208891 Probabilistic Graphical Models **Spring 2019**

Lecture 2: December 23

Lecturer: Donlapark Pornnopparath

2.1 Overview of the course

- **Representation**: Our goal is to represent a real-world phenomenon in a form of a tractable graphical model that reflects our understanding of the domain. We will be using tools from graph theory and probability to study the relations between the variables and complexity of the model.
- **Inference**: Suppose that a graphical model is present. Our goal now is to find a way to preform a real-world task this model. We will focus on two of the most common inference task:
  - *Marginal inference*: We want to calculate the marginal probability of a variable, given a background event E:
$$\mathbb{P}(y|E=e) = \sum_{x_1,x_2,\ldots,x_n} \mathbb{P}(y, x_1, x_2, \ldots, x_n|E=e).$$
  - *Maximum a posteriori (MAP)*: We want to find the most likely configuration that gives rise to an observed event. For example, if we observed that the user clicked the link ($y=1$) then we want to find a situation which is most probable i.e.
$$\max_{x_1,x_2,\ldots,x_n} \mathbb{P}(x_1, x_2, \ldots, x_n|y=1)$$

  Sometimes, these inference tasks are intractable and we have to resort to approximation methods. As these algorithms became very crucial for high-dimensional inference, we will go over these algorithms in details in this part of the course.
- **Learning**: The last part focuses on building the model from the real-world data. Most of the time we will turn this into a problem of optimizing an objective function, and sometimes our learning algorithm will be repeatedly making inference in an attempt to match the inferred values with the actual outcome.

2.2 Conditional probability

The conditional probability of event A after observing event B is given by
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)},$$
which is equivalent to the **chain rule**:
$$\mathbb{P}(A,B) = \mathbb{P}(A|B)\mathbb{P}(B),$$
and in general, when we have n random events,
$$\mathbb{P}(A_1, A_2, \ldots, A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\ldots\mathbb{P}(A_n|A_1, A_2, \ldots, A_{n-1}). \tag{2.1}$$
Often, there will be a background event C in our consideration, in which case we have
$$\mathbb{P}(A|B,C) = \frac{\mathbb{P}(A,B|C)}{\mathbb{P}(B|C)}.$$

2.3 Independence of random variables

Notation: Let X and Y be random variables. The notation $X \perp\!\!\!\perp Y$ denotes that X is independent of Y.

Definition 2.1. Let X, Y and Z be random variables. Then we say that X is conditionally independent of Y given Z if

$$\mathbb{P}(X|Y,Z) = \mathbb{P}(X|Z).$$

In this case, we write $X \perp\!\!\!\perp Y|Z$.

Equivalently, we have the following decomposition:

$$\mathbb{P}(X,Y|Z) = \mathbb{P}(X|Y,Z)\mathbb{P}(Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z).$$

Note that this condition is symmetric: if $X \perp\!\!\!\perp Y|Z$ then $Y \perp\!\!\!\perp X|Z$.

All the identities that we have been mentioned also hold for continuous variables. For example, for continuous X, Y, Z with $X \perp\!\!\!\perp Y|Z$, we have

$$f_{XY|Z}(x,y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

and for continuous $X_1, X_2, \ldots, X_n$ we have the chain rule

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = f_{X_1}(x_1)f_{X_2|X_1}(x_2,x_1)\cdots f_{X_n|X_1,\cdots,X_{n-1}}(x_n|x_1,\cdots,x_n).$$

To compute the conditional probability from a given data, we often need the **Bayes' rule**

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

and in the continuous case,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X_Y}(x|y)f_Y(y)}{\int f_{X_Y}(x|y)f_Y(y)\ dy}.$$

2.4 Bayesian network

In the case that we can represent the dependencies between variables, it is natural to represent the model as a *directed acyclic graph (DAG)*. From the chain rule (2.1),

$$\mathbb{P}(X_1, X_2, \ldots, X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\ldots\mathbb{P}(X_n|X_1, X_2, \ldots X_{n-1}). \tag{2.2}$$

By inserting our knowledge into the model, we reduce each term on the right hand side so that it depends only on a few variables:

$$\mathbb{P}(X_i|X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i|X_{A_i}), \tag{2.3}$$

where we used the notation $X_A = \{X_j : j \in A\}$ and $A_i \subseteq \{1, 2, \ldots i-1\}$. This is called a *Bayesian network*.

