CS281A/Stat241A: Statistical Learning Theory

chapter 2 to chapter 3

Lecturer: Michael I. Jordan *Scribes: Hao Zhang and Peng Zhao*

1 MAX-ELIMINATE AND MAX-SUM-OF-PRODUCT

We'd like to compute the maximum probability of X_F, conditioned on the shaded nodes X_E. In particular, this is:

$$\max_x P^E(X_F)$$

and we are also interested in the state of X_F that achieves the maximum conditional probability, i.e.,

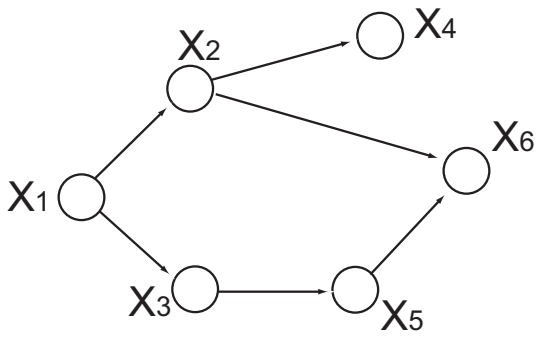
$$\text{argmax}_x P^E(X_F)$$

Recall the notion of "un-normalized" conditional $P^E(X_F) \equiv P(X_F, X_E)\delta(X_E, \bar{X_E})$. We know that it doesn't have to sum to 1 but is proportional to the conditional probability $P(X_F|\bar{X_E})$.

Referring to the canonical example in class, we see that

$$\begin{aligned} \max_x P^E(X) &= \max_{X_1} ... \max_{X_6} P(X_1)P(X_2|X_1)...P(X_6|X_2, X_5) \\ &= \max_{X_1} P(X_1) \max_{X_2} P(X_2|X_1)... \max_{X_6} P(X_6|X_2, X_5) \end{aligned}$$

which resembles the form we encountered in inference, with summation replaced by max. Analogously, we can define intermediate terms as $m_6^*(x_2, x_5)$ and $m_5^*(x_2, x_3)$, etc.



The underlying properties that allowed us to draw the analog is the distributive law, common to both the summation and the max operator, that is, $ac + bc = (a + b)c$ and $\max(ac, bc) = \max(a, b)c$. These algebraic structures are called "semi-rings". And there's a version of ELIMINATE (on arbitrary graph) for any semi-ring. Or, alternatively, one can easily generalize what property those operators must have for the ELIMINATE algorithm to carry through.

On trees, a version of SUM-OF-PRODUCT algorithm also exists for the max operator.

The next step is to compute the argmax. Here,the situation is bit more complicated, depending on whether the argmax is unique.

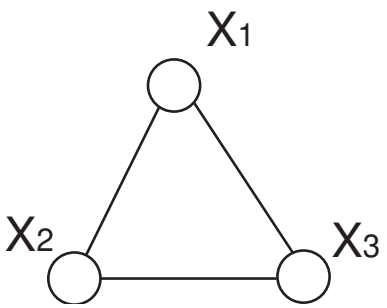
CASE 1: Think of the conditional probabiliy $P^E(X)$ being a table of values. If the maximum is unique, and we know that

$$P^*(X_1^*) = \max_{X \setminus X_i} P^E(X)$$

achieves the maximum. Then the choice of X_1 at the maximum has to be X_1^* by uniqueness of the maximum.

CASE 2: If there are multiple maximum points, then the problem of finding the maximizing configuration is harder. For example, consider a graphical model of three nodes forming a triangle, each of the node can be in state 0 or 1, and the potentials are

$$\phi(X_i) = [.5; .5] \qquad \phi(X_i, X_j) = (1 \quad .2; .2 \quad 1)$$



This graphical model effectively favors the three nodes to be in the same state. Hence the most probable state configurations are: (0,0,0) and (1,1,1). However, when MAX-ELIMINATE is applied until only one node is left, it may as well be in a 0 state or a 1 state to yield the same maximum value. This ambiguity can be resolved. The added complication is to keep track of the argmax along the way of elimination and trace back the state once the state on the final node is fixed.

