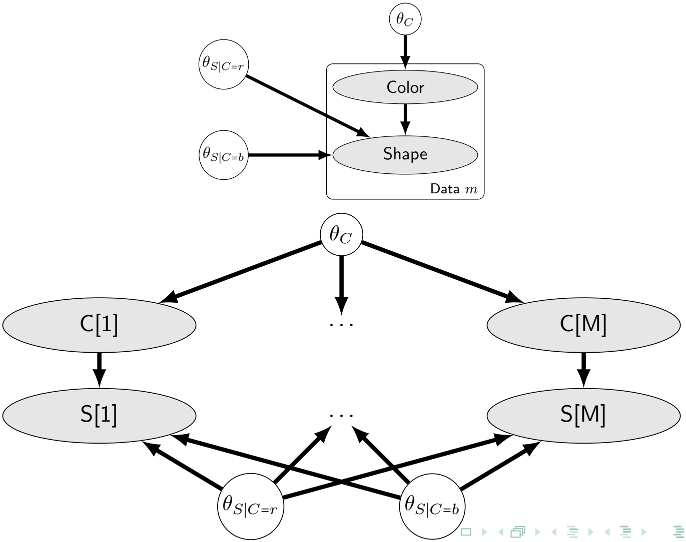
Mean $= \frac{\alpha}{\alpha+\beta}$**; Var:** $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$**; Skew Right:** $\alpha > \beta$

Skew Left: $\alpha < \beta$**; No Skew:** $\alpha = \beta$

Dirichlet Prior and Posterior

- The Dirichlet distribution is a **generalisation** of the Beta distribution.
- If the prior, $P(\theta)$, is Dirichlet then the posterior, $P(\theta, \mathcal{D})$, is Dirichlet.
- If $P(\theta)$ is $Dir(\alpha_1, \ldots, \alpha_K)$, then $P(\theta \mid \mathcal{D})$ is $Dir(\alpha_1 + M[1], \ldots, \alpha_K + M[K])$ where $M[k]$ is the number of occurances of x^k.
- This means that the posterior has a **compact description** and therefore makes clear computation and representation.

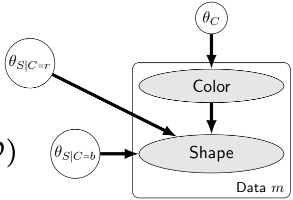
Local Decomposition of Bayesian Averages



Bayesian Prediction Averages

$$P(\theta_C, \theta_{S|C}) = P(\theta_C)P(\theta_{S|C})$$
$$P(\theta_{S|C}) = P(\theta_{S|C=r})P(\theta_{S|C=b})$$

$$P(\theta_{S|C}|\mathcal{D}) = P(\theta_{S|C=r}|\mathcal{D})P(\theta_{S|C=b}|\mathcal{D})$$



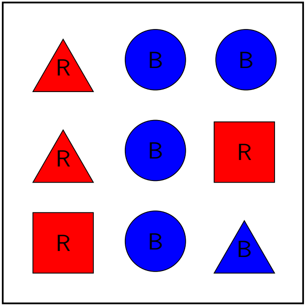
$P(\theta|\mathcal{D})$ decomposes nicely!

$$P(\theta|\mathcal{D}) = \prod_i \prod_{pa_{X_i}} P(\theta_{X_i|pa_{X_i}}|\mathcal{D})$$

If $P(\theta_{X|\mathbf{u}})$ is Dirichlet then:

$$P(X_i[M+1] = x_i \mid \mathbf{U[M+1]} = \mathbf{u},\ \mathcal{D}) = \frac{\alpha_{\mathbf{x_i}|\mathbf{u}} + \mathbf{M}[\mathbf{x_i}, \mathbf{u}]}{\sum_{\mathbf{i}} \alpha_{\mathbf{x_i}|\mathbf{u}} + \mathbf{M}[\mathbf{x_i}, \mathbf{u}]}$$

