Lecture 7

Model Accuracy

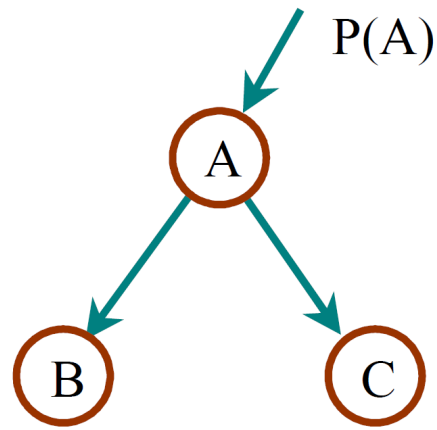
Joint Probability

We noted previously that:

The joint probability of any Bayesian network is the product of the conditional probabilities given the parents with the prior probabilities of the roots.

For a simple three node tree:

$$P(A\&B\&C) = P(A) \times P(B|A) \times P(C|A)$$



Given a data point (a_i, b_j, c_k) the joint probability $P(a_i, b_j, c_k)$ tells us how well that point fits the model.

Model Accuracy

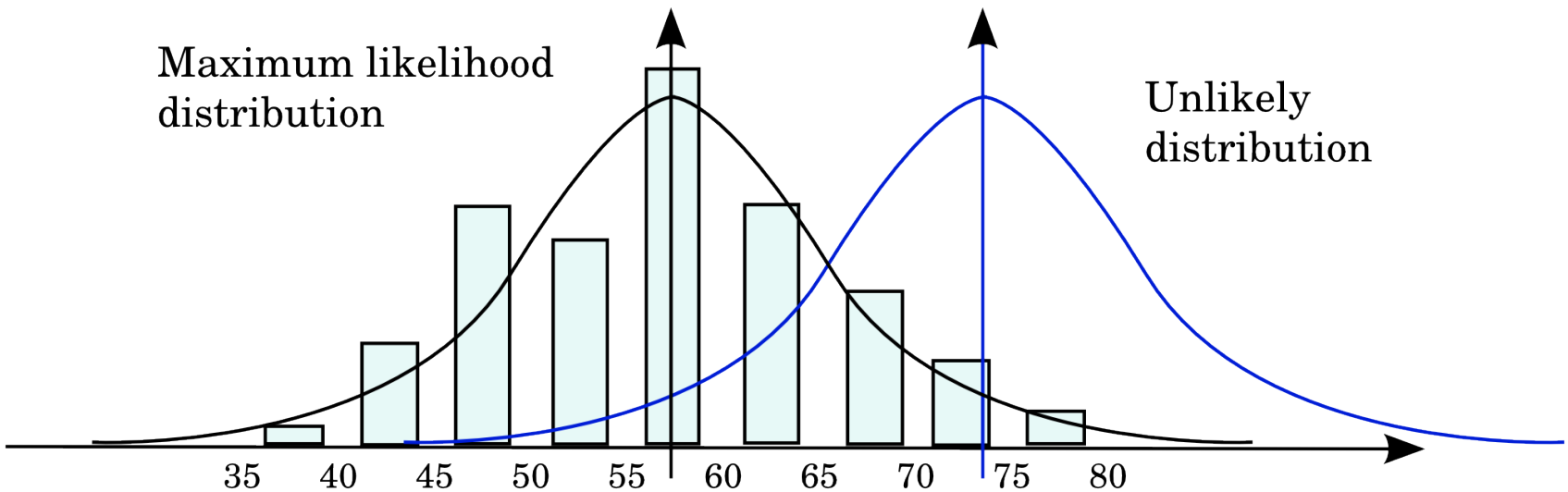
Suppose we have a data set Ds and a Bayesian network Bn, and we write the joint probability of the network as $P(Bn)$ then we can define the likelihood of the data set given the network as:

$$P(Ds|Bn) = \prod_{data} P(Bn)$$

The higher the value of $P(Ds|Bn)$ the closer the data distribution is to the joint probability distribution of the variables.

Maximum Liklihood

The concept of model accuracy is closely related to maximum liklihood.



We find the product of probabilities of each data point.

