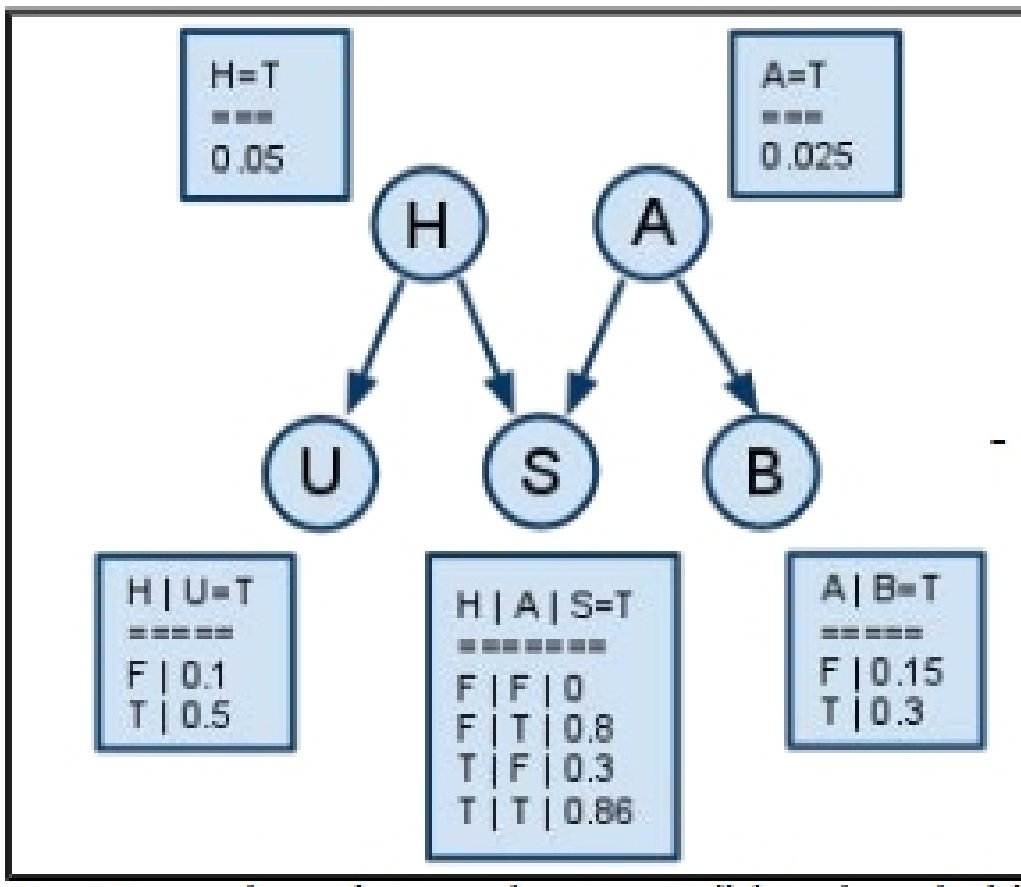
Probability, Bayes Nets, Naive Bayes, Model Selection

Major Ideas:

1. Intro to Bayes nets: what they are and what they represent.
2. How to compute the **joint probability** from the Bayes net.
3. How to compute the **conditional probability** of any set of variables in the net.
    - Marginalization and Exact Inference
    - Bayes Rule (backward inference)
4. Naive Bayes - classification using Bayes Nets
5. Bayesian Model Selection / Structure Search
6. Generative versus Discriminative Models
7. (Optional) D-Separation Rules for determining conditional independence in Bayes Nets
8. (Optional) Noisy OR

Bayes Nets are a compact way to represent the Joint Distribution of a set of Random Variables. The nodes represent Random Variables. Random variables are variables that provides a mapping from values to probabilities.

We have the following Bayes net from recitation. There are five random variables (to simplify we've removed the MIT variable)



H = {T, F} There is a **H**ack on Campus
A = {T, F} There is an **A**rt exhibit on Campus
U = {T, F} We see a **U**niformed police man.
S = {T, F} We see a **S**culpture
B = {T, F} We see **B**usiness People in suits on Campus

Next to each node, you have conditional probability tables (**CPT**), these represent the conditional probability of the underlying Random Variable conditioned on its parents.

The CPT for nodes H and A have only one value because (not H) or (not A) is simply one minus the value shown. Variables U, S, B have CPTs that are dependent on their parent variables. Generally, in a CPT, we use the first n-1 columns to denote the settings of the "given" Variables. If the variable is binary, the last column is the True value of probability of the variable for the settings of the given variables. If a variable is multivalued, then When dealing with binary variables we don't show the implicit 1-p column.

For Example S's CPT fully expanded is:

H	**A**	**S**=T	S=F (not shown)
F	F	0	1
F	T	0.8	0.2

T	F	0.3	0.7
T	T	0.86	**0.14**

So using the CPT above. What is the probability of *NOT seeing a sculpture* if we know there is a Hack and there is an Art exhibit? Or what is P(S=F | H = T, A = T) Answer: 0.14

A Matter of Parameters

Sometimes questions may ask "*how many parameters*" are in a given Bayes Net? By number of parameters we really mean the number of CPT entries. This is because the Bayes Net is fully specified only when all the parameters are assigned some numerical value. Note that the hidden columns like P(S=F|H,A) does not count in that number, because those entries can be gotten by 1-P (S=T|H,A).

