CS281 Section #4 Notes, v 1.1

1 Neural networks

1.1 Exercises

- See the PyTorch demo here: [walkthrough]

2 Directed Graphical Models

2.1 Review

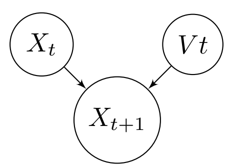
A *directed graphical model* G over V random variables is a way to factorize a probability distribution $p(x_{1:V})$. They are also called bayes nets or probabilistic graphical models. In addition to giving the way the joint distribution factorizes, they also encode the *conditional independence* structure of the variables, which we'll explain below.

$$p(x_{1:V}|G) = \prod_{i=1}^{V} p(x_i|x_{\text{parents}(i)})$$

Gaussian Bayes Net

In class, we worked with the following graphical model: suppose we have a node x_i and its parent nodes $x_{\text{parents(i)}}$ representing continuous variables. Let the parent variables be distributed as Gaussians. Here, we choose to model x_i as a linear function of its parents with Gaussian noise: $p(x_i \mid x_{\text{parents(i)}}) = \mathcal{N}(x_i; \text{linear}(x_{\text{parents(i)}}), \sigma_i^2)$

Example 1. X is position. $X(t+1) \sim X(t) + X(t)dt + \epsilon$



Formally,

$$p(x_i \mid x_{pa(i)}) = \mathcal{N}(x_i; \mu_i + w_i^T x_{pa(i)}, \sigma_i^2)$$

is a linear Gaussian conditional probability distribution (CPD).

In lecture, we show that multiplying all these CPDs results in a jointly Gaussian distribution,

$$\prod_{i=1}^{V} p(x_i \mid x_{pa(i)}) = p(X) = \mathcal{N}(X; \mu, \Sigma)$$

It's straightforward that $X = [\mu_1, ..., \mu_V]$. What remains is to find the covariance matrix. For review from lecture, this is how we do it: for simplicity, rewrite the conditional probability distribution in the following way:

$$x_i = \mu_i + \sum_{j \in x_{pa(i)}} w_{ij}(x_j - \mu_j) + \sigma_i z_i$$

$$z_i \sim \mathcal{N}(0, 1) \quad \text{(Gaussian random noise)}$$

Note that w_{ij} can be 0 if x_j is not a parent of x_i. For the "root" nodes, with no parents, all the coefficients are 0. Let $S = \text{diag}(\sigma)$, rewriting this again in matrix-vector form:

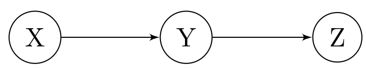
$$\begin{aligned}
\mathbf{x} - \mu &= W(\mathbf{x} - \mu) + S\mathbf{z} \\
\mathbf{x} - \mu - W(\mathbf{x} - \mu) &= S\mathbf{z} \\
(x - \mu)(I - W) &= S\mathbf{z} \\
\mathbf{x} - \mu &= (I - W)^{-1} S\mathbf{z}
\end{aligned}$$

$$\mathbf{\Sigma} = \text{Cov}[\mathbf{x} - \mu] = \text{Cov}[(I - W)^{-1} S\mathbf{z}] = (I - W)^{-1} S \,\text{Cov}[\mathbf{z}] S (I - W)^{-1T}$$

which implies that the variance is $(I - W)^{-1} S^2 (I - W)^{-1T}$.

Conditional Independence Statements

Consider a joint distribution over discrete random variables X, Y, Z, with each discrete R.V. taking 10 possible values. Without knowing anything else about X, Y, Z, in order to represent the joint distribution $P(X, Y, Z)$ in a table, we would need to store $10^3 - 1 = 999$ values (since we know their sum is 1). But suppose we know we have a dependence among X, Y, Z that looks like the following DAG:



Then the joint distribution simplifies to $P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$. $P(X)$ requires storing 9 values, $P(Y|X)$ requires 90 values (9 probabilities for each possible value of X), as does $P(Z|Y)$, for a total of 189 values necessary to specify the joint distribution. If we're trying to estimate the distribution by e.g. counting the number of occurrences of each of the variables, this means that we split our data into 189 pieces rather than 999, which can decrease error. The same idea holds for continuous distributions.

D-Separation

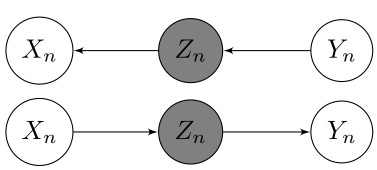
Conditional independencies can disappear or appear with new knowledge (when a random variable is observed). Z is said to *d-separate* X and Y if information about Z renders X and Y conditionally independent: $X \perp Y \mid Z$. In general, we can also talk about the *d-separatedness* of two random variables, given a collection.

Formally: Let X, Y, Z be disjoint subsets of nodes in a DAG G. A path between X and Y is given by a sequence of edges that connects a node in X to a node in Y (directionality doesn't matter). We say a path from x_i to y_i is *active* if dependencies flow from one end to another (information about variable x_i influences our belief about variable y_i and vice versa). If a path from x_i to y_i is not active, it is said to be *blocked*, so that observing information about x_i does not affect our beliefs about y_i.

