COSC 348:
Computing for Bioinformatics

Lecture 8:
Introduction into probability theory

Lubica Benuskova

http://www.cs.otago.ac.nz/cosc348/

Probability

- Probability summarizes our uncertainty about the world
  - E.g., there is a probability of $P = 0.6$ that patient's toothache is caused by a cavity in the tooth.
  - That is, we believe there is an 60% chance a patient with a toothache has a cavity.
  - The rest $P = 0.4$ (40%) summarizes all other causes.
- Probability $0 \leq P \leq 1$ corresponds to a **degree of belief** in the truth of a given proposition:
  - $P = 1$ corresponds to a belief that a given sentence is *true*
  - $P = 0$ corresponds to a belief that a given sentence is *false*

Random variable

- **Random variable *X***: basic element of the probability theory. Describes the property of the world such that it assumes concrete values with some probability, i.e. the values of X are assigned with some probability.
- According to the set of all possible values:
  - **Boolean random variables**: have the values ⟨*true, false*⟩. Example: *Cavity*
  - **Discrete random variables**: exhaustive and countable domain of mutually exclusive values. Example: *Nucleotide* = ⟨*A, C, T, G*⟩
  - **Continuous random variables**: values from the real numbers, either the entire line or some subset. Example: *Height* = 164cm

Evidence

- Agent's beliefs depend on its observations to this date.
- These observations constitute the **evidence**, on which probability assertions are made
  - Example: Before drawing a card from a shuffled pack, the agent assigns $P = 1/52$ to a drawn card to be the ace of spades. After looking at the drawn card, the probability of the same proposition is either = 0 or 1.
- Before the evidence is obtained we talk about the **prior** or unconditional probability; after the evidence is obtained we talk about **posterior** or conditional probability.
- Probabilities can change when more evidence is acquired.

Prior (unconditional) probability

- Before the evidence is obtained we talk about **prior** or unconditional probability of a proposition a, written as $P(a)$, that corresponds to belief prior to arrival of any evidence.
  - Elementary proposition is the assignment of value, e.g. *Weather* = *sunny*
- **P**(*Weather*) denotes the **vector** of probability values for each individual state of the weather, the so-called **prior probability distribution**
- E.g., for the random variable *Weather* = ⟨*sunny, rainy, cloudy, snow*⟩, the prior probability distribution reads: **P**(*Weather*) = ⟨ 0.72, 0.1, 0.08, 0.1⟩ (probabilities are normalized, i.e. they sum to 1)

Posterior (conditional) probability

- After the evidence is obtained we talk about **posterior** probability
  - e.g., $P(cavity \mid toothache) = 0.6$, given that *toothache* is **all** I know
- Notation: $P(a \mid b)$, where a and b are (any) propositions, reads as "the probability of a, given **all we know** is b"
- If we know more, e.g., *cavity* is also given, then we (trivially) have $P(cavity \mid toothache \wedge cavity) = 1$. (Note: $\wedge$ is symbol for logical &)
- New evidence may be irrelevant, allowing simplification, e.g., $P(cavity \mid toothache \wedge sunny) = P(cavity \mid toothache) = 0.6$

Product rule

- Definition of posterior probability *in terms of prior probabilities*

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

- The last equation can be rewritten as the so-called **product rule**

$$P(a \wedge b) = P(a|b)\ P(b) = P(b|a)\ P(a)$$

  – Meaning for a and b to be true, we need b to be true and we also need a to be true given b or we need a to be true and we also need b to be true given a

- Posterior probabilities are vehicles of probabilistic inference

Bayes' rule

- Main formula of probabilistic reasoning, derived from the product rule.

- Recall the definition of the product rule

$$P(a \wedge b) = P(a\,|\,b)\ P(b)$$
$$P(a \wedge b) = P(b\,|\,a)\ P(a)$$

- Equating the two right-hand sides, and dividing by $P(a)$ we get **the Bayes' rule**

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

Bayes' rule: example

- What is the conditional (posterior) probability that a patient has meningitis when his/her neck is stiff? $P(m \mid s)$=?
  – The doctor knows that meningitis causes a stiff neck in 50% of cases, that is $P(s \mid m) = 0.5$
  – The doctor knows the prior probability of meningitis $P(m) = 1/50000$
  – The doctor knows the prior probability of stiff neck $P(s) = 1/20$

- Applying the Bayes' rule: $P(m\,|\,s) = \frac{P(s\,|\,m)\,P(m)}{P(s)} = 0.0002$

- That is, we expect 1 in 5000 patients with stiff neck to have meningitis.

Bayes' rule: cause and effect

- Bayes' rule for variables X and Y

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

- Can be rewritten as a rule for cause and effect

$$\mathbf{P}(Cause\,|\,Effect) = \frac{\mathbf{P}(Effect\,|\,Cause)\,\mathbf{P}(Cause)}{\mathbf{P}(Effect)}$$

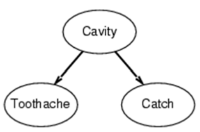
- Useful for assessing diagnostic probability from causal probability, because it is easier to know the conditional probabilities of effect given the cause.