Bayesian Estimation: Example



Dirichlet Prior $\alpha = 2$

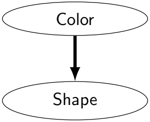
Color	
R	B
$\alpha/2$	$\alpha/2$

Color	
R	B
α_1	α_2

	Shape		
Color	T	S	C
R	$\alpha/6$	$\alpha/6$	$\alpha/6$
B	$\alpha/6$	$\alpha/6$	$\alpha/6$

	Shape		
Color	T	S	C
R	α_3	α_4	α_5
B	α_6	α_7	α_8

Color	
R	B
0.5	0.5



	Shape		
Color	T	S	C
R	0.5	0.5	0
B	0.2	0.1	0.7

$$P(R) = \frac{\alpha_1 + M[R]}{\alpha + M} = \frac{1+4}{2+9}$$

$$P(B) = \frac{\alpha_2 + M[B]}{\alpha + M} = \frac{1+5}{2+9}$$

$$P(\blacktriangle \mid R) = \frac{\alpha_3 + M[\blacktriangle, R]}{\alpha_{\text{red}} + M[R]} = \frac{\frac{2}{6}+2}{1+4}$$

$$P(\blacksquare \mid R) = \frac{\alpha_4 + M[\blacksquare, R]}{\alpha_{\text{red}} + M[R]} = \frac{\frac{2}{6}+2}{1+4}$$

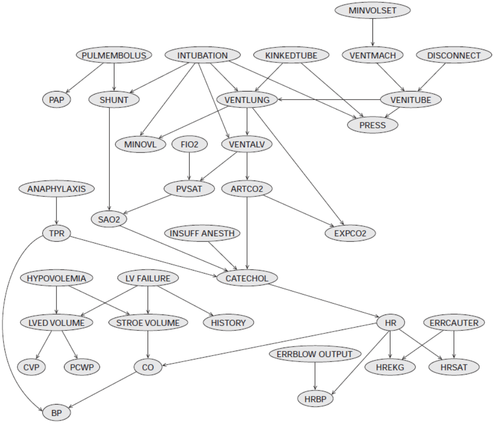
$$P(\bullet \mid R) = \frac{\alpha_5 + M[\bullet, R]}{\alpha_{\text{red}} + M[R]} = \frac{\frac{2}{6}+0}{1+4}$$

$$P(\blacktriangle \mid B) = \frac{\alpha_6 + M[\blacktriangle, B]}{\alpha_{\text{blue}} + M[B]} = \frac{\frac{2}{6}+1}{1+5}$$

$$P(\blacksquare \mid B) = \frac{\alpha_7 + M[\blacksquare, B]}{\alpha_{\text{blue}} + M[B]} = \frac{\frac{2}{6}+0}{1+5}$$

$$P(\bullet \mid B) = \frac{\alpha_8 + M[\bullet, B]}{\alpha_{\text{blue}} + M[B]} = \frac{\frac{2}{6}+4}{1+5}$$

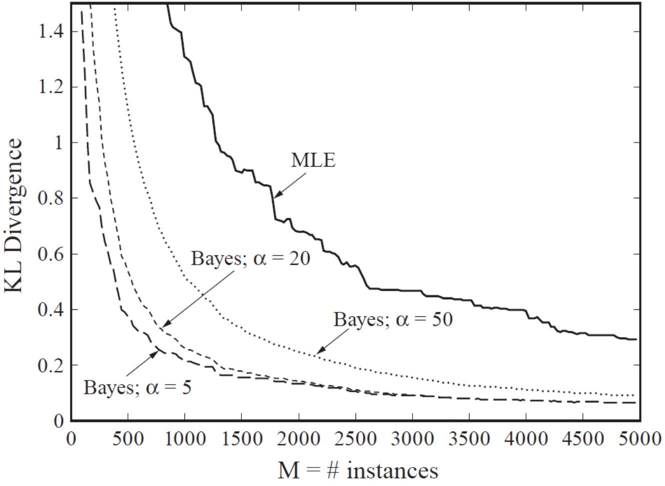
ICU Alarm Network



- **Pulmembolus** - bloodcloth in the lung
- **Shunt** - flap that allows bloodflow in the lung
- **Intubation** - Tube in throat to help breath
- **HypoVolemia** - body loses fluid

