Machine Learning,ML

Prof. Carl Gustaf Jansson
Prof. Henrik Boström
Prof. Fredrik Kilander
Department of Computer Science and Engineering
KTH Royal Institute of Technology, Sweden

Lecture 13
Bayes (ian) Belief Networks

So welcome to the third lecture of this third week the course of machine learning. The theme of this lecture will be Bayes or Bayesian belief networks. So starting with the general characteristics of a representation, a Bayesian belief network abbreviated as BBN has a number of synonyms you can call Bayes network or Base Model network, or Belief Network or you can call it a Decision network or some lengthy elaboration like probabilistic directed acyclic graphical model. So essentially BBN is a probabilistic graphical model representing the set of variables and their conditional dependencies and as a structure it is a directed acyclic graph or that DAG. So a BBN enables us to model and reason about uncertainty. BBN's accommodate both subjective probabilities and probabilities based on objective data, but we in both cases we talk more on assertive probabilities rather than probabilities based on frequencies. The most important use of BBN's is in revising probabilities in the light of actual observations of events. So if you look at this little example to the right where you have three variables, rain it's raining or not, there is a sprinkler on or not, the grass is wet or the grass is not wet. And so one can say there is a this kind of graph represents some causal dependency, so the rain can make the grass wet directly, the absence of rain can affect a person to put on the sprinkler, and when the sprinkler is on the grass gets wet. Yeah also so in a way you can look at consider this kind of network in a forward manner starting from the course and looking at the consequences, but you can also do it backwards so you can see "Oh is the grass wet" but is it probably for the grass it is actually wet and then when you get data on oh but the grass is wet, you can infer probabilities of the likelihood of certain courses whether it's more likely that the sprinkler was on or that it was really raining. So the backward reasoning here is more of the end purpose of this kind of representation. So one key theorem from probability theory is called Bayes theorem this theorem as a fundamental importance for the way we want to use Bayesian networks for problem solving actually Bayes theorem talks about the situation where we have two variables A and B where B is

dependent on A or we can say A causes B. So about three kinds of probabilities that can be considered here it's the independent probability of A we call it P(A) and similar for B we have the prior probability of B, P(B) which means considering the probability for B in isolation. Then we can talk about the conditional probability so we can see initially can talk about the conditional probability for B given A and the and vice versa. So the condition for A given B. Finally we can talk about the joint probability for both A and B considered together and the joint probability for A and B can be proven to be computed as the multiplication of P(B|A) and the conditional probability of B given A.

$$P(A|B) = P(B|A) * P(A) / P(B)$$

And what is more interesting is how we can reason in this kind of Network backward so it's the case that for many domains it's more likely that we know the data to support the conditional probability for B given A which means is the for in the forward Direction, for reasoning from cause to effect.

So therefore it would be very practical to have a theorem that can infer the opposite probability the reasoning from effect to cause instead of directly observing it. So I am happy that the Bayes theorem does exactly that same, what the Bayes theorem says that we can infer the conditional probability of A the cause given B the effect to be exactly the conditional probability of B given A time's the prior probability of A divided by the prior probability of B.

So the intuitive meaning of course then is that Bayes Theorem really defines how one can infer a conditional probability for a course giving us a probability of a symptom, what you can see to the far right here is just a graphical intuitive way of defining that kind of proof that can be given for this theorem.

So let's now look at the core components of this representation and so Nodes represent variables in the Bayesian sense as described earlier can be observable quantities, Hidden variables or hypotheses edges represent conditional dependencies. So each node is associated with the probability function that takes as input a particular set of probabilities for values for the nodes parent variables and outputs the probability of the values of the variable represented by a node. So a prior probability if we look at the example to the right, is the same we looked at earlier we have rain we have the sprinkler on we have grasswet. So the probably of rain is given then by the two probabilities for the true feature values so if rain is a

feature it can have the value true we the 0.2 probability and it could have a value false with its 0.8 probability. So then if we look to a conditional probability in this case we look at the probability of the sprinkler is on given that it's rain, so we can see here that we have two cases so either it doesn't rain, the rain is false and then it's a 0.4 probability to the sprinklers on and it's a 0.6 probability that that was sprinkler is not on. In the opposite case when it when it when it's really raining it's a much lower probability that it's the sprinklers on so it's very rarely do you put on the sprinkler so in 0.1 case you have a sprinkler on. So to these kinds of small tables are really the key information connected with each node in this kind of belief network either prior or conditional and you see depending on how many connections there are how many input edges there are 2 node you get a larger a table because typically you describe the probability function here in terms of a table, of course also it depends on how many feature values there so if there are it happen to be more than true and false while he was an ordinal feature then also the table grows in in size. You also see on this line that the joint probability function there is a way to calculate that from the prior probabilities and the conditional probabilities and one can do that in a format manner so as been remarked earlier and it's more likely in the main that we information about all these probabilities in the forward Direction starting from causes and moving towards effects.

