Study Notes for AI (CS 440/520)
Lecture 10: Intro to Decision and Probability Theory

Corresponding Book Chapters: 13.2-13.3-13.4-13.5-13.6
Note: These notes provide only a short summary and some highlights of the material covered in the corresponding lecture based on notes collected from students. Make sure you check the corresponding chapters. Please report any mistakes in or any other issues with the notes to the instructor.

1 Uncertainty and probability theory

So far we have dealt with problems were sentences and propositions were either a true or a false. Once variables were assigned values, we were certain about the truth value of a sentence. This is not always true, however, in many real applications.
Consider the following example from the area of medical diagnosis:

$$\forall\ p \quad Symptom(p, Toothache) \Rightarrow Disease(p, Cavity).$$

Although this rule might sound reasonable, it is often wrong. There are many reasons why this sentence is wrong:

- Laziness: It takes too much work to list the complete list of consequences, e.g.,

$$\forall\ p\ Symptom(p, Toothache) \Rightarrow Disease(p, Cavity) \vee Disease(p, GumDisease) \vee Disease(p, Abscess)$$

- Ignorance: Perhaps medical science has no complete theory for the domain. Or we have not run all the necessary tests so as to apply medical science.

1.1 Probability Theory

Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance, e.g., there is an 80% chance that a patient has a cavity if patient has tooth ache.

How can we come up with probabilities that represent real events and their dependencies? Typically, there are two sources of information:

- Statistical data, e.g., after testing 1000 patients with toothache, 800 of them had a cavity
- From general rules, e.g., by construction there is a $\frac{1}{32}$ probability to roll 6 & 6 with dice.

Note that we still deal with problems were sentences are either true or false in reality. In probability theory, however, our degree of belief about a sentence varies between 0 and 1.

We can classify all probabilities either as prior (unconditional) or posterior (conditional) probabilities. The difference between them is that the latter are calculated after acquiring evidence.

1.2 Decision Theory

So how do we take decisions when we have uncertainty. Consider the following problem: How soon should we leave from our home to get to the airport?

- Action A_{90}: Leave 90 minutes prior to the departure of your flight. Assume the action has 95% chance of succeeding and a short waiting time.
- Action A_{120}: Leave 120 minutes prior to the departure of your flight. Assume the action has 99% chance of succeeding and a longer waiting time.

In order to decide which one of the above actions is the most beneficial we must first assign a utility to the end states that these actions can potentially bring us. For example, if it is catastrophic to miss the airplane, then the second action is preferable, otherwise we might prefer the first one. Assigning appropriate utilities to end states is the objective of utility theory. The combination of utility theory and probability theory is defined as decision theory:

$$DecisionTheory = ProbabilityTheory + UtilityTheory \tag{1}$$

Consequently, the question that arises is what does an intelligent agent do in the context of decision theory. We will call such agents **decision-theoretic agents**, they are agents that act intelligently under uncertainty. A decision-theoretic agent will select the action that yields the highest expected utility, averaged over all possible outcomes of the action (Maximum Expected Utility principle, MEU). In particular, the agent will follow these steps:

- Update a belief-state distribution based on actions/percepts (answering the question: what is the most probable state the world is currently in?)
- Calculate outcome of actions given action descriptions and current belief state (compute the utility of each action)
- Select action with the highest expected utility given probabilities of outcomes and utility information
- Return action

In order to reach the point where we can describe concrete decision-theoretic agents, we will first review the tools of probability theory.

2 Axioms of probability and Bayes rule

The following nomenclature is useful for the following discussion:

- Random variables: each variable represents a subset of the world that we don't know (e.g. Cavity: do or do I not have a cavity in my left wisdom tooth?)
- Domain: possible values each random variable can acquire
- Propositions: as in logic, sentences that may be true or false for assignments of values to variables. We can use propositional logic to create more complex propositions from random variables (e.g. $cavity \land \neg toothache$)
- Boolean random variables: true/false
- Discrete random variables: discrete values
- Continuous random variables: continuous values
- Atomic event: an assignment of values to all random variables

2.1 Conditional probability

As mentioned before, probabilities can be further classified as conditional and unconditional. Conditional probabilities are calculated "given, |" values for some other probabilities. In other words, they indicate what is the probability of some event if all that we know is that the other event holds. For example the term $P(a|b)$ represents the probability of a given that all we know is b. Unconditional probabilities, however, are calculated regardless of any other events. $P(a)$ represents the probability of a being true in the absence of any other information.
The mathematical definition of conditional probability, also known as the **product rule** is the following:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \tag{2}$$

It can also be defined via the use of a scaling factor (e.g., when we are interested in comparing $P(cavity|threshold)$ and $P(\neg cavity|threshold)$ the denominator does not effect which one of the two probabilities is greater):

$$P(a|b) = \alpha \times P(a \wedge b) \quad \text{where} \quad \alpha = \frac{1}{P(b)} \tag{3}$$

