10-708: Probabilistic Graphical Models 10-708, Spring 2014

1 : Introduction

Lecturer: Eric P. Xing *Scribes: Yuxing Zhang, Tianshu Ren*

1 Logistics

There is no exam in this course, and the grade breakdown is as follows:

- Homework assignments (40%)
  - There will be 5 assignments in total
- Scribe duties (10%)
  - Needs to be submitted within one week after the lecture, people should sign up on the spreadsheet first
- Short reading summary (10%)
  - Should be handed to TA in paper at the beginning of each lecture
- Final project (40%)
  - Could be theoretical or algorithmic research, could also be applications of PGM algorithms to real problems
  - Group project with team size 3
  - Final presentation will be in class, oral presentation, around 30 minutes

2 Abstract Definition of Graphical Models

Graphical model is characterized by the following three components:

- Graph: G is a graphical representation that can be used to describe the relationship between variables. In the context of graphical models, it characterizes the dependencies/correlations between random variables.
- Model: $\mathcal{M}_G$ provides a mathematical description of the graph. It is the basis for quantitative analysis of a graphical model.
- Data: $D \equiv \{X_1^{(i)}, X_2^{(i)}, ..., X_m^{(i)}\}_{i=1}^N$ are our observations of the variables in the graph. Specifically, $X_j^{(k)}$ corresponds to the k_{th} observation of random variable X_j.

Graphical model: $\mathcal{M}_G$, where G is a graph, connects data from hypothesis, instantiations from random variables, which easily encodes the dependencies among the random variables, especially in high dimensional problems with many random variables.

3 Fundamental Questions

3.1 Representation

- How to describe the problem in a proper way?
- How to encode domain knowledge in to a graph? (good/bad hypothesis)

3.2 Inference

- How to answer queries (especially about some hidden variables) using the model and the data?
- The queries are not deterministically, but probabilistic inference like asking probability $P(X|D)$.

3.3 Learning

- How to obtain the "best" model from the data? Normally, we will assume a tree structure of the graph, then somehow maximize a score function to get the parameters of the model such as

$$\mathcal{M} = \arg\min_{\mathcal{M} \in M} F(D; \mathcal{M})$$

4 A Heuristic Example of Graphical Model

Consider a scenario where we want to specify the probability distribution of 8 random variables. Without any prior knowledge, we can specify the joint pdf by a table with 2^8 row entries, and that is all we can do. This is not only expensive, but also almost impossible because we may not have past observations on all the 2^8 configurations.

However, sometimes we may have some prior knowledge about the relationship between the random variables. For instance, the 8 random variables may be states of parts of a biological system and we have prior knowledge about their relationships. e.g. X_1 and X_7 may be biologically separated that they cannot directly influence each other. These dependency relationships can be characterized by a directed graph where random variable X depends on Y if Y is its parent. Figure. 1 shows one such scenario.

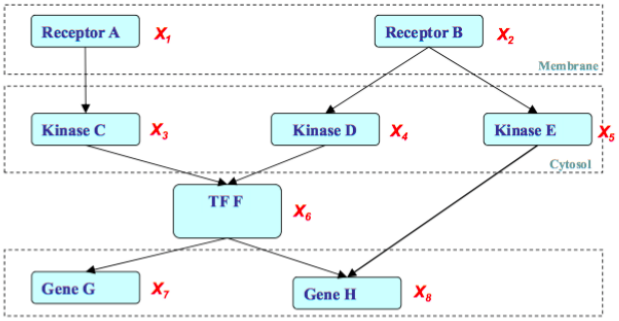

Figure 1: Dependencies among variables

Note that with the known dependencies and the graph, we can easily use the factorization rule to simplify our description of joint distribution. Specifically in this case,

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

For this joint distribution, we only need to specify probabilities for 18 cases, i.e. $P(X_1 = 0), P(X_2 = 0), P(X_3 = 0|X_1 = 0), P(X_3 = 0|X_1 = 1), \ldots,$ etc. This is number is significantly smaller than the 2^8 number of cases in a naive specification of joint distribution.