So let's then look at a minute at related examples so you recognize three of the variables, rain sprinkler and wet grass what we did now was that we included a fourth variable called cloudy and changed the structure so that cloudy effects sprinkler and cloudy effects rain but as you see the connection between rain and sprinkler disappeared so this kind of alternative Network is based for discussing another phenomena which we call conditional independence and I will talk about that in the next slide.

Conditional independence means that nodes that are not connected by any path represent variables that are conditionally dependent on each other so as you see in in there in the alternative example where we changed the structure, sprinkler and rain is no longer related because there is no path between them so therefore sprinkler and rain are considered as conditionally independent and it's kind of obvious that the conditional dependency of sprinkler given cloudy is one thing and doesn't change anything to include rain, so the condition sprinkler given cloudy and rain is to say as the first. So in this case the probabilities for the network can be calculated as below on this slide and as you can see the nice thing here is that we can calculate the joint probability for all variables just by following the structure of the network in a formal manner, so we essentially the joint probability is this product of the

prior probability for the independent variable times the conditional probabilities for the two middle variables that all depend on cloudy and finally on the conditional probability of wet grass given the middle two. So this is a nice property of this kind of networks that we can do this kind of forward reasoning just following the structure of the network and using the probability table probability function in each node. So the examples we've shown earlier here are of course very small so to exemplify the principles however in realistic cases and here on this slide I've included a more realistic one these kind of networks are large and can be very huge however and due to the properties of the Bayesian networks that will try to illustrate in the last few slides it should be scalable and manageable, but the example here is more the one you would face in reality. Yeah so now I want to talk about a few structural issues concerning this kind of networks, so one key thing is to understand that a Bayesian network cannot include a cycle so you see here the left example, you see a valid the direct did a cyclic net graph and even if one of the arrows seems to point upwards this is more of depiction problem and and doesn't really cause any harm however as you see in the right example and the arrow between A and C is now directed in the opposite direction creating a cycle and hopefully you have understood from the way this kind of reasoning behaves is that it's strictly based on working either entirely in the forward direction or in the backward direction, and it's impossible to handle at this kind of cycle.

So the next structural aspect I want to mention is that of course when you study the same networks you can see patterns and there are patterns of different kinds but there are few very basic patterns and as you see here are three examples Sequence which is kind of obvious convergence where we're two variables cause a third and divergence where one variable have to two causes, so at that point we will not do much about this and it doesn't really we're not going to use the knowledge of patterns like this but when you handle big networks and especially if you want to modify networks and extend networks these structural aspects and the ability to treat these cases in a uniform fashion is more important.

So if we turn to problem solving for this kind of representation it's already said that the main idea is to use the networks to infer probabilities of causes from the probabilities of effects so because they are Bayesian network is a complete probabilistic model of the variables and relationship describing effects in terms of courses. Inferences typically aims to update the beliefs concerning the causes in the light of the new evidence, so backward inferences in a Bayesian network can be viewed as the answering all queries about the state of a subset of variables (hypothesis variables) or the variables that are the potential causes when other

variables are observed typically those considered the evidence variables. And the main vehicle for making this inference in the backward direction so to say is the use of Bayes theorem which actually states that the conditional probability of the hypothesis given some evidence is equal to the conditional probability of the evidence giving the hypothesis times the prior probability of the evidence divided by the prior probability of hypothesis.

$$P(H|E)=P(E|H)*P(E)/P(H)$$

So this Bayes theorem enables the carrying out of one inference step backwards in the structure. But because of the homogeneity of the structure this kind of inference can be recursively applied throughout the whole structure using the Bayes theorem in each step.

So what does learning mean in this kind of representation yeah actually there are two kinds of possibilities for learning, one kind of learning we call it a parameter learning and as you understand the only parameters basic parameters that we have are the conditional probabilities given a fixed variable structures, so we then assume we have a fixed number of variables and we have a fixed number of edges connecting the units corresponding to the variables, and in for every node we have this kind of table that describes the probability function for the conditional probability for that node. And essentially the low-hanging fruit here is of course to be able to manipulate and enhance the quality of these tables across the network so it better reflects the an adequate decision making for the domain and that of course can then be based on available data but this is kind of obvious very standardized learning process. What is more tricky of course is to learn in the sense of changing the structure and one level there is of course to assume that we have the same variables which means that we have the same nodes but we can modify the structure of edges among the number the number of edges among the two nodes, even more advanced to the same is the addition of new variables, typically is not so likely that we want to modify the low level evidence we can say this is the input layer of the structure and not necessarily so the most abstract hypothesis it's rather that it's more likely that we want to create some in-between variables which are non-observables and we could call them Hidden and you will see that there is a parallel here in the thinking about the world networks where we also will talk about input layer output layer and hidden layers. So this was the end of this lecture thanks for your attention and the next lecture will be on the topic of Neural Networks so thank you and good bye.