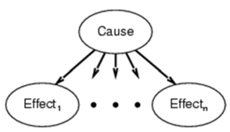
Conditional independence

- In reality, *a single cause can directly influences a number of effects*, all of which are conditionally independent given the cause.

- Thus, the full joint distribution can be then written as

$$\mathbf{P}(Cause \wedge Effect_1 \wedge \ldots \wedge Effect_n) = \alpha\,\mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i\,|\,Cause)$$





Bayesian network: specification

1. A set of random variables makes up the **nodes**;
2. A set of directed links (**arrows**) connects pairs of nodes;
3. The meaning of an arrow: X_j has a direct influence upon X_i.
4. If there is an arrow from node (variable) X_j to node (variable) X_i, the variable X_j is said to be a parent of X_i;
5. Each node X_i has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node.

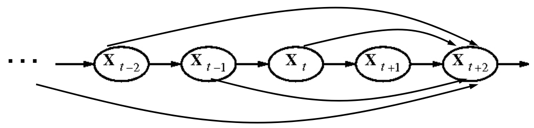
Conditional probability table (CPT)

- A node with ***no parents*** has only one row with ***prior probabilities*** of each possible value of variable.
- Each row of probabilities must sum to 1, but we often omit the column for negation because $P(\neg X) = 1 - P(X)$
- Distribution of probabilities associated with each ***node with parents*** is called **conditional probability table (CPT)**
- Each row in CPT contains the conditional probability for *a conditioning case (*which is a combination of values for the parent nodes) and sums to 1.

Markov models

- Markov models are statistical models that describe the change of **states,** i.e. random variable *X*, **in time** using probability.
- In an *ordinary Markov model*, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters.
- In general, the current state may depend on all previous states, thus:

$$\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1} \wedge \mathbf{X}_{t-2} \wedge \ldots \wedge \mathbf{X}_0)$$



Markov assumption

- Markov assumption: the current state depends only on a finite history of previous states (Markov process, Markov chain).
- Markov process of the k^{th} -order: Process, in which the current state depends on *k* previous states, and not on any earlier states, thus:

$$\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{0:t-1}) \Rightarrow \mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1} \wedge \mathbf{X}_{t-2} \wedge \ldots \wedge \mathbf{X}_{t-k})$$

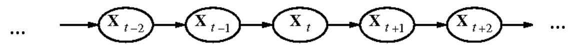
  – Notation *j*:*k* will be used to denote the sequence of time steps from time *j* to time *k* (inclusive).

Markov process of the 1st order

- The 1st-order Markov process: the current state depends only on the previous state, and not on any earlier states.
- The laws describing how the state evolves over time are contained entirely within the conditional distribution, called the **transition model**

$$\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{0:t-1}) \Rightarrow \mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1})$$

- The topology of the Bayesian network for state transitions:



Example: model of language

- Let random variable (state) be the variable *Word*, which can have these discrete values:
  – *Word* = {From women's eyes this doctrine I derive: they sparkle still the right Promethean fire; they are the books, the arts, the academes, that show, contain and nourish all the world: else none at all in ought proves excellent. }
- Sequence of *n* words is denoted by $w_1 \ldots w_n$, and w_t denotes the word at position *t* of the sequence.
- What is the probability of this particular sequence of words?

$$P(w_1 \wedge \ldots \wedge w_n) = ?$$

Statistical model of language

- The expression for the probability of this sequence with the use of the product rule for *n* variables reads:

$$P(w_1 \wedge \ldots \wedge w_n) = \prod_{t=1}^{n} P(w_t \mid w_1 \wedge \ldots \wedge w_{t-1}) =$$

$$= P(w_n \mid w_1 \wedge \ldots \wedge w_{n-1}) \ldots P(w_3 \mid w_1 \wedge w_2) P(w_2 \mid w_1) P(w_1)$$

- Most of these terms are very difficult to estimate or compute. Thus we have to simplify.

Bigram model of language

- **Bigram model** is a Markov model of the 1st order.

$$P(w_1 \wedge \ldots \wedge w_n) \approx \prod_{t=1}^{n} P(w_t \mid w_{t-1})$$

$$P(w_1 \wedge \ldots \wedge w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)\ldots P(w_n \mid w_{n-1})$$

- Calculation of **transition probabilities**: count the number of times each word pair occurs in a **representative** corpus, and use the counts to estimate the transitional conditional probabilities.
  - if "they" appears 1000 times and is followed by "are" 30 times then the estimate of the probability of transition is $P(are_t \mid they_{t-1}) = 30 / 1000 = 0.003$.

Trigram and other models

- **Trigram model** corresponds to

$$P(w_1 \ldots w_n) \approx \prod_{t=1}^{n} P(w_t \mid w_{t-1}\, w_{t-2})$$

- For trigram and bigram models we must deal with *zero counts*. In these cases we use a small nonzero number or the process of *smoothing* gives a non-zero probability to such instances.

- Bi- and trigram models are sensitive only to a local context and local syntax, however they fail for long distance relationships.

- Grammar and syntax models exist that are better.