Credit: *Koller Textbook [2009], pp 750*

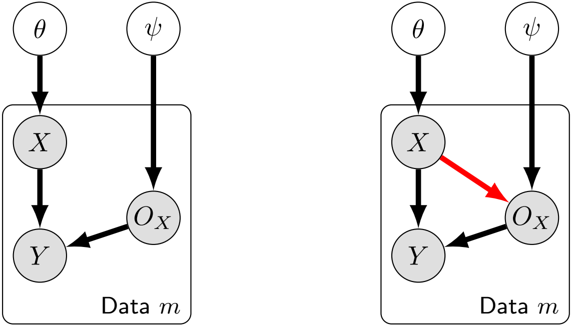
Values of Alpha



Credit: *Koller Textbook [2009], pp 751*

Mechanism Behind Incomplete Data

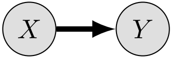
- Often we need to deal with **incomplete data**.
- This can occur mainly in three situations:
  1. Omitted fields in data collections process (e.g. blank field)
  2. Observations were not made (e.g. Medical tests)
  3. Some variables are hidden (e.g. quality of life)



(a) Randomly missing (b) Deliberately missing

Likelihood Function for Complete Data

- Recall the likelihood for complete data:

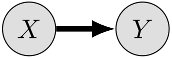


$$L(\theta : \mathcal{D}) = \theta_{\mathbf{x^1}}^{\mathbf{M[x^1]}} \theta_{\mathbf{x^0}}^{\mathbf{M[x^0]}} \theta_{\mathbf{y^1|x^0}}^{\mathbf{M[x^0,y^1]}} \theta_{\mathbf{y^0|x^0}}^{\mathbf{M[x^0,y^0]}} \theta_{\mathbf{y^1|x^1}}^{\mathbf{M[x^1,y^1]}} \theta_{\mathbf{y^0|x^1}}^{\mathbf{M[x^1,y^0]}}$$

- E.g. For samples: $\mathcal{D} = \{(x^0, y^0), (x^0, y^1), (x^1, y^0)\}$

$$\begin{aligned}
L(\mathcal{D} : \theta) &= P(x^0, y^0) P(x^0, y^1) P(x^1, y^0) \\
&= P(x^0) P(y^0|x^0) P(x^0) P(y^1|x^0) P(x^1) P(y^0|x^1) \\
&= \theta_{x^0} \cdot \theta_{y^0|x^0} \cdot \theta_{x^0} \cdot \theta_{y^1|x^0} \cdot \theta_{x^1} \cdot \theta_{y^0|x^1} \\
&= (\theta_{x^0} \cdot \theta_{x^0} \cdot \theta_{x^1}) \cdot (\theta_{y^0|x^0} \cdot \theta_{y^1|x^0} \cdot \theta_{y^0|x^1}) \\
&= (\theta_{x^0}^2 \cdot \theta_{x^1}) \cdot (\theta_{y^0|x^0} \cdot \theta_{y^1|x^0} \cdot \theta_{y^0|x^1})
\end{aligned}$$

Multimodal Likelihood Function



- Now suppose we have incomplete data:
- For samples: $\mathcal{D} = \{(?, y^0), (x^0, y^1), (?, y^0)\}$

$$
\begin{aligned}
L(\mathcal{D} : \theta) &= P(y^0)P(x^0, y^1)P(y^0) \\
&= \left(\sum_{x \in Val(X)} P(x, y^0) \right)^2 P(x^0)P(y^1|x^0) \\
&= \left(\theta_{x^0} \cdot \theta_{y^0|x^0} + \theta_{x^1} \cdot \theta_{y^0|x^1} \right)^2 \theta_{x^0} \cdot \theta_{y^1|x^0}
\end{aligned}
$$

- **NOT** unimodal
- **NOT** decomposed as product of likelihoods
- **NOT** in closed form (solved in a finite number of steps)
- **REQUIRES** probabilistic Inference (for sum-product)

