

Boring Title: Undirected Graphical Models, Exciting Title: Markov Random Fields

1 Independence properties are simpler than in DAGS

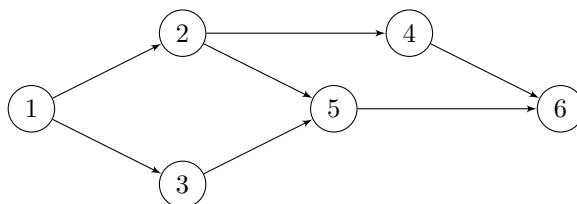
- **global Markov property**: for node sets A, B, C in graph G , $\mathbf{x}_A \perp \mathbf{x}_B \mid \mathbf{x}_C$ iff removing all nodes in C leaves no path connecting any node in A to any node in B
- **undirected local Markov property**: given its set of immediate neighbors (i.e. its **Markov blanket**), a node t is conditionally independent of all other nodes
- **pairwise Markov property**: two nodes conditionally independent given the rest if no direct edge between them
- global implies pairwise and vice versa

Example 1. Markov blanket for UGMs is defined above. What does the Markov blanket in DGMs include?

2 DGM to UGM Conversion

For each pair of parent nodes that share a child, marry them by adding an edge. Then drop the direction of all edges. Finally, use the CI properties of UGMs to answer a CI query more easily than in a DGM. At worst, we lose a few CI assumptions. However, we should not introduce new ones.

Example 2. Convert the following DGM to UGM:



3 Distributions and Graphs

- graph G is an **I-map** of distribution p if $I(G) \subseteq I(p)$ where $I(\cdot)$ is “the set of conditional independencies encoded by...”. **Question:** is a naive bayes DGM with extra connections among the x ’s an I-map for $p(x, y)$ with the typical naive Bayes assumptions?
- G is a **perfect map** of p if $I(G) = I(p)$
- DGMs, UGMs perfect maps for different (but overlapping) sets of distributions
- Graphs only specify the conditional independence structure, nothing about the parameters, distributions.
- Shape of graph determines difficulty of inference

4 Parameterization of UGMs

Each edge in a DGM represents a normalized conditional probability distribution. What do undirected edges represent?

- A **clique** is a set of nodes such that each node has an edge with all others in the set. A **maximal clique** is the clique of the largest size that a node belongs to.
- Each clique has a **potential function** that assigns each possible configuration a score. Considering just 8 nodes with binary values, this is 2^8 states to score for just the one clique.
- We use **log-potentials** $\theta_c(x_c)$. The book uses potentials $\psi_c(x_c) = \exp[\theta_c(x_c)]$
- joint distribution over \mathbf{x} 's proportional to product over all potentials

$$p(x_1, \dots, x_T) = \exp \left[\sum_c \theta_c(x_c) - A(\theta) \right] \\ \propto \prod_c \exp [\theta_c(x_c)] = \prod_c \psi_c(x_c)$$

Very important: UGMs are **globally normalized** ($A(\theta)$ term) instead of **locally normalized** like each edge of a DGM. More about this below in CRFs.

Example 3. Draw the UGM corresponding to the Naive Bayes DGM and describe how you would form the potential functions

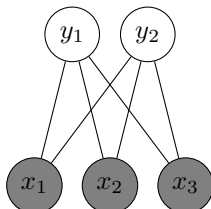
5 Ising model

Example 4. Let X_s be the Bernoulli random variable associated with node s on a graph $G = (V, E)$ with $|V| = m$. X_s takes on the **spin** value $\{-1, +1\}$. This could correspond to magnet orientation, pixel value, and many other things. X_s and X_t interact if they have a direct edge in the graph (we expect a certain relationship between neighboring pixels). Let $\theta_{st} \in \mathbb{R}$ be the strength of the edge (s, t) and θ_s be the marginal potential for node s (determined by an external field). Write down the joint probability of the set of \mathbf{x} in terms of potential functions that correspond to the model above. It should look like an exponential family. How do you calculate $A(\theta)$? What is the distribution's dimension?

6 Conditional Random Fields

- Clique potentials functions over hidden configurations y conditioned on observed features x
- Can be thought of as a **structured output** form of logistic regression
- CRF advantage over MRF like logistic regression vs. naive Bayes, i.e. don't need to model things that we always observe
- **x are not necessarily independent**

An example with pairwise cliques:



See a time series CRF at end of notes.

7 Stereo Depth Reconstruction CRF

Example 5. Given two images \mathbf{x}_L and \mathbf{x}_R taken at a small known angle difference, recover the depth measurement of each pixel by estimating the position of pixel $\mathbf{x}_L(i, j)$ in \mathbf{x}_R , which will have a different horizontal value and the same vertical value. This difference, call it y_s , is the **disparity** at i, j . By computing each such disparity, a disparity map shows an estimate of depth. Write down a suitable potential function $\psi_s(y_s, \mathbf{x})$. How could you define a $\psi_{st}(y_s, y_t)$ to enforce similar disparity in neighboring pixels?