So how many parameters does our "Hack or Art show" network have?
#parameters = #cpt entries in network = 1 + 1 + 2 + 4 + 2 = 10

Joint probability or computing the probability of a specific world state

Suppose we know there is a hack on campus, and there isn't an art exhibit, we see a uniformed officer, a sculpture, and we don't see any Business men.
The probability of such a world is: P(H=T, A=F, U=T, S=T, B=F)

We can easily compute the Joint probability from a Bayes net!
For any Bayes Net:

$$p(V_1 \ldots V_n) = \prod_{i=1}^{n} p(V_i | Parents(V_i))$$

In other words, Bayes Nets is really an encoding of the conditional dependencies of a set of random variables. All variables are independent of other variable given their parents.
For our example:

$$p(H,A,U,S,B) = p(H)p(A)p(U|H)p(S|H,A)p(B|A)$$

So for the specific setting above:

P(H=T, A=F, U=T, S=T, B=F) = P(H=T)P(A=F)P(U=T|H=T)P(S=T|H=T, A=F)P(B=F|A=F)
= 0.05 x (1-0.025) x 0.5 x 0.3 x (1-0.15) = 0.006215625

Marginalization over the Joint

Example: How to compute the probability of P(S)?
We can compute any arbitrary probabilities from joint probabilities by the method of "marginalization" = summing out variables that we don't want.

$$P(S) = \sum_{H,A,U,B} P(H,A,U,S,B)$$
$$P(S) = \sum_{H,A,U,B} P(H)P(A)P(U|H)P(S|H,A)P(B|A)$$
$$P(S) = \sum_{H}\sum_{A}\sum_{U}\sum_{B} P(H)P(A)P(U|H)P(S|H,A)P(B|A)$$

Next move the sums so that a sum is placed only before all the terms that depend on it.
e.g. P(S|H,A) depends on sum A and sum H so those sums occur left of it:

$$P(S) = \sum_{A} P(A) \sum_{H} P(H)P(S|H,A) \sum_{U} P(U|H) \sum_{B} P(B|A)$$

Notice that: $\sum_{B} P(B|A) = 1$, and same for $\sum_{U} P(U|H) = 1$!

So dropping the B and U terms we get the final summation:

$$P(S)=\sum_A P(A)\sum_H P(H)P(S|H,A)$$

Which works out to:

$$P(S)=P(H)P(A)P(S|H,A)+P(\overline{H})P(A)P(S|\overline{H},A)+P(H)P(\overline{A})P(S|H,\overline{A})$$
$$+P(\overline{H})P(\overline{A})P(S|\overline{H},\overline{A})$$

= 0.0347

Ancestor (Sub)Graph: a subgraph of the Bayes Net where only variables of interest and their ancestors are drawn

Ancestor Graph Shortcut:
Any variable that is not in the "ancestor" graph for the set of variables of interest we can remove from the summation.

Example:
In computing **P(S)**, **S** is the variable of interest, **H, A** are both ancestors of S, but not **U** or **B**. So **P(U|H)**, **P(B|A)** can be safely dropped.
In computing **P(S | B)**, **S, B** are variables of interest, **H, A** are ancestors of S, and B. So **P(U|H)** can be dropped from the summation.

General Method for computing any P(X) from a Bayes Net

1. Write P(X) as a marginalization sum over joint probabilities.
2. Cross out/ignore terms that are not in the ancestor graph of X.
3. Expand the summation and simplify

Example:

$$P(B)=\sum_{A,H,S,U}P(H,A,U,S,B)$$ Ancestor graph of B only include variables B and A.

$$P(B)=\sum_{A,H,S,U}\cancel{P(H)}\cancel{P(A)}\cancel{P(U|H)}P(B|A)P(A)$$

$$P(B)=P(B|A)P(A)+P(B|\overline{A})P(\overline{A})$$

General Method for computing any P(X|Y) from a Bayes Net:

1. Write P(X, Y) as a marginalization sum over joint probabilities (without summing over X, Y).
2. Cross out/ignore terms that are not in the ancestor graph of X and Y.
3. Write P(Y) as a marginalization sum over joint probabilities.
4. Cross out/ignore terms that are not in the ancestor graph of Y.
5. Perform division: P(X,Y)/P(Y), and simplify.

Example: What is P(S|B)?

$$P(S|B)=\frac{P(S,B)}{P(B)}$$ From the definition of conditional Probability

Both of these terms can be computed via marginalization over the Joint.

$$P(S,B)=\sum_{A\in T,F}\sum_{H\in T,F}P(A)P(H)P(S|H,A)P(B|A)$$

$$P(B)=\sum_{A\in T,F}P(B|A)P(A)$$

Both terms above have been simplified using the ancestor graph trick.