EECS 445: Machine Learning **Fall 2016**

Discussion 7: Probabilistic Graphical Models

Written by: Benjamin R. Bray

This week in lecture, we introduced **Probabilistic Graphical Models**, which are a very elegant and powerful way to model the *independence structure* of complex data. So far, we have covered only simple probabilistic models like linear regression, logistic regression, and Naive Bayes. Graphical models provide us with a language for expressing more complex probabilistic models. Here are a few advantages of using graphical models:

- **Transparency.** All modeling assumptions are explicitly encoded in the graph structure of a probabilistic model and the accompanying distributions.
- **Interpretability.** compared to i.e. neural networks or random forest models, whose learned parameters are hard to interpret. In addition, because graphical models are *probabilistic*, we can easily determine confidence intervals for our estimates, which matters when machine learning algorithms are used to make important real-world decisions (diagnosis, trading, self-driving cars, etc.).
- **Modularity.** Graphical models can be easily combined and modified, and it is easy to adapt old models to new applications. Sometimes it's even useful to "nest" one graphical model inside another!

Fair Warning: We'll be using probability extensively in this section of the course, and it's important that you have a good grasp of the basics! Conditional independence and Bayes' rule will be especially important. Check Piazza for a list of concepts we expect you to know.

Bayesian networks, which we define below, can be interpreted as a compact way of encoding conditional independence assumptions. First, recall the concept of independence:

Definition 7.1. *We say events or variables X and Y are **independent**, written $(X \perp Y)$, if*

$$P(X \mid Y) = P(X) \quad or \quad P(Y) = 0$$

While independence is a useful property, it is not often that we encounter two independent events in practice. More commonly, two events are independent only if we have information about a third event.

Definition 7.2. *We say events or variables X and Y are **conditionally independent** given Z, which we write as $(X \perp Y \mid Z)$, if*

$$P(X|Y,Z) = P(X|Z)$$

Example 7.1. *(Koller, Probabilistic Graphical Models) For example, suppose we want to reason about the chance that a student is accepted to graduate studies at Stanford or MIT. Denote by Stanford the event "admitted to Stanford" and by MIT the event "admitted to MIT." In most reasonable distributions, these two events are not independent. If we learn that a student was admitted to Stanford, then our estimate of her probability of bveing accepted at MIT is now higher, since it is a sign that she is a promising student.*

Now, suppose both universities base their decisions only on the student's grade point average, and we know that our student has a GPA of A. In this case, we might argue that learning that the student was admitted to Stanford should not change the probability that she will be admitted to MIT: her GPA already encodes all relevant information. Formally, we have

$$P(MIT \mid Stanford, GradeA) = P(MIT \mid GradeA)$$

In this case, we say MIT is conditionally independent of Stanford given GradeA.

Excercise 7.1. *Let X, Y, Z, W be events or random variables. Show that conditional independence satisfies the following properties:*

- ***Symmetry:*** $(X \perp Y \mid Z) \implies (Y \perp X \mid Z)$
- ***Decomposition:*** $(X \perp Y, W \mid Z) \implies (X \perp Y \mid Z)$
- ***Weak Union:*** $(X \perp Y, W \mid Z) \implies (X \perp Y \mid Z, W)$
- ***Contraction:*** $(X \perp W \mid Z, Y) \wedge (X \perp Y \mid Z) \implies (X \perp Y, W \mid Z)$

7.1 Bayesian Networks

A Bayesian network is one type of graphical model that encodes a set of independence assumptions among a set of random variables with a graph. At a high level, the structure of a Bayesian network tells us which relationships between variables we believe to be most important for solving the problem at hand. Here is the formal definition of a Bayesian network from lecture:

Definition 7.3. *A **Bayesian network** $\mathcal{G}$ is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$. Let $\text{Parents}(X_k)_{\mathcal{G}}$ denote the parents of X_k in $\mathcal{G}$ and $\text{NonDesc}(X_k)_{\mathcal{G}}$ denote the variables in the graph that are not descendants of X_k. Then $\mathcal{G}$ encodes the following set of conditional independence assumptions, called the **local independencies**, and denoted by $\mathcal{I}_\ell(\mathcal{G})$.*

$$\textit{For each variable } X_k : (X_k \perp \text{NonDesc}_{\mathcal{G}}(X_k) \mid \text{Parents}_{\mathcal{G}}(X_k))$$

This definition looks scary and technical, but over the next few weeks we will work with many examples and working with these graphs will become second-nature. For now, whenever you get confused, remember this one important fact:

> Every variable $X_k \in \mathcal{G}$ is conditionally independent of its nondescendants, given its parents.