8 Time series models

Time-dependent patterns. Be it a generative or discriminative model, we may be interested in questions like: how likely is this (underlying) sequence given another (observed) one. Some examples discussed in class: assigning discrete phoneme labels to continuous speech signals, assigning a continuous label (inferred trajectory of airplane) to noisy continuous signals obtained from radar.

9 Example in detail: Hidden Markov Model

Assume that the underlying process is a discrete Markov process. Each state is connected to the data via an **emission model**. There are K discrete states and the transition probability between them is given by the matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$, which fully parametrize the categorical dist. for y_{t+1} .

Example 6. Memorylessness An important hallmark of a Markov process is that the next state only depends on the current state, and not the previous states. This property, called "memorylessness", is formalized by the Chapman-Kolmogorov equation. If we denote the state of the system at time t y_t , then

$$p(y_{t+1}|y_t, y_{t-1}, \dots, y_1) = p(y_{t+1}|y_t).$$

Is this property illustrated in the UGM representation?

9.1 Parameters of the HMM

Given data \mathbf{X} , the likelihood of a state sequence \mathbf{y} is given by:

$$P(\mathbf{X}, \mathbf{y}) = p(y_1) \prod_{t=2}^T p(y_t | y_{t-1}) \prod_{t=1}^T p(\mathbf{x}_t | y_t).$$

Note that the three components are:

- $p(y_1)$ the initial probabilities. Usually it is parameterized by a K -sized vector π .
- $p(y_t | y_{t-1})$ the transition probabilities. It is parameterized by the transition matrix \mathbf{A} .
- $p(\mathbf{x}_t | y_t)$ are the emission probabilities. In a typical setting, for each y_t there is a multivariate normal (MVN) distribution parameterized by $\{\Sigma_k, \mu_k\}$.
- Stay within exponential family with categorical and Gaussian distributions

Maximum likelihood estimate (MLE) for HMMs is difficult because the joint distribution does not factorize over T (why?). In fact, if there are K states and T data points, we will have to consider all K^T paths to do maximum likelihood exactly. More on this in a few weeks.

10 HMMs with continuous state space: Kalman filtering

In an HMM, we considered the scenario where continuous observations are generated by a process transitioning in a discrete state space. Expanding on HMMs, **Kalman filtering** considers continuous observations that are generated by continuously changing processes. Specifically, we assume to be getting noise sensor data $\mathbf{X} \in \mathbb{R}^D$ that comes from a source $\mathbf{Y} \in \mathbb{R}^D$. We can use similar notations from our discussion of HMMs. The components are:

- $p(\mathbf{y}_t | \mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t | \mathbf{A}\mathbf{y}_{t-1}, \Gamma)$ are analogous to transition probabilities in HMMs. Alternatively, we can write this as a linear equation (this model is sometimes called a **linear dynamical system**).

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{w}_t,$$

where \mathbf{w}_t is a zero-mean Gaussian noise term.

- $p(\mathbf{x}_t | \mathbf{y}_t) = \mathcal{N}(\mathbf{x}_t | \mathbf{C}\mathbf{y}_t, \Sigma)$ are the emission probabilities. It can be similarly written in linear form:

$$\mathbf{x}_t = \mathbf{C}\mathbf{y}_t + \mathbf{v}_t.$$

Since all the conditional probabilities are Gaussian, and linear transformations of Gaussian distributions are still Gaussians, Kalman filters are relatively easy to analyze.

11 Maximum-entropy Markov Models (MEMMs) and CRFs

Consider a directed MEMM where $p(y|x, w) = \exp [\log p(y_1) + \log p(y_1 | x_1) + \sum_{t=2}^T \log p(y_t | y_{t-1}, x_t, w)]$ is the probability of a sequence of labels given some data. MEMMs and the other time series models above are so-called **locally normalized** models. Each conditional probability distribution, such as $p(y_t | y_{t-1})$, sums to one. In a **globally normalized** models, each undirected edge can represent an un-normalized potential function. We can compute a product over clique potentials using a variety of functions, and then use a normalizing term.

$$p(y|x, w) = \exp \left[\sum_t \theta^{obs}(y_t, x_t) + \sum_{t=2}^T \theta^{trans}(y_t, y_{t-1}) - A(\theta) \right]$$

Example 7. How do you calculate $A(\theta)$?

Example 8. In practical applications, we sometimes need signals that come later to disambiguate previous signals. I.e., we need an information flow from \mathbf{x}_t to \mathbf{y}_{t-1} . Analyze the time series models that we have discussed. Is this possible in all the models?