Expectation Maximisation Algorithm

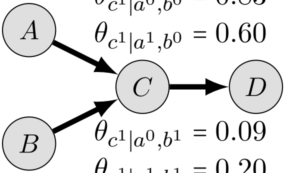
- Expectation Maximisation is a specialised approach to optimising likelihood functions.
- The approach as follows:
  1. "fill in" the missing values arbitrarily.
  2. Use the complete data learning procedure to estimate the parameters
  3. Then estimate the missing values with the new parameters
  4. Continue with steps 2 and 3 until convergence.

Intuition

- EM algorithm estimates expected sufficient statistics using completed data instances.
- It then finds the parameters that maximize the likelihood with respect to these statistics.

Understanding Expectation Maximisation

Suppose we had the following Bayesian network with parameters:



$\theta_{a^1} = 0.3$

$\theta_{b^1} = 0.9$

$\theta_{c^1|a^0,b^0} = 0.83$

$\theta_{c^1|a^1,b^0} = 0.60$

$\theta_{c^1|a^0,b^1} = 0.09$

$\theta_{c^1|a^1,b^1} = 0.20$

$\theta_{d^1|c^0} = 0.1$

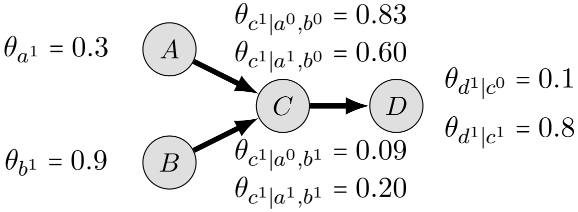
$\theta_{d^1|c^1} = 0.8$

- In the fully observable case MLE for $\hat{\theta}_{d^1|c^0}$ is:

$$\hat{\theta}_{d^1|c^0} = \frac{M[d^1, c^0]}{M[c^0]} = \frac{\sum_{m=1}^{M} \mathbb{1}\{\xi[m]\langle D, C\rangle = \langle d^1, c^0\rangle\}}{\sum_{m=1}^{M} \mathbb{1}\{\xi[m]\langle C\rangle = \langle c^0\rangle\}}$$

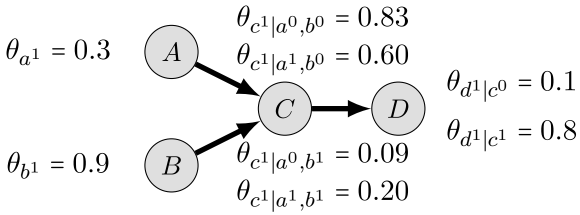
- In the incomplete data case we **cannot calculate** the value of the indicator function.

Understanding Expectation



- Suppose we have the following instance in the data: $\mathcal{D} = \{\langle a^1, ?, ?, d^0 \rangle\}$
- Then there are 4 possible completions of this data:
  1. $\langle a^1, \mathbf{b^0}, \mathbf{c^0}, d^0 \rangle$
  2. $\langle a^1, \mathbf{b^0}, \mathbf{c^1}, d^0 \rangle$
  3. $\langle a^1, \mathbf{b^1}, \mathbf{c^0}, d^0 \rangle$
  4. $\langle a^1, \mathbf{b^1}, \mathbf{c^1}, d^0 \rangle$

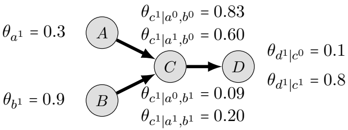
Understanding Expectation



- We can calculate the likelihood of each case given the parameters:
  1. $P(b^0, c^0 \mid a^1, d^0, \theta) = (0.3 \cdot 0.1 \cdot 0.4 \cdot 0.9)/P(a^1, d^0 \mid \theta)$
  2. $P(b^0, c^1 \mid a^1, d^0, \theta) = (0.3 \cdot 0.1 \cdot 0.6 \cdot 0.2)/P(a^1, d^0 \mid \theta)$
  3. $P(b^1, c^0 \mid a^1, d^0, \theta) = (0.3 \cdot 0.9 \cdot 0.8 \cdot 0.9)/P(a^1, d^0 \mid \theta)$
  4. $P(b^1, c^1 \mid a^1, d^0, \theta) = (0.3 \cdot 0.9 \cdot 0.2 \cdot 0.2)/P(a^1, d^0 \mid \theta)$

  Do you remember how to calculate $P(a^1, d^0 \mid \theta)$?