A Simple Example

Suppose we have the following data set and model:

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2

$$P(X) = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}$$

$$P(Y|X) = \begin{bmatrix} 1 & 1/3 \\ 0 & 2/3 \end{bmatrix}$$

$$P(Bn) = P(X\&Y) = P(X)P(Y|X)$$

$$P(Ds|Bn) = \prod_{data} P(Bn) = (1/4 \times 1) \times (3/4 \times 1/3) \times (3/4 \times 2/3) \times$$

$$(3/4 \times 2/3) = 1/64 = 0.0156$$

A Simple Example

Suppose we now choose a different model for the same data:

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2

$$P(X) = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}$$

$$P(Y) = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

$$P(Bn) = P(X\&Y) = P(X)P(Y)$$

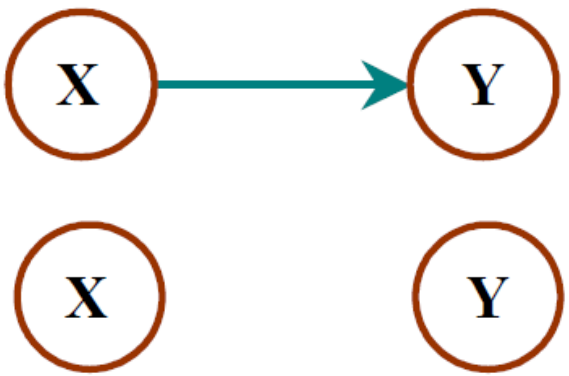
$$P(Ds|Bn) = \prod_{data} P(Bn) = (1/4 \times 1/2) \times (3/4 \times 1/2) \times (3/4 \times 1/2) \times$$

$$(3/4 \times 1/2) = 0.0066$$

Log Likelihood

The value of $P(Ds|Bn)$ goes down dramatically with the number of data points. Consequently it is more common to use the log likelihood of the data set given the model:

$$log_2(P(Ds|Bn))$$



$$P(Ds|Bn) = 0.0156$$
$$log_2 P(Ds|Bn) = -6$$

$$P(Ds|Bn) = 0.0066$$
$$log_2 P(Ds|Bn) = -7.24$$

Log Likelihood

Note that when using log liklihood the product becomes a sum i.e.:

$$P(Ds|Bn) = \prod_{data} P(Bn)$$

becomes:

$$log_2 P(Ds|Bn) = log_2 (\prod_{data} P(Bn))$$

and to avoid underflow we take the log of each joint probability and add them up:

$$log_2 P(Ds|Bn) = \sum_{data} log_2 P(Bn)$$

Minimum Description Length

- It is not surprising that the first model represents the data better. It uses six conditional probabilities rather than four.
- So although $log_2 P(Ds|Bn)$ gives us a measure of how well a network represents a data set it is not sufficient to compare competing models since it will always be maximised by adding more arcs.
- Instead we invoke the minimum description length principle which is to choose the smallest model that represents the data accurately.
- Thus we define an MDLScore which weighs size and accuracy.

Model Size

- A simple measure of model size is to count the number of parameters required to represent the conditional and prior probabilities.
- Each prior probability is a vector of m probability values. However since it is a probability distribution and sums to 1 it can be represented by $m-1$ parameters.
- Each link matrix has $n \times m$ probability values, but each column is a probability distribution, hence it can be represented by $(n-1) \times m$ parameters.

Representing the parameters

- We know that with M bits we can represent integers between 0 and $2^M - 1$.
- Conversely integers up to R can be represented by $\lceil log_2 R \rceil$ bits. We can also calculate the average number of bits required to represent integers up to R which is $(log_2 R)/2$.
- Suppose that we have N data points, and instead of representing our parameters as probabilities we represent them by frequency counts, then we will need on average $(log_2 N)/2$ bits to represent each parameter.
- This is the normal measure used to define the size of the model parameters.

Model Size

Suppose that the model has been defined using N points and that we require $|Bn|$ parameters to be able to construct the probabilities in the prior probability vectors and the conditional probability matrices. Then the model size can be defined as:

$$Size(Bn|Ds) = |Bn|(log_2 N)/2$$

Model Size Example

In our simple example above we used the following data set and model:

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2

$$P(X) = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}$$

$$P(Y|X) = \begin{bmatrix} 1 & 1/3 \\ 0 & 2/3 \end{bmatrix}$$

$P(X)$ can be represented by one parameter.
$P(Y|X)$ can be represented by two parameters.
Hence $|Bn| = 3$ and the number of data points $N = 4$
Thus:

$$Size(Bn|Ds) = |Bn|(log_2 4)/2 = 3 \times 2/2 = 3$$

Problem

What is the model size for the independent model?

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2

$$P(X) = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}$$

$$P(Y) = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

Problem

What is the model size for the independent model?

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2

$$P(X) = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}$$

$$P(Y) = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

$P(X)$ can be represented by one parameter.
$P(Y)$ can be represented by one parameter.
Hence $|Bn| = 2$ and the number of data points $N = 4$
Thus:

$$Size(Bn|Ds) = |Bn|(log_2 4)/2 = 2 \times 2/2 = 2$$

Overall Model Score

Since we wish to find the smallest model that will represent the data accurately we define our measure as:

$$MDLScore(Bn\&Ds) = |Bn|(log_2 N)/2 - log_2 P(Ds|Bn)$$

We saw in the examples above that the lower the network liklihood the larger the negative log likelihood becomes. Hence the network with the lowest MDL score is the best.

