

# Introduction to Statistical Data Science

October 18, 2018

## 1 Lecture 1

### Basic Rules of Probability

Frequently we will say  $p(x) \propto f(x)$  for some non-negative function  $f(x)$

Then we can conclude that:

$$p(x) = \frac{f(x)}{\sum_y f(y)}$$

For joint distribution,

$$\sum_x p(x) = \sum_x \sum_y p(x, y) = 1$$

### Independence

If  $p(x|y) = p(x)$  for all states of  $x$  and  $y$ , then the variables  $x$  and  $y$  are said to be independent as  $x \perp\!\!\!\perp y$ .

If  $x$  and  $y$  are independent, then  $x$  and  $y$  are uncorrelated. However, in general,  $x$  and  $y$  are uncorrelated, then cannot conclude that  $x$  and  $y$  are independent.

### Conditional Independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

$$p(\mathcal{X}, \mathcal{Y} | \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z}) p(\mathcal{Y} | \mathcal{Z})$$

and

$$p(\mathcal{X} | \mathcal{Y}, \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z})$$

Conditional independence does not imply marginal independence:

$$p(x, y) = \sum_z p(x|z)p(y|z)p(z) \neq \sum_z p(x|z)p(z) \sum_z p(y|z)p(z)$$

## 2 Lecture 2

### Graphs

Definition:

A graph consists of nodes (vertexes) and undirected or directed links (edges) between nodes.

Path:

A path from  $X_i$  to  $X_j$  is a sequence of connected nodes starting at  $X_i$  and ending at  $X_j$ . (no direction)

Directed Acyclic Graph:

Graph in which by following the direction of the arrows a node will **never** be visited **more than once**.

Parents and Children:

$X_i$  is a parent of  $X_j$  if there is a link from  $X_i$  to  $X_j$ .  $X_i$  is a child of  $X_j$  if there is a link from  $X_j$  to  $X_i$ .

Ancestors and Descendants:

The ancestors of a node  $X_i$  are the nodes with a directed path ending at  $X_i$ . The descendants of  $X_i$  are the nodes with a directed path beginning at  $X_i$ .

**Undirected Graph:**

Clique:

A clique is a fully connected subset of nodes.

Maximal Clique:

Clique which is not a subset of a larger clique.

Connected graph:

There is a path between every pair of vertices.

**Connected components:**

In a non-connected graph, the connected components are the connected-subgraphs.

**Connectedness: Singly-connected**

There is only one path from any node  $a$  to another node  $b$

**Multiply-connected**

A graph is multiply-connected if it is not singly connected.

## **Belief Networks (Bayesian Networks)**

A belief network is a **directed acyclic graph** in which each node is associated with the conditional probability of the node given its parents.

### **Processing the network**

Firstly write the whole joint distribution such as:

$$p(A, R, E, B) = p(A|R, E, B)p(R|E, B)p(E|B)p(B)$$

Then, according to the assumption, remove some independent variable from the joint distribution. **It does matter that the order of joint distribution influence the processing.**

### **Uncertain Evidence**

In soft/uncertain evidence the variable is in more than one state, with the strength of our belief about each state being given by probabilities. For example, if  $y$  has the states  $dom(y) = \{\text{red, blue, green}\}$  the vector  $(0.6, 0.1, 0.3)$  could represent the probabilities of the respective states

In the calculation, we can do this: Given  $P(A = tr) = 0.7$

$$p(B = tr|\tilde{A}) = \sum_A p(B = tr|A)p(A|\tilde{A})$$

### **Independence**

If  $C$  has more than one incoming link, then  $A \perp\!\!\!\perp B$  and  $A$  is not conditional independent with  $B$  under  $C$  condition. In this case  $C$  is called collider. If  $C$  has at most one incoming link, then

$A \perp\!\!\!\perp B|C$  and A is not independent with B. In this case  $C$  is called non-collider.

### **d-connected/separated**

X and Y are d-connected by Z if there is any path from X to Y that is not blocked by Z

If all of the paths are blocked then we say X and Y are d-separated by Z.

### **Markov Equivalence**

skeleton

Formed from a graph by removing the arrows.

immorality

An immorality in a DAG is a configuration of three nodes, A,B,C such that C is a child of both A and B, with A and B not directly connected.

### **Markov Equivalence**

Two graphs represent the same set of independence assumptions if and only if they have the same skeleton and the same set of immoralities.

### **BN representation**

BN cannot represent whatever independence statements are present in  $p$ .

Fundamentally, the actual **numerical distribution**  $p$  **contains much more information than a graph can represent.**

## **Markov Network**

A Markov Network is an undirected graph in which there is a potential (non-negative function)  $\psi$  defined on each maximal clique.

The joint distribution is proportional to the product of all clique potentials:

$$P(Y) = \frac{1}{Z} \prod_C \Psi_C(Y_C)$$
$$z = \sum_Y \prod_C \Psi_C(Y_C)$$

where the  $\Psi_C(Y_C)$  is a potential function (strict positive). The potential function would usually indicated as

$$\Psi_C(Y_C) = \exp\{-E(Y_C)\}$$

## Iterated conditional Modes

To find the minimum of a function  $f(x_1, \dots, x_D)$ , firstly make a guess  $x_1^*, \dots, x_D^*$ . Then look at the single variable  $x_i$  keeping all others fixed. Then set  $x_i^* = \operatorname{argmin}_{x_i} f(x_i)$ . Then repeat this, sweeping through all variables, usually in a random order. Then repeat the whole process until convergence.

## Boltzmann machine

### Ising model

Given  $x_i \in +1, -1$ , then the joint distribution would be:

$$p(x_1, \dots, x_D) = \frac{1}{Z} \prod_{i \sim j} \phi_{ij}(x_i, x_j)$$

$$\phi_{ij}(x_i, x_j) = e^{-\frac{1}{2T}(x_i - x_j)^2}$$

## Rule for independence in Markov Networks

Remove all links neighboring the variables in the conditioning set  $Z$ .

If there is no path from any member of  $x$  to any member of  $y$ , then  $x$  and  $y$  are conditionally independent given  $z$ .

## Factor Graphs

A square node represents a factor (non negative function) of its neighboring variables. FGs are most commonly used for inference.

## Markov Models

### Time-Series

A time-series is an ordered sequence:

$$x_{a:b} = \{x_a, x_{a+1}, \dots, x_b\}$$

For the time series data, we need a model  $p(v_{1:T})$ . For the causal consistency, it is meaningful to consider the decomposition.

$$p(v_{1:T}) = \prod_{t=1}^T p(v_t | v_{1:t-1})$$

with the convention  $p(v_t | v_{1:t-1}) = p(v_1)$  for  $t = 1$ .

Only the recent past is relevant:

$$p(v_t | v_1, \dots, v_{t-1}) = p(v_t | v_{t-L}, \dots, v_{t-1})$$

where  $L \geq 1$  is the order of the Markov chain.

$$p(v)$$