The following relations are equivalent to the product rule:

$$P(a \wedge b) = P(a|b) \times P(b) \tag{4}$$

$$P(a \wedge b) = P(b|a) \times P(a) \tag{5}$$

2.2 Axioms of probability

Axioms of probability are fundamental rules to standardize probability problems. **Axioms of probability** (Kolmogorov's axioms)

- $0 <= P(a) <= 1, \quad \forall a$
- $P(true) = 1, \ P(false) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Other useful rules can be derived from Kolmogorov's axioms and the definition of conditional probability:
e.g what is $P(\neg a)$?
From axiom 3: $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a) \Rightarrow$
From axioms 1, 2: $P(true) = P(a) + P(\neg a) - P(false) \Rightarrow$
$1 = P(a) + P(\neg a) \Rightarrow$
$P(\neg a) = 1 - P(a)$

In general, $\sum_i P(D = d_i) = 1, \ D =< d_1, ..., d_n >$

2.3 Inference using full joint distributions

When we have available joint probabilities ($P(x, y) = P(X \wedge y)$), we can calculate a particular prior probability by applying **marginalization**:

$$P(x) = \sum_{y \text{ value of } Y} P(x, y) \tag{6}$$

Using this technique, we basically "sum out" the joint probabilities containing the variable we are interested in and the other variable, for all the values of the other variables.

Consider for example the full joint probability distribution for the cavity problem in Figure 1. If we are interested in computing $P(cavity)$ then we must sum the probabilities in the top row cells on the table. This operation will correspond to computing: $\sum_{catch,toutache} P(cavity, catch, toothache)$, which is what the marginalization rule specifies.

	touthache		¬touthache	
	catch	¬catch	catch	catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Figure 1: The full joint probability distribution for the cavity problem.

If we combine marginalization with the product rule, then we acquire the **conditioning** rule:

$$P(x) = \sum_{y \text{ value of } Y} P(x|y)P(y) \tag{7}$$

A more general version of the conditioning rule is the following:

$$P(x|z) = \sum_{y \text{ value of } Y} P(x,y|z)P(y|z)$$

2.4 Independence

Consider the joint probability distribution:

$$P(\underbrace{toothache, catch, cavity, \overbrace{weather}^{values: rainy, cloudy, sunny, snow}}_{\text{32 probabilities needed to represent this joint distribution}})$$

Note, however, that weather does not effect whether a patient has a tooth ache, a cavity or a catch. The weather variable is **independent** from the remaining variables in this joint distribution. In this case, the above probability can be simplified to the following (by applying the product rule):

$$P(weather|ache, catch, cavity) \times P(ache, catch, cavity) = P(\underbrace{weather}_{\text{4 values needed}}) \times P(\underbrace{ache, catch, cavity}_{\text{8 values needed}})$$

In general for independent variables a and b:

$$P(a|b) = P(a) \text{ or } P(b|a) = P(b) \text{ or } P(a \wedge b) = P(a)P(b)$$

2.5 Bayes rule

Bayes rule is one of the most useful rules in statistics, it allows to formulate conditional probability problems through reversed conditional dependence:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} \tag{8}$$

A more general version of the Bayes rule, which allows for a background evidence e, is the following:

$$P(Y|X,e) = \frac{P(X|Y,e)P(Y|e)}{P(X|e)}$$

If we are interested in comparing the probabilities $P(b|a)$ and $P(\neg b|a)$ then we can avoid the requirement of knowing the denominator ($P(a)$) and treat it as a normalization factor:

$$< P(b|a),\ P(\neg b|a) >=< \alpha P(a|b)P(b),\ \alpha P(a|\neg b)P(\neg b) >$$

Let's apply the Bayes rule in a medical diagnosis application:

Meningitis causes stiff neck 50% of the time
1/50,000 people have meningitis
1 out of 20 people have stiff neck

$$P(m|s) = \frac{P(s|m) \times P(m)}{P(s)} = \frac{0.5 \times 1/50,000}{1/20} = 0.0002$$

We can avoid using $P(s)$ and instead treat it as normalization factor:

$$P(m|s) = \alpha < P(s|m) \times P(m), \qquad \underbrace{P(s|\neg m)}_{\text{but we have to know this variable}} \qquad \times P(\neg m) >$$

2.6 Conditional independence

Another example:

$$P(cavity|toothache \wedge catch) = \alpha P(toothache \wedge catch|cavity)P(cavity)$$

Tooth ache and catch are NOT independent in general. They are, however, if we know whether we have a cavity or not as they are effects of the same cause. In this case, the variables are conditionally independent and the following is true:

$$P(ache \wedge catch|cavity) = P(ache|cavity) \times P(catch|cavity)$$

In general, if variables x and y are conditionally independent:

- $P(x,y|z) = P(x|z)P(y|z)$
- $P(x|y,z) = P(x|z)$
- $P(y|x,z) = P(y|z)$

When we have a joint distribution of cause and effect variables, then the Naive Bayesian model can be used to simplify the computation:

$$P(Cause, Effect_1, ..., Effect_n) = P(Cause)\prod_i P(Effect_i|Clause) \tag{9}$$

It is called naive because it is often used in cases where the effect variables are not truly conditionally independent but this simplifying assumption is being made.