Introduction to Statistical Data Science

October 13, 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 (vertixes) 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:

Xi is a parent of X_j if there is a link from X_i to X_j. Xi 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 α 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 $P(A = tr) = 0.7$

$$p(B = tr|\widetilde{A}) = \sum_{A} p(B = tr|A)p(A|\widetilde{A})$$

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**.