5 Why Having a Graph

From the example above, there are some clear advantages of using graphical models over a table.

5.1 Simpler joint distribution

As we have discussed above, even for only 8 binary random variables, the table for the joint distribution is huge (exponential in the number of random variables), and it would be way too expensive to do that for hundreds of random variables. Besides, using a table to represent the joint distribution does not provide any insight for the problem, nor does it use any domain knowledge in the representation. This is one of the reasons why people use graphical models to represent joint distributions.

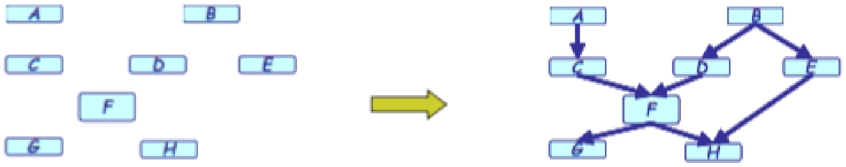

Figure 2: Graphical model leads to simpler joint distribution

5.2 Easier learning and inference

For a table representation of a joint distribution, it is often impossible to learn the probabilities of every outcome from the data, since many instantiations of the random variables don't even show up with limited amount of data, especially for some combinations with extremely low probability.

Then there is also a problem with doing inference, because if we have a giant table and we want to ask about some marginal distributions, we need to sum up a subset of all the rows in the table to compute the probability, which is expensive when the number of random variables is large.

5.3 Enables integration of domain knowledge

In some cases, we assume no prior knowledge of data. But incorporating domain knowledge makes us make more reasonable assumption and simplifies learning/inference. Graphical model provides a platform for inter-disciplinary communication and a simple solution to integrate domain knowledge.

5.4 Enables data integration

Sometimes data come from different sources or they are of different types. It is generally hard to combine these data in some reasonable way, but graphical makes it possible and easy by specifying dependencies and join the distributions together by taking product of their own distributions. For instance, we can combine text, image and network data into holistic social media data using graphical models.

5.5 Enables rational statistical inference

With graphical models, we can do Bayesian learning and inference by adding a prior to the model parameters, which will capture the uncertainty about the model itself. Computing the posterior distribution of the parameters and take the posterior mean will give us an Bayes estimator.

So informally, probabilistic graphical model is a smart way to write, specify, compose and design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with structured semantics. Formally speaking, it 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.

6 Types of Graphical Model

6.1 Bayesian Network (Directed Probabilistic Graphical Models)

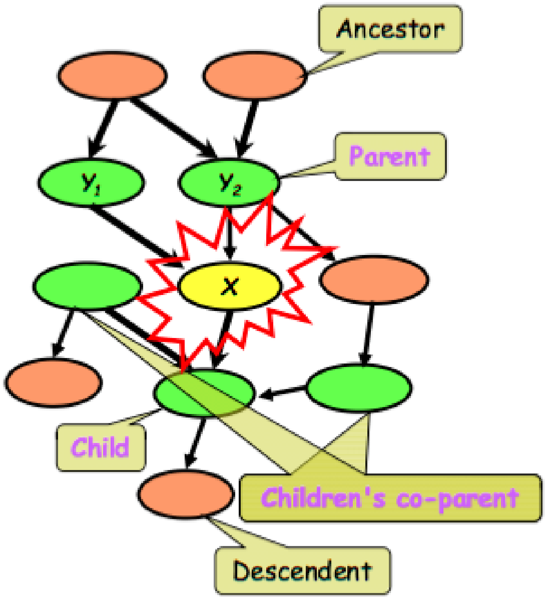

Figure 3: An example of Bayesian Network

The structure of a Bayesian Network is that of directed acyclic graph (DAG). Such representation describes causality relationships between variables and facilitates a generative process of the random variables in the graph. Figure 3 illustrates one such graph. One important observation is that X is conditionally independent of the orange random variables given the green random variables. The green random variables are called the Markov blanket of X. Note that X and Y_1, Y_2 should form a well defined conditional probability distribution such that $P(x|y_1, y_2) \geq 0$, $\forall x, y_1, y_2$ and $\sum_x P(x|y_1, y_2) = 1$.