Calculating the Normalising Constant



$$P(a^1, d^0 \mid \theta) = P(a^1) \sum_{b \in Val(b^0, b^1)} P(b) \sum_{c \in Val(c^0, c^1)} P(c \mid a^1, b) P(d^0 \mid c)$$

$$= P(a^1) \sum_{b \in Val(b^0, b^1)} P(b) \Big(P(c^0 \mid a^1, b) P(d^0 \mid c^0) + P(c^1 \mid a^1, b) P(d^0 \mid c^1) \Big)$$

$$= P(a^1) \Big(P(b^0) \big(P(c^0 \mid a^1, b^0) P(d^0 \mid c^0) + P(c^1 \mid a^1, b^0) P(d^0 \mid c^1) \big)$$

$$+ P(b^1) \big(P(c^0 \mid a^1, b^1) P(d^0 \mid c^0) + P(c^1 \mid a^1, b^1) P(d^0 \mid c^1) \big) \Big)$$

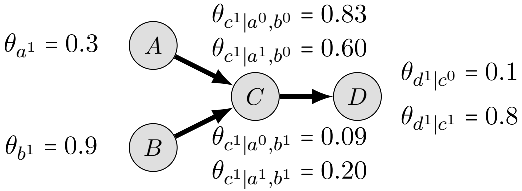
$$= 0.3 \Big(0.1 \big(0.4 \cdot 0.9 + 0.6 \cdot 0.2 \big) + 0.9 \big(0.8 \cdot 0.9 + 0.2 \cdot 0.2 \big) \Big)$$

$$= 0.3 \Big(0.1 \big(0.48 \big) + 0.9 \big(0.76 \big) \Big)$$

$$= 0.3 \big(0.732 \big)$$

$$= \boxed{0.2196}$$

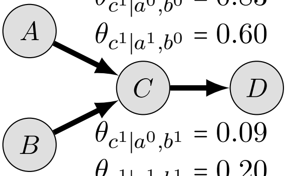
Understanding Expectation



- Calculate the Q function: expected value of the complete-data log-likelihood based on current estimates:
  1. $Q(\langle b^0, c^0 \rangle) = P(b^0, c^0 \mid a^1, d^0, \theta) = \frac{(0.3 \cdot 0.1 \cdot 0.4 \cdot 0.9)}{0.2196} = 0.0492$
  2. $Q(\langle b^0, c^1 \rangle) = P(b^0, c^1 \mid a^1, d^0, \theta) = \frac{(0.3 \cdot 0.1 \cdot 0.6 \cdot 0.2)}{0.2196} = 0.0164$
  3. $Q(\langle b^1, c^0 \rangle) = P(b^1, c^0 \mid a^1, d^0, \theta) = \frac{(0.3 \cdot 0.9 \cdot 0.8 \cdot 0.9)}{0.2196} = 0.8852$
  4. $Q(\langle b^1, c^1 \rangle) = P(b^1, c^1 \mid a^1, d^0, \theta) = \frac{(0.3 \cdot 0.9 \cdot 0.2 \cdot 0.2)}{0.2196} = 0.0492$
- Therefore the most likely assignment to $\mathcal{D} = \{\langle a^1, ?, ?, d^0 \rangle\}$ is $\mathcal{D} = \{\langle a^1, b^1, c^0, d^0 \rangle\}$

Understanding Expectation

$\theta_{a^1} = 0.3$ $\qquad$ $\theta_{b^1} = 0.9$



$\theta_{c^1|a^0,b^0} = 0.83$
$\theta_{c^1|a^1,b^0} = 0.60$
$\theta_{c^1|a^0,b^1} = 0.09$
$\theta_{c^1|a^1,b^1} = 0.20$

