Bayesian Networks (Bayes network, belief network, probabilistic directed acyclic graphical model)

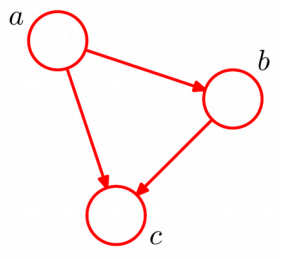
Graphical models are useful in visualizing probabilistic models. A graph consists of *nodes* which are connected by *links*. Each node represents a random variable (or a group of variables) and the links express probabilistic relationships between the variables. The graph then captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors each depending only on a subset of the variables.

Here, we are interested in Bayesian networks (bayesian belief nets), also known as directed acyclic graphs, in which the links of the graphs have a particular directionality indicated by arrows. The acyclic means that there must be no directed cycles.

Let's look at a joint distribution p(a,b,c) over three variables a, b, and c. By applying Bayes rules we get

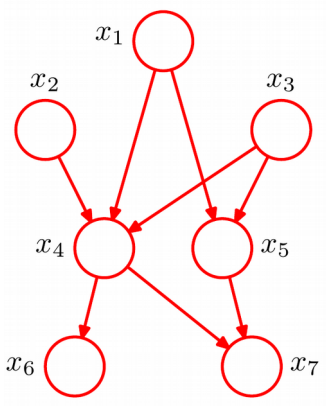
$$p(a,b,c) = p(c|a,b)p(a,b) = p(c|a,b)p(b|a)p(a)$$

Note that this decomposition holds for any choice of the joint distribution. Let's look at the right-hand side of this as a simple graphical model by first introducing a node for each variable and then drawing arrows according to the conditional probabilities. e.g. *c* depends on *a* and *b* so we draw an arrow from both *a* and *b* towards *c*.



If there is a link going from a node *a* to a node *b*, then we say that node a is the *parent* of node *b*, and we say that node *b* is the *child* of node *a*. Note that left-hand side of the equation is symmetrical with respect to the three variables *a*, *b*, and *c*, whereas the right-hand side is not. We have implicitly chosen a particular ordering, namely *a*, *b*, *c*, and had we chosen a different ordering we would have obtained a different decomposition and hence a different graphical representation.

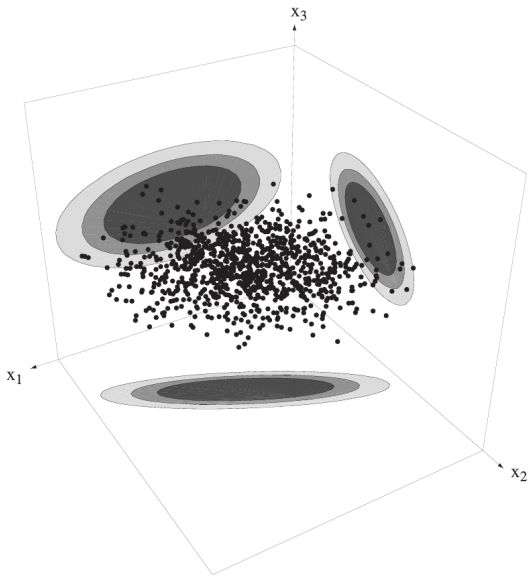
A bit more tricky example could look like



$p(x_1)p(x_2)p(x_3)p(x_4 \mid x_1, x_2, x_3)p(x_5 \mid x_1, x_3)p(x_6 \mid x_4)p(x_7 \mid x_4, x_5)$

Can you see how the corresponding equation is built?

If for two random variables a and b we have p(a,b) = p(a)p(b) then we say that the variables are *statistically independent*. In the image below x_1 and x_3 are independent. You can see it by observing the distributions and that the distributions are slanted except between x_1 and x_4. (Assuming that both variables are normally distributed.) The major axes are parallel to major ellipsoid axes. (The independence cannot be always seen from the distributions as sometimes there are non-linearities or complex distributions that make the observation difficult.)



The problem of Bayesian networks is that if we don't know all the relations between variables (or the amount of variables is large), building the network becomes difficult and time consuming. To learn the relations, there's a need for more and more data → curse or dimensionality. In addition, Bayesian networks cannot handle cyclic relationships. However, Bayesian networks are useful for solving problems when data is scarce as a way to prevent overfitting.

For our exercise, we need one more concept: marginalization (or summing out, "sum rule"). We can recover the probability distribution of any single variable from a joint distribution by summing (discrete case) or integrating (continuous case) over all the other variables.

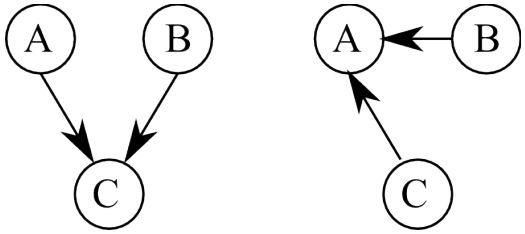
The Rules of Probability

sum rule $$p(X) = \sum_{Y} p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X).$$

(From Bishop: Pattern recognition and and machine learning. The book is NOT required for the course but a good reference if, after the course, you want to know more.)

On this lesson, you are given some data and following graphs. We should determine which one is correct.



We can do that by comparing a joint probability distribution of our data and what Bayes formulas would predict the distribution to be.

Let's first look at the case in which C depends on A and B. Because A and B are independent, we must calculate p(A,B,C) = p(C | A,B) p(A,B) = p(C | A,B) p(A) p(B).

From the data we can calculate joint probability p_c(A,B,C). Then we can use the sum rule to sum over different variables.

$$p_c(A,B) = \sum_C p_c(A,B,C)$$

$$p_c(A) = \sum_B p_c(A,B)$$

$$p_c(B) = \sum_A p_c(A,B)$$

We still don't know the p(C | A,B) but we get it from our data p(C | A,B) = p_c(a,b,c) / p_c(a,b).

Thus we need to calculate p_{bayes}(A,B,C) = (p_c(a,b,c) / p_c(a,b)) p_c(A) p_c(B). Now we can compare p_c(A,B,C) – p_{bayes}(A,B,C) and see if values are close to zero.

In the second graph, A depends on B and C, and B and C are independent. Thus p(A,B,C) = p(A | B,C) p(B,C) = p(A | B,C) p(B) p(C).

By comparing the results we see that the last graph is obviously wrong in addition to the fact that it doesn't seem to predict all the outcomes.

What you should learn from here: How to use Bayes networks and that with Bayes networks, you can do calculations really fast and relatively easy. However, you need to know something about your data before hand. If you know nothing, Bayes calculations cannot help you and finding the correct Bayes network for large networks can be nearly impossible.