Definition 2.2. A Bayesian network is a directed acyclic graph $G = (V, E)$ which specifies a random variable X_i for each node $i \in V$ and a conditional probability distribution $\mathbb{P}(X_i|X_{A_i})$ where A_i is the set of parents of i. Moreover, G specifies the following conditional independence:

$$X_i \perp\!\!\!\perp X_{D_i^C}|X_{A_i},$$

where D_i^C is the set of nondescendants of i.

We can see from (2.2) and (2.3) that the joint probability distribution over all variables is defined by the Bayesian network. Moreover, it can be factored as a product of the conditional probability distribution specified by G.

2.4.1 Simple Bayesian networks

As a demonstration, we compute the conditional probability imposed by the most basic Bayesian networks.

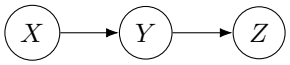

Figure 2.1: a little sequence.

A little sequence

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y|X)\mathbb{P}(Z|Y).$$

In this case, $X \perp\!\!\!\perp Z|Y$, which follows from

$$\mathbb{P}(X|Y, Z) = \frac{\mathbb{P}(X, Y, Z)}{\mathbb{P}(Y, Z)} = \frac{\mathbb{P}(X)\mathbb{P}(Y|X)\mathbb{P}(Z|Y)}{\mathbb{P}(Z|Y)\mathbb{P}(Y)} = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(Y)} = \mathbb{P}(X|Y).$$

Note that this does not imply that $X \perp\!\!\!\perp Z$. For example, let $X =$ `it was summer`, $Y =$ `hot weather` and $Z =$ `a fan was turned on`. Given that the weather was hot, then the fan would be turned on regardless of the season. However, it is more likely that one will turn on a fan during the summer.

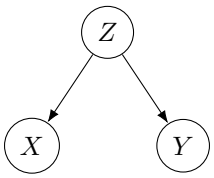

Figure 2.2: a little tree.

A little tree

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(Z)\mathbb{P}(X|Z)\mathbb{P}(Y|Z)$$

It follows that $X \perp\!\!\!\perp Y|Z$ from

$$\mathbb{P}(X|Y, Z) = \frac{\mathbb{P}(X, Y, Z)}{\mathbb{P}(Y, Z)} = \frac{\mathbb{P}(Z)\mathbb{P}(X|Z)\mathbb{P}(Y|Z)}{\mathbb{P}(Y, Z)} = \frac{\mathbb{P}(X|Z)\mathbb{P}(Y, Z)}{\mathbb{P}(Y, Z)} = \mathbb{P}(X|Z).$$

Again, this does not imply that $X \perp\!\!\!\perp Y$. For example, let $Z =$ `hot weather`, $X =$ `ice cream was bought` and $Y =$ `a fan was turned on`. Once we know the past weather condition, the event of ice cream being bought does not affect the chance of the fan being turned on. However, without knowing the weather condition, observing that one went out and bought ice cream would reduce the chance of the fan being turned on.

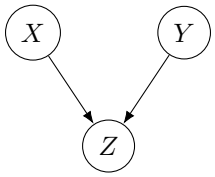

Figure 2.3: a little V.

A little V

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y),$$

which implies that $X \perp\!\!\!\perp Z$ from

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X, Y)\mathbb{P}(Z|X, Y) = \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y).$$

This does not imply that $X \perp\!\!\!\perp Y|Z$. For example, if $X =$ `sprinkler was turned on`, $Y =$ `it was raining` and $Z =$ `the lawn is wet`. Even though X and Y are independent, if the lawn is wet and it was raining a few moment ago, then the chance that the sprinkler was turned on would be lowered.