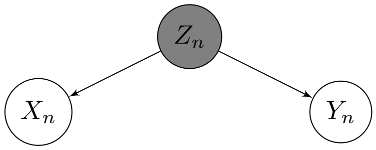
Then $X \perp Y \mid Z$ if every path between X and Y is *blocked*, and we say that X and Y are d-separated given Z.

There are broadly two ways a path can be blocked:

- Nonconverging arrows on a node in Z.
  $\exists\{\rightarrow C \rightarrow\}$ or $\{\leftarrow C \leftarrow\}$ or $\{\leftarrow C \rightarrow\}, C \in Z$

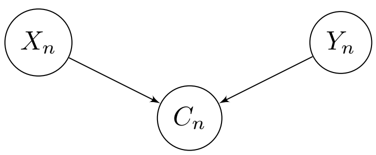


Example: X_n, Y_n, Z_n represents mitochondrial DNA (passed down only from the mother) from grandmother, mother, and an individual. Observing mother's mitochondrial DNA renders X_n and Y_n independent.



Example: Let Z_n be parents' genotype, and let X_n, Y_n be genotypes of their 2 children. When the parents' genotype is given, the children's genotypes are rendered independent: knowing information about X_n does not affect Y_n since we have already observed Z_n. Note that if Z_n was unobserved, then indeed X_n and Y_n are no longer necessarily conditionally independent: without knowing the parents' genotype, information about X_n *would* influence our beliefs about her sibling Y_n.

- Converging arrows on a node $\notin Z$
  $\exists\{\rightarrow C \leftarrow\}, C \notin Z$

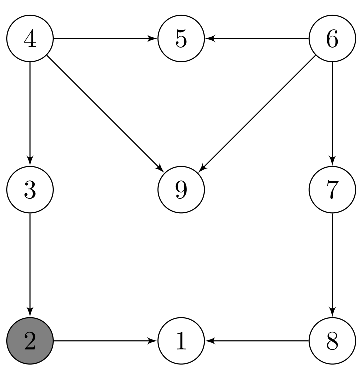


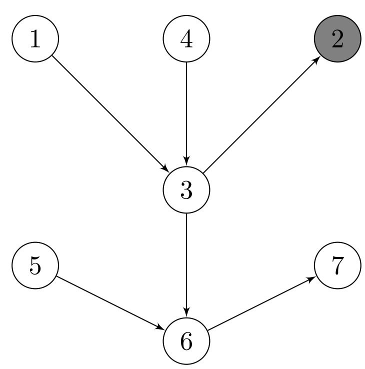
Example: Let X_n, Y_n be the outcomes from 2 fair dice rolls, and let C_n be the sum. A priori, the fair dice rolls X_n, Y_n are indeed independent. However, if C_n, the sum of the two rolls, becomes known, X_n and Y_n are no longer independent, since information about X_n would influence our beliefs about Y_n.

A node like C_n is called a V-node. If a V-node or a child of a V-node is observed, this will induce a conditional dependency on the parent nodes.

2.2 Exercises

1. For the following graphs, determine the largest set $X_{i:j}$ such that $X_1 \perp X_{i:j} \mid X_2$

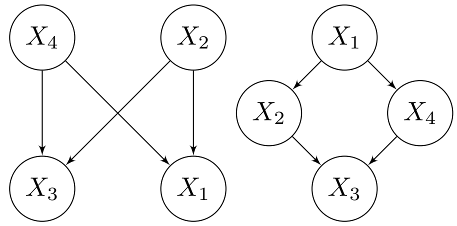




2. Suppose there is a distribution across $X_{1:4}$ such that the only independencies are: $X_1 \perp X_3 \mid \{X_2, X_4\}$ and $X_2 \perp X_4 \mid \{X_1, X_3\}$.

   Example for context: It was just NYFW (New York Fashion Week). Let X_i be the color of model i's shirt. Model 1 and 2 are friends. Model 2 and 3 are friends. Model 3 and 4 are friends. Model 1 and 4 are friends. For sartorial reasons, no model friends can have the same color shirt, lest they be photographed together in the street.

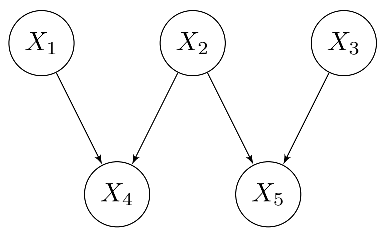
Can you represent this distribution as a DAG?



Both of these DAGS $X_1 \perp X_3 \,|\, \{X_2, X_4\}$ but $X_2 \not\perp X_4 \,|\, \{X_1, X_3\}$.

3. Let X_1, X_2, X_3 represent the outcomes of 3 independent binary RV's. Let $X_4 = 1\{X_1 = X_2\}, X_5 = 1\{X_2 = X_3\}$.

   (a) Draw the DAG



   (b) Under what circumstance is $X_4 \perp X_5$, if any? (Note $X_4 \perp X_5$ is not implied by this GM)