Week5 Report

Zhengyi Wang
wangzhen17@mails.tsinghua.edu.cn

Abstract

This week, we talked about the fundemental of machine learning. Our talk is basically on *Problistic Graphic Model* and related topics.

1 PGM

PGM, short for Problistic Graphic Model, refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables.

1.1 Bayersian Network

BN, short for Bayersian Network, is a directed graph whose vertexes are the events and edges are about the casuality relationship. The graph should be topological sorted.

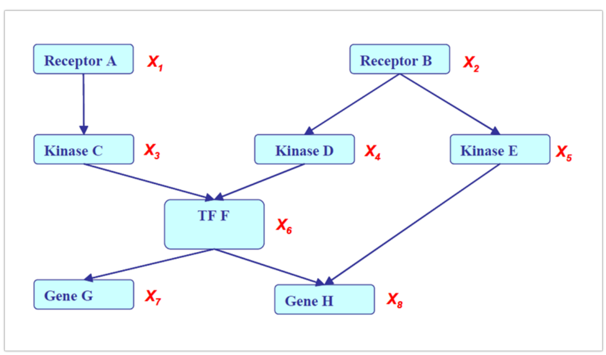

Figure 1: A BN example

Figure 1 shows an example of BN. In which we can represent the joint distribution as

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) =$$

$$P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3, X_2)P(X_7|X_6)P(X_8|X_5, X_6)$$

1.1.1 Independency in BN

Global Independency Given all other nodes, X is independent of any of the other nodes except its parents, children and children's co-parents.

Local Independency Each node is independent of its non-descendants given its parents.

Reading Group on Machine Learning (May 26th, 2019).

1.2 Markov Random Field

MRF, short for Markov Random Field, is another kind of graphic model. Different from Bayersian Network, it's an undirected graph whose vertexes also represent the events. However its edges show the correlation between events rather than casuality relationship.

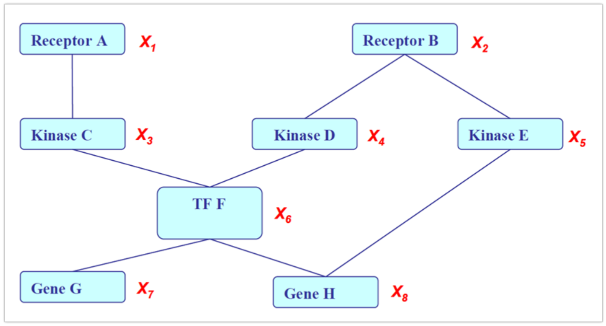

Figure 2: A MRF example

Figure 2 shows an example of MRF. In which the joint distribution can be discribed as followed,

$$\frac{1}{Z}e^{E(X_1)+E(X_2)+E(X_3,X_1)+E(X_4,X_2)+E(X_5,X_2)+E(X_6,X_3,X_4)+E(X_7,X_6)+E(X_8,X_5,X_6)}$$

In which, E is called energy function.

1.2.1 Cliques

For G={V,E}, a complete subgraph, *clique*, is a subgraph G'={V'⊆V,E'⊆E} such that nodes in V' are fully interconnected.
A *maximal clique* is a complete subgraph *s.t.* any superset V" of V' is not complete.

1.2.2 Definition of MRF

An *undirected graphic model*, also known as *Markov Random Field*, *Markov Networks*, represents a distribution $P(X_1, ..., X_n)$ defined by an undirected graph H and a set of positive potential function Φ_c associated with the cliques (which can be written as maximum cliques) of H s.t.

$$P(x_1, ..., x_n) = \frac{1}{Z}\Pi_{c \in C}\Phi_c(x_c)$$

where C is the set of cliques. And Z is for nomalization.

1.2.3 Independence in MRF

We can learn from the form of the probability distribution of MRF that a node in MRF is only dependent on its neighbors.

$$P(X_u = x_u|X_v, v \neq u) = P(X_u = x_u|X_v, v \in Neighborhood_u)$$

From that we have the following character about MRF's independence.

Global Markov Independence A probability distribution satisfies the global Markov property if for any disjoint A, B, C, such that B separates A and C, A is independent of C given B.

Local Markov Independence For each node $X_i \in V$, there is a unique Markov blanket of X_i, denoted by MB_{X_I}, which is the set of neighbors of X_i in the graph (those that share an edge with X_i). The local Markov independencies associated with H is

$$l_l(H) = \{X_i \perp V - X_i - MB_{X_i}|MB_{X_i} : \forall i\}$$

Pairwise Markov Independecne

$$I_p(H) = \{X \perp Y | V/\{X, Y\} : \forall \{X, Y\} \notin E\}$$

1.2.4 Exponential Form

Constraining clique potential to be positive can be inconvenient, so we represent clique potential $\Phi_c(x_c)$ in an unconstrained form using a real-value "energy function" $\phi_c(x_c)$

$$\Phi_c(x_c) = e^{\phi_c(x_c)}$$

This gives the joint a nice additive form

$$p(x) = \frac{1}{Z} e^{-\sum_{c \in C} \phi_c x_c} = \frac{1}{Z} e^{-H(x)}$$

in which H(x) is called *free energy*. In phisics it's called *Boltzman Distribution*. In statistics it's called *log-linear model*.

2 Exact Inference

Exact inference is one kind of inference method to use our PGM model to inference and predict.

2.1 Elimination Algorithm

The elimination algorithm is used for exact inference.

2.1.1 Basic Idea about Elimination Algorithm

The idea is simple. If we want to compute the distribution about some of the variables, we can eliminate other variable step-by-step using maginalization.

2.1.2 Complexity about Elimination Algorthm

The complexity is highly dependent on the elimination order and is **exponential**.