This method of scoring models by subtracting the liklihood of the model from its size was originally developed by Gideon Schwarz in 1978 and is called the Bayesian Information Criterion (BIC).

There is another similar score called the Akaike Information Criterion (AIC) which we will not cover.

A (slightly) bigger example

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2
x_2	y_2
x_1	y_1
x_1	y_1
x_2	y_2

$$P(X) = \begin{bmatrix} 3/8 & 5/8 \end{bmatrix}$$

$$P(Y|X) = \begin{bmatrix} 1 & 1/5 \\ 0 & 4/5 \end{bmatrix}$$

$$\begin{aligned} log_2 P(Ds|Bn) &= 3 \times log_2(3/8) + log_2((5/8) \times (1/5)) + \\ &\quad 4 \times log_2((5/8) \times (4/5)) \\ &= 3 \times log_2 3 - 9 - 3 - 4 = -11.25 \end{aligned}$$

A (slightly) bigger example

Choosing an independent model:

X	Y
x_1	y_1
x_2	y_1
x_2	y_2
x_2	y_2
x_2	y_2
x_1	y_1
x_1	y_1
x_2	y_2

$$P(X) = \begin{bmatrix} 3/8 & 5/8 \end{bmatrix}$$

$$P(Y) = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

$$\begin{aligned} log_2 P(Ds|Bn) &= 3 \times log_2(3/16) + log_2(5/16) + 4 \times log_2(5/16) \\ &= 3 \times log_2 3 - 12 + log_2 5 - 4 + 4 \times log_2 5 - 16 \\ &= -15.6 \end{aligned}$$

A (slightly) bigger example

The model size has gone up reflecting the fact that the data set is twice as big:

$$\begin{aligned} \text{Connected:} \quad & |Bn|(log_2 N)/2 = 3 \times 3/2 = 4.5 \\ \text{Independent:} \quad & |Bn|(log_2 N)/2 = 2 \times 3/2 = 3 \end{aligned}$$

Which gives the following MDL (or BIC) scores:

$$\begin{aligned} \text{Connected:} \quad & MDLScore(Bn\&Ds) = 4.5 + 11.25 = 15.75 \\ \text{Independent:} \quad & MDLScore(Bn\&Ds) = 3 + 15.6 = 18.6 \end{aligned}$$

The data set shows a strong correlation between X and Y and the MDL metric indicates that the connected network, though bigger is preferable.

Searching for the best network

- Our MDLScore opens up the possibility of choosing between competing networks.
- We could devise heuristic tree search algorithms to find the best network. For example with four variables A, B, C and D:
  - There is independent network (root of the search tree)
  - There are C_2^4 networks with one arc: $A-B$, $A-C$, $A-D$, $B-C$, $B-D$ and $C-D$
  - It looks as if we might be up against a combinatorial explosion!
- How many networks are there in total?

Size of the search tree

- We can estimate the size of the search tree by noting that n variables can be connected by a possible:

$$(n-1)+(n-2)+...2+1 = n \times (n-1)/2 \text{ arcs.}$$

- Since any possible arc can be present or not, there are a total of $2^{(n(n-1)/2)}$ possible networks.
- Unfortunately this is rising exponentially and for 8 variables there are 268,435,456 nodes in the search tree
- So exhaustive search is not possible for non trivial networks, and even singly connected networks soon become infeasable.

Heuristic Search

One possibility is to prune the tree during the depth first search using some heuristic rules. Possibilities could include:

1. If a network has a higher MDLScore that its parent remove it from the search tree.
2. Add arcs only if the mutual information between the variables is high
3. At each level expand only networks with low MDL scores

These strategies can dramatically reduce the search time, but they all run the risk of finding a sub-optimal solution. Given that the spanning tree algorithm gives us an aproximate model, tree search does not seem to add much value.

The MDL Score is not an absolute measure

$$MDLScore(Bn\&Ds) = |Bn|(log_2 N)/2 - log_2 P(Ds|Bn)$$

Suppose we have a data set Ds with N items, and we test a Bayesian Network Bn against it. Let $log_2 P(Ds|Bn) = s$.