$\theta_{d^1|c^0} = 0.1$
$\theta_{d^1|c^1} = 0.8$

- Suppose we have another incomplete instance: $\mathcal{D} = \{\langle ?, b^1, ?, d^1 \rangle\}$, then:
  1. $Q'(\langle a^0, c^0 \rangle) = P(a^0, c^0 \mid b^1, d^1, \theta) = \frac{(0.7 \cdot 0.9 \cdot 0.91 \cdot 0.1)}{0.1675} = 0.342$
  2. $Q'(\langle a^0, c^1 \rangle) = P(a^0, c^1 \mid b^1, d^1, \theta) = \frac{(0.7 \cdot 0.9 \cdot 0.09 \cdot 0.8)}{0.1675} = 0.271$
  3. $Q'(\langle a^1, c^0 \rangle) = P(a^1, c^0 \mid b^1, d^1, \theta) = \frac{(0.3 \cdot 0.9 \cdot 0.8 \cdot 0.1)}{0.1675} = 0.129$
  4. $Q'(\langle a^1, c^1 \rangle) = P(a^1, c^1 \mid b^1, d^1, \theta) = \frac{(0.3 \cdot 0.9 \cdot 0.2 \cdot 0.8)}{0.1675} = 0.258$

Understanding Maximisation

This expectation step givers us an **augmented data set**, $\mathcal{D}^+$, with **likelihood weightings**. $\mathcal{D}^+$ consists of:

$$\cup_m = \{\langle \mathbf{o}[m], \mathbf{h}[m] \rangle \ : \mathbf{h}[m] \in Val(\mathbf{H}[m])\},$$

where each data case, $\langle \mathbf{o}[m], \mathbf{h}[m] \rangle$, has a weighting $Q(\mathbf{h}[m] \mid \mathbf{o}[m], \theta)$.

- Now we compute **expected sufficient statistics**:

$$\bar{M}_\theta[\mathbf{y}] = \sum_{m=1}^{M} \sum_{\mathbf{h}[m] \in Val(\mathbf{H}[m])} Q(\mathbf{h}[m]) \mathbb{1}\{\xi[m]\langle \mathbf{Y} \rangle = \mathbf{y}\}$$

Hence, we calculate:

$$\tilde{\theta}_{d^1|c^0} = \frac{\bar{M}_\theta[d^1, c^0]}{\bar{M}_\theta[c^0]}$$

For $\mathcal{D} = \{\langle a^1, ?, ?, d^0\rangle, \langle ?, b^1, ?, d^1\rangle\}$ we apply:

$$\bar{M}_\theta[\mathbf{y}] = \sum_{m=1}^{M} \sum_{\mathbf{h}[m] \in Val(\mathbf{H}[m])} Q(\mathbf{h}[m]) \mathbb{1}\{\xi[m]\langle \mathbf{Y}\rangle = \mathbf{y}\}$$

$$\begin{aligned}
\bar{M}_\theta[d^1, c^0] &= Q'(\langle a^0, c^0\rangle) + Q'(\langle a^1, c^0\rangle) \\
&= 0.342 + 0.129 = 0.471 \\
\bar{M}_\theta[c^0] &= Q(\langle b^0, c^0\rangle) + Q(\langle b^1, c^0\rangle) + Q'(\langle a^0, c^0\rangle) + Q'(\langle a^1, c^0\rangle) \\
&= 0.0492 + 0.8852 + 0.342 + 0.129 = 1.4054
\end{aligned}$$

$$\tilde{\theta}_{d^1|c^0} = \frac{\bar{M}_\theta[d^1, c^0]}{\bar{M}_\theta[c^0]} = \frac{0.471}{1.4054} = 0.335$$

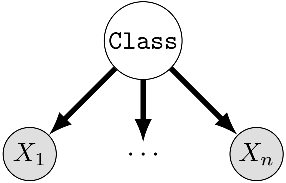
Expectation Maximisation Properties

The EM algorithm has some useful properties:

1. Each iteration is **guaranteed to improve** the log-likelihood function of the current set of the parameters to the data.
2. EM is **guaranteed to converge** to a local maximum, local minimum, or saddle point;
3. The convergence point is a fixed point of the likelihood function, which is essentially **always a local maximum**.

