BAYES RULE AND BAYESIAN NETWORKS. CONTRAST WITH FREQUENTIST APPROACHES

1. Bayes Rule

The Economist article (In Praise of Bayes, dated September 30, 2000) introduces the concept of Bayes' Rule with the following example: *The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (i.e., the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become a near-certainty that the sun will always rise.*

Bayes Rule gives an approach to incorporate new information or evidence into your model (of the world) and adjust your beliefs accordingly. In mathematical terms it states

$$P(X = x|e) = \frac{P(e|X = x) \times P(X = x)}{P(e)} \tag{1}$$

We are trying to calculate the probability that the random variable X takes the value x given some event e has occurred. Bayes' Rule states that this can be decomposed as the conditional probability of the event e occurring given X takes the value x (i.e. $P(e|X = x)$), multiplied by the probability of the random variable taking that specific value: $P(X = x)$. We then need to divide out the probability of this event occurring as the event can occur whether the random variable X takes the specific value or not.

From the total probability rule, we know that $P(e)$ can be expressed as $P(e|X = x) + P(e|X \neq x)$ since either event $X = x$ occurs or it doesn't occur. These are the only two, mutually exclusive possibilities with respect to the random variable X. This allows us to state Bayes' Rule in yet another form:

$$P(X = x|e) = \frac{P(e|X = x) \times P(X = x)}{P(e|X = x) \times P(X = x) + P(e|X \neq x) \times P(X \neq x)} \tag{2}$$

Another form in which Bayes' rule might be encountered is as follows:

$$P(X = x|e) \times P(e) = P(e|X = x) \times P(X = x) = P(e, X = x) \tag{3}$$

This last form is sometimes very useful.

In Blackboard you will find a link to a non-technical overview of Bayes' Rule. Please read that to get a more intuitive understanding of this very important rule.

2. Bayesian Networks: Introduction

We'll briefly introduce the concept of Bayesian networks and graphical models. This is a very active research topic and a lot of this material is beyond the scope of this course. However, these are incredibly useful techniques and we'll briefly look at them this week. Bayesian Networks are one of the most exciting advances in machine learning in the recent decades. The key idea is to represent the world as a collection of random variables with a joint probability distribution. The structure of the distribution encodes your beliefs about the relationships between the variables.

The key idea of Bayesian Networks is to

- Represent the world as a collection of random variables: $X_1, \ldots, X_n$, together with their conditional independence relationships.
- Learn the probability distribution from data: $p(X_1, \ldots, X_n)$.
- Use the learnt probability distribution to perform inference when evidence is presented. E.g. $p(X_i|X_1 = x_1, \ldots, X_n = x_n)$.

A Bayesian Network is specified by a directed acyclic graph (DAG) $G(V, E)$ such that there is one node V for each random variable X_i and one conditional probability distribution, conditioning the random variable X_i on its parents, $X_{pa(i)}$. The joint probability density function is then described as follows:

$$p(X_1, \ldots, X_n) = \prod_{i \in V} p(X_i|X_{pa(i)}) \tag{4}$$

Let's look at an example Bayesian Network. We have 5 random variables. (D, I, G, S, L) – which correspond to *Difficulty, Intelligence, Grade, SAT Score, Letter of Recommendation.* If the student gets a good grade in that course, the student has a higher probability of receiving a good recommendation letter. Likewise, if the student in Intelligent, they have a higher chance of receiving a good SAT score. In general, higher the intelligence, the better their course grade and more difficult the course, the harder it is to get a good grade. This information can be encoded in the figure shown below. In addition, the conditional probability tables are shown next to the corresponding nodes. This graph represents a Bayesian network and factors the joint probability function $p(d, i, g, s, l) = p(d)p(i)p(g|i, d)p(s|i)p(l|g)$.

3. Bayesian versus Frequentist approaches to statistical inference

There are two approaches to statistical inference: Bayesian and Frequentist. As you can gather from online sources, there is a healthy debate between the two. In the past frequentist approaches held sway but over time Bayesian approaches have gradually gotten more accepted as it is able to answer some questions that are not easily answered with Frequentist approaches. It may be impossible to fit the entire controversy in the course notes. We'll provide a gist and give you pointers to research further on these approaches.

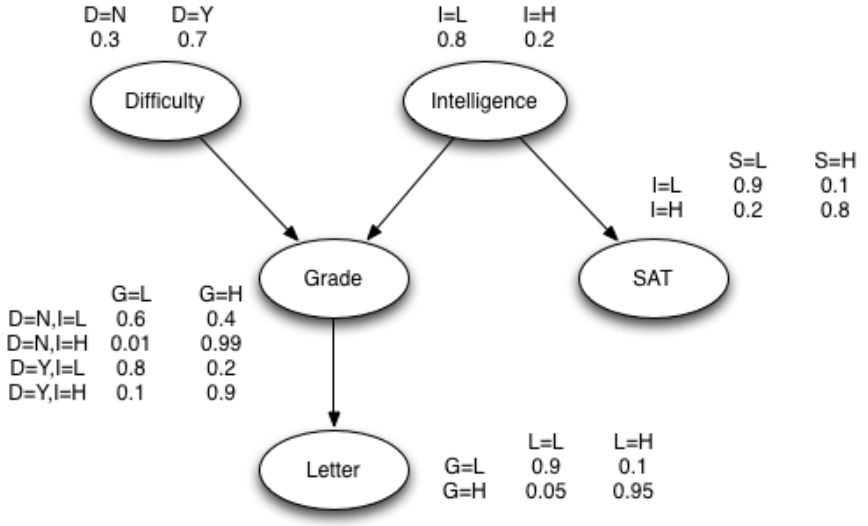

FIGURE 1. A Simple Bayesian Network

In many cases, both approaches agree and make reasonable claims. You may prefer one or the other. In general, Bayesian approaches have been more successful in the recent times and are considered main stream now.

In a frequentist view of the world, you assume have some data D and you are interested in finding out $P(D|H)$. That is given that hypothesis H is true, what is the probability of seeing such a data set D. There is no ambiguity about the hypothesis itself. Either the hypothesis is true or it is not. This works in many cases. For example, either you have cancer or not and given some blood test, you want to estimate the probability of getting that specific blood test result given the hypothesis of having cancer versus the hypothesis of not having cancer.

However, consider the following case: If you eat 9 servings of fruits and vegetables and exercise 30 minutes a day, will you get cancer when you turn 60? Can you say with certainty that you will or will not get cancer? Bayesian reasoning helps you in these types of situations which a frequentist approach finds difficult to handle. In a Bayesian approach, you have some data D and you are interested in finding out the probability of the hypothesis H given that you observed data D – that is, $P(H|D)$. Given that data D about your exercise and eating habits, what is the probability of getting cancer? Hopefully, Bayesian reasoning shows that it is a very low probability.

That is the crux of the difference between these two approaches. In Blackboard, you'll find some links to further your understanding between these two approaches.