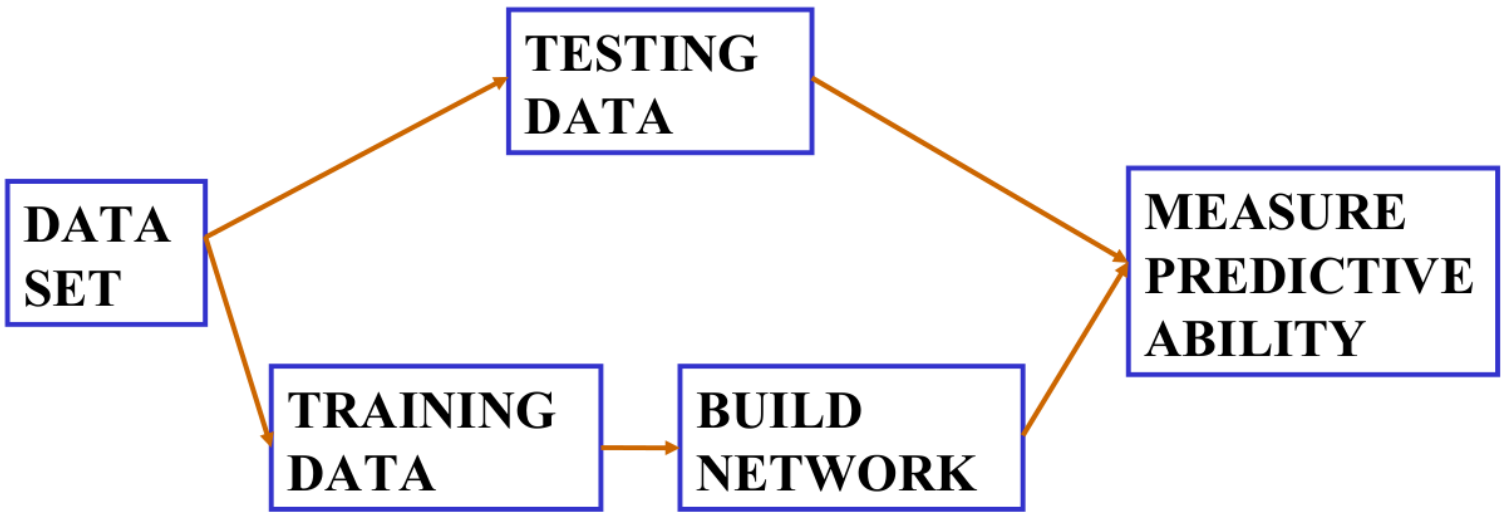
Now create a new data set Ds' by duplicating every data point in Ds. Ds and Ds' will have identical probability distributions, but $log_2 P(Ds|Bn) = 2s$

Since N has been doubled $log_2 N$ increases by 1 and the Size measure increases by the number of parameters.

Overall the MDLScore increases significantly, though the model and data distributions have not changed.

Measuring Predictive Accuracy

An alternative approach to model accuracy is to measure a network's predictive ability.



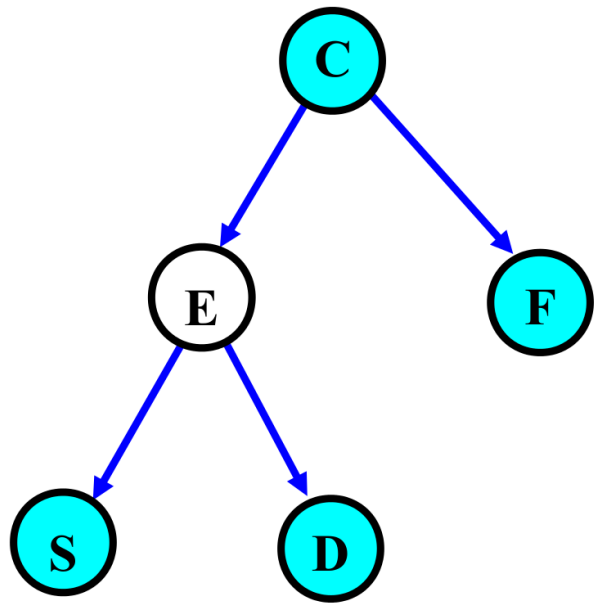
Measuring Predictive Accuracy

Choose a target variable eg E.

For each point in the testing data set:

- Instantiate C, S, D and F
- Calculate $P'(E)$
- Calculate $|P'(E) - D(E)|$ (where $D(E)$ is the data value of E

The smaller the average $|P'(E) - D(E)|$ the more accurate the network



Small is beautiful

- The joint probability of the variables in a Bayesian Network is simply the product of the conditional probabilities and the priors of the root(s).
- If the network is an exact model of the data then it must represent the dependency exactly.
- However, using a spanning tree algorithm this may not be the case.
- In particular, we may not be able to insert an arc between two nodes with some dependency because it would form a loop.
- The effect of unaccounted dependencies is likely to be more pronounced as the number of variables in the network increases.

Minimal Spanning Tree Approach

The predictive aproach can be used to minimise the number of variables in the spanning tree. (Enrique Sucar)

1. Build a spanning tree and obtain an ordering of the nodes starting at the root.
2. Remove all arcs
3. Add arcs in the order of the magnitude of their dependency
4. If the predictive ability of the network is good enough (or the nodes are all joined) stop:
   otherwise go to step 3