Bayesian Clustering

- Another application of EM is for Bayesian clustering
- This approach assumes the data is a mixture distribution and **uses the hidden variable** to separate its components.



$$\bar{M}_\theta[c] = \sum_{m=1}^{M} P(c \mid x_1[m], \ldots, x_n[m], \theta^t), \quad \theta_c^{t+1} = \frac{\bar{M}_\theta[c]}{M}$$

$$\bar{M}_\theta[x_i \mid c] = \sum_{m=1}^{M} P(c, x_i \mid x_1[m], \ldots, x_n[m], \theta^t), \quad \theta_{x_i \mid c}^{t+1} = \frac{\bar{M}_\theta[x_i, c]}{\bar{M}_\theta[c]}$$

K-Means Clustering

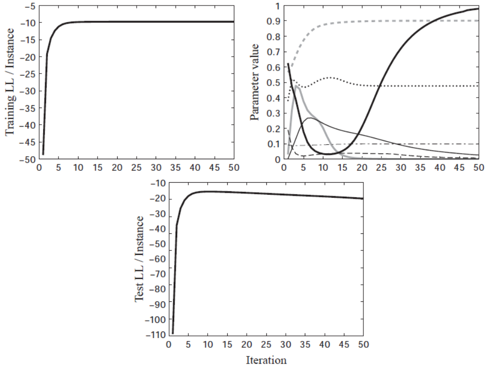
- An alternative to using a soft assignment is using a **hard assignment**.
- Given θ^t, we assign the following for each instance m:

$$c[m] = \underset{c}{\operatorname{argmax}} P(c \mid x[m], \theta^t)$$

- This results in $(\mathcal{D}^+)^t = \langle \mathcal{D}^+, \mathcal{H}^t \rangle$
- Thereafter, we compute **regular sufficient statistics** from $(\mathcal{D}^+)^t$ and computing the parameters.
- Hard EM assumes that data is generated from a **single Gaussian distribution**, which is not appropriate for complex settings,
- Each point will gravitate to the closest class, also called **K-means clustering**.

Convergence of Expectation Maximisation

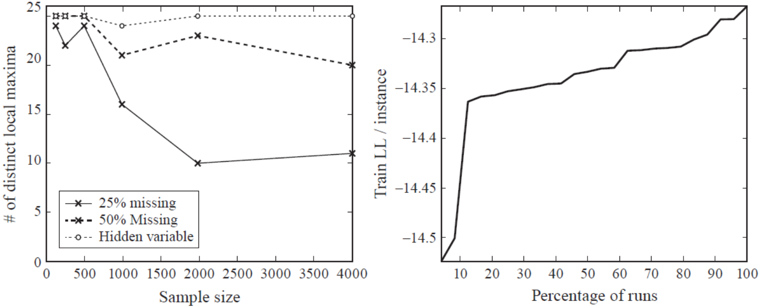
- EM maximises a **(bounded) log-likelihood function**, ensuring its guaranteed convergence.



Test set log-likelihood drops from overfitting, model complexity, or training-test data disparity.

Credit: *Koller Textbook [2009], pp 885*

Convergence of Expectation Maximisation



Credit: *Koller Textbook [2009], pp 886*

Summary

- We can **learn a model** for density estimation and knowledge discovery
- When learning we must clearly establish whether the learned model captured P^* using **experimental protocols**.
- Given issues with reliability of MLE, Bayesian estimation offers a much more useful **trade off** between evidence and priors
- Parameter estimation can be accomplished in both **complete and incomplete** data.
- EM is a powerful tool which has **practical properties**.
- However, the **convergence** of EM needs to be carefully assessed.