Such an approach resembles dynamic programming, and it can be said that optimization is harder than inference, which only involves sum of products.

2 Statistical side of graphical models

The key question is how to define and find parameters.

Parametrization: in the directed graph case,

$$P(X|\theta) = \prod_{i=1}^{n} P(X_i|X_{\pi_i}, \theta_i)$$

where $\theta = (\theta_1, ..., \theta_n)$. Note that the distributions $P(\cdot)$'s are no longer arbitrary but are indexed (parameterized) by θ.

In the undirected graph case,

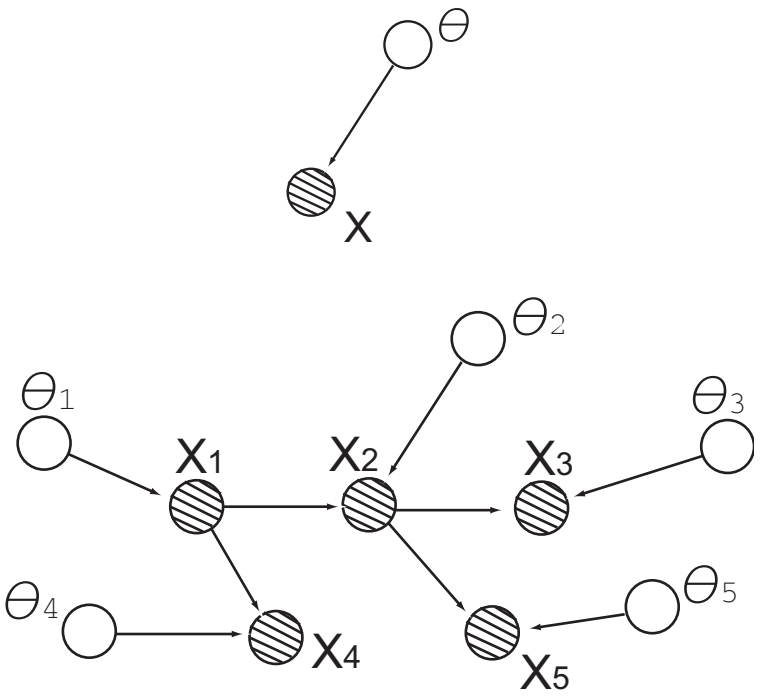
$$P(X|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \phi_c(X_c|\theta_c)$$

Given a model $P(X|\theta)$, the central question is to find an estimate of θ. In a Bayesian setting, θ is considered to be a random variable in the model.

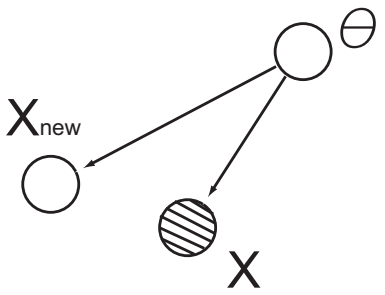
A key fact is Bayes rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

where $P(\theta|X)$ is called the posterior, $P(X|\theta)$ is the likelihood, and $P(\theta)$ is the prior.



In the task of prediction, we'd like to predict a new random variable X_{new} based on the previous observations X. A Baysian approach is the model the data as coming from a distribution parameterized by θ.



Then,

$$P(X_{new}|X) \propto P(X_{new}, X) = \int_{\theta} P(X_{new}, X, \theta)$$

The joint distribution factorizes as

$$P(X_{new}, X, \theta) = P(X_{new}|\theta)P(X|\theta)P(\theta)$$

and we can reach

$$P(X_{new}|X) = \int_{\theta} P(\theta|X)P(X_{new}|\theta)$$

The $P(\theta|X)$ term can be thought of as "voting weights" and the $P(X_{new}|\theta)$ term can be thought of "voters" for all possible values of X_{new}.