To write out the joint distribution of random variables represented by a Bayesian network, we usually use factorization rule, e.g. for Figure 1

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

6.2 Markov Random Fields (Undirected Probabilistic Graphical Models)

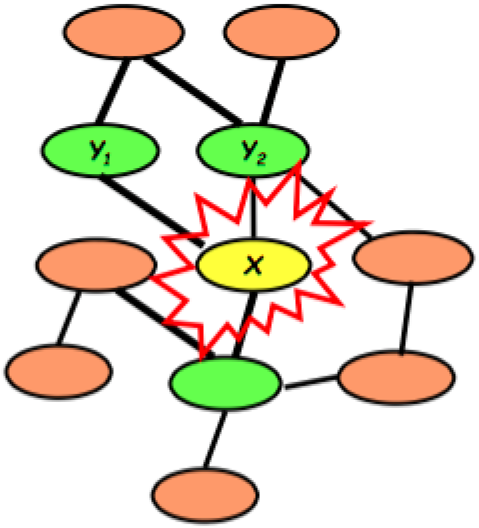

Figure 4: An example of Markov Random Field

The structure of a Markov Random Fields is that of undirected graph. Such representation describes correlations between variables, but not an explicit way to generate samples. Figure 4 illustrates one such graph. Here, a node (X)is conditionally independent of every other node (orange) given its directed neighbors (green). The joint distribution is completely determined by local contingency functions (potentials) and cliques. In contrast, the joint distribution of Bayesian network is determined by local conditional distributions and DAG.

The joint distribution is thus factored as, e.g. for Figure 1 (replace directed arrow with undirected lines)

$$\begin{aligned} P(X_1, X_2, ..., X_8) = \frac{1}{Z} exp\{&E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) \\ &+ E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\} \end{aligned}$$

6.3 Equivalence Theorem

For a graph G, Let $\mathcal{D}_1$ denote the family of all distributions that satisfy $I(G)$, let $\mathcal{D}_2$ denote the family of all distributions that factor according to G, then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

6.4 Relations to ML tasks

- Density estimation (parametric and nonparametric methods)

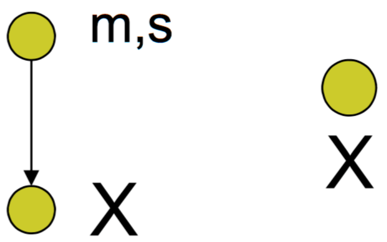

Figure 5: A graphical model representation of density estimation problem

X is the observed variable generated from an unknown distribution with parameters m and s. The goal is to estimate m and s from X.

- Regression (linear, conditional mixture, nonparametric)

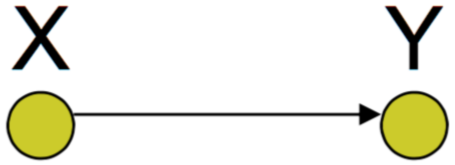

Figure 6: A graphical model representation of regression problem

X and Y are two random variables. The goal is to estimate Y from X.

- Classification (Generative and discriminative approach)

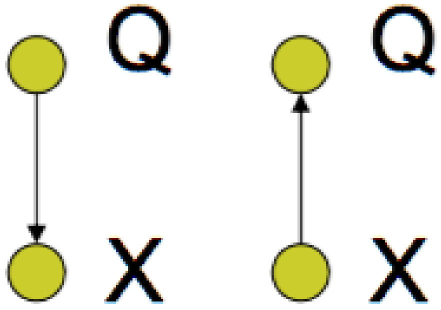

Figure 7: A graphical model representation of classification problem

The left of Figure 7 represents classification problem where the model is generative. Q is the class variable that determines the generation of X. The goal is to estimate Q from which X is generated. The right of Figure 7 represents classification problem where the model is discriminative. The goal is to assign each observation of X a class Q.

- Clustering

  Similar to classification, but we do not observe Q in training set.

7 Summary

Probabilistic graphical model is not a model but a language for communication, computation and development. It provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms. Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism. So the graphical model framework provides a way to view all of these systems as instances of a common underlying formalism.