Excercise 7.2. *Why must we require that the graphs be acyclic? What happens if there are cycles?*

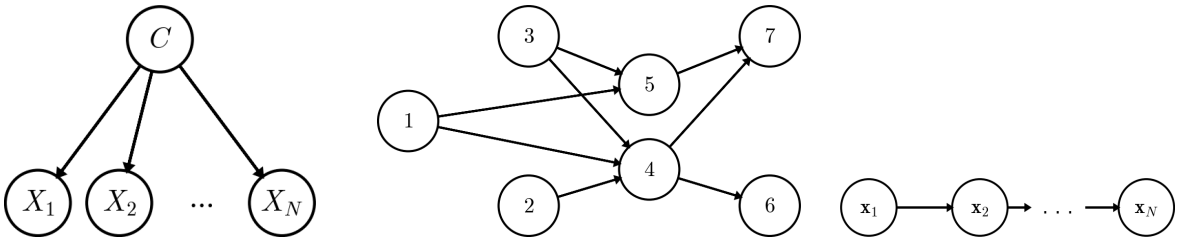
Excercise 7.3. *Identify the parents and nondescendants for each node in every graph in Figure 7.1.*

7.1.1 Example: Naive Bayes

Naive Bayes is the simplest nontrivial example of a graphical model. The corresponding Bayesian network, shown in Figure 7.1a, was briefly discussed in lecture, but let's go through it a little more carefully.

Excercise 7.4. *List all local independencies $\mathcal{I}_{\mathcal{G}}$ encoded by the graph in Figure 7.1a.*

Example 7.2. *Show directly that the independence assumptions made by Bayesian network in Figure 7.1a correspond exactly the Naive Bayes assumption.*



(a) Naive Bayes (b) Topological Sort (c) Markov Chain

Figure 7.1: Three simple Bayesian networks.

7.2 From Networks to Factorizations

Theorem 7.4 (Factorization)**.** *Suppose the network $\mathcal{G}$ is an I-map for P, that is, the distribution P satisfies the independence assumptions contained in $\mathcal{G}$. $X_1, \ldots, X_n$ is any **topological sorting** of the nodes of $\mathcal{G}$, then P **factorizes** according to $\mathcal{G}$ in the following way:*

$$P(X_1, \ldots, X_n) = \prod_{k=1}^{n} P(X_k \mid \text{Parents}_{\mathcal{G}}(X_k))$$

Proof: Once we have topologically sorted nodes $X_1, \ldots, X_n$, the chain rule for probability yields

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n|X_1, \ldots, X_{n-1}) = \prod_{k=1}^{n} P(X_k \mid X_1, \ldots, X_{k-1})$$

The nodes are topologically sorted, so all parents of X_k are in the range $X_1, \ldots, X_{k-1}$, and the rest of these nodes must be nondescendants. Because X_k is conditionally independent of its nondescendants given its parents, we can delete the nondescendants frmo this list when conditioning. Therefore, $P(X_k \mid X_1, \ldots, X_{k-1}) = P(X_k \mid \text{Parents}_{\mathcal{G}}(X_k))$. For a more rigorous proof, see the lecture notes. ■

Example 7.3. *After establishing the factorization theorem, it is clear that Figure 7.1a is indeed Naive Bayes.*

The factorization theorem is incredibly useful for building **generative models**, which describe our assumptions about how the observed data is generated. Using the factorization theorem, we can interpret any Bayesian network as a generative model.

7.2.1 Example: Markov Chains

A **Markov chain** is a simple probabilistic model used for describing sequences, usually discretized in either time or space. The corresponding graphical model is shown in Figure 7.1c. Markov chains are often called *local* or *memoryless* models because if the value at position k in the chain is known, then past values do not influence our predictions of future values. While not entirely true in general, this assumption can simplify inference in complex settings like weather forecasing, finance, and natural language processing. Usually, we assume all variables in a Markov chain have the same type.

Excercise 7.5. *List all local independence assumptions made by the Markov Chain model in Figure 7.1c.*

Excercise 7.6. *By the factorization theorem, the joint distribution for a Markov chain factorizes as*

$$P(X_1, \ldots, X_n) = \prod_{k=1}^{n} P(X_k \mid \text{Parents}_{\mathcal{G}}(X_k)) = \prod_{k=1}^{n} P(X_k \mid X_{k-1})$$

Using this factorization, interpret Markov Chains as a generative model. It may help to think of the random variables $X_1, \ldots, X_n$ as words in a sentence, a robot's position at time t, or the weather forecast $\{Rainy, Sunny\}$ on day d. Draw a transition diagram for the model.

7.3 From Factorizations to Networks

If we know the independence structure of a joint distrubution over some variables, we can write down a corresponding Bayesian network that encodes these independencies. Remember that there are multiple valid answers! The complete graph on n nodes is a valid Bayesian network for any variables $X_1, \ldots, X_n$.

Excercise 7.7. *Draw a Bayesian network for each factored joint distribution below.*

- $P(A, B, C, D)$
- $P(A, B, C) = P(A)P(B|A)P(C|B)$
- $P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D)P(E|C, D)$

We can also be more concrete! Suppose we are given a generative model, complete with distributions that tell us how to generate data. We can use this information to come up with a Bayesian network that fits the independence assumptions of the generative model.

Example 7.4. *TODO*