Session 9

I. Announcements [5 minutes]

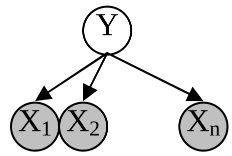
- **Homework 4 is online and is due November 4th**
  - Get started early and get ahead of the game.
- Exam statistics:

```
Number of grades reported:    111
Mean:                        60.1
Standard deviation:          14.9
Minimum:                     17.0
1st quartile:                50.5
2nd quartile (median):       63.0
3rd quartile:                69.0
Maximum:                     90.0
Max possible:               100.0
Distribution:
  0.0  -  5.0:             0
  5.0  - 10.0:             0
 10.0  - 15.0:             0
 15.0  - 20.0:             1     *
 20.0  - 25.0:             1     *
 25.0  - 30.0:             0
 30.0  - 35.0:             3     ***
 35.0  - 40.0:             2     **
 40.0  - 45.0:             5     *****
 45.0  - 50.0:            14     *************
 50.0  - 55.0:            12     ***********
 55.0  - 60.0:            16     ***************
 60.0  - 65.0:            10     **********
 65.0  - 70.0:            22     ********************
 70.0  - 75.0:            11     **********
 75.0  - 80.0:            10     **********
 80.0  - 85.0:             3     ***
 85.0  - 90.0:             0
 90.0  - 95.0:             1     *
```

II. Introduction to Bayes Nets

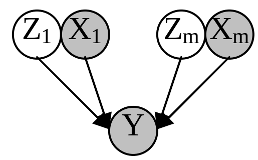
Structure of Bayes Nets

- The structure of a network contains the essential information about the conditional independence of the random variables.
- There are many reoccurring structures that capture common assumptions.
    - <u>Naïve Bayes Model</u>



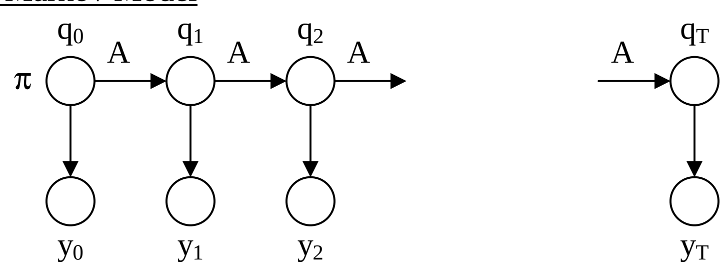
(a) conditionally independent features

    - <u>Noisy Or Model</u>



$$Y = \begin{cases} 1 & \bigvee_{j=1}^{m} \left(X_j \wedge \neg Z_j \right) \\ 0 & \mathit{otherwise} \end{cases}$$

    - <u>Hidden Markov Model</u>



- These models are very important in a branch of AI known as Statistical Machine Learning where we try to learn their parameters from observations of real-world phenomenon we assume follow a given model.
    - Inconsistencies between the exact model are often secondary to the effects captured in the structure of the model.
    - Independence assumptions often don't hold in the real world, but the models still perform well due to the approximate independence exhibited.

Foundations

- ***Conditional Independence*** **–** implies that two variables X,Y are independent given variable Z:

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z) \qquad P(X \mid Y,Z) = P(X \mid Z)$$

- **Bayes' Rule –** application of product rule that allows diagnostic beliefs to be derived from casual beliefs:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \qquad P(Y \mid X,e) = \frac{P(X \mid Y,e)P(Y \mid e)}{P(X \mid e)}$$

<u>DRUNK DRIVING EXAMPLE</u>

- **Naïve Bayes Model –** a single cause Y directly influences a number of events X_i that are all conditionally independent given the cause:

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y)\prod_i P(X_i \mid Y)$$

  - Often works in situations where conditional independence does not hold.
  - **SPAM FILETER**

Chain Rule of Probability Theory – In general,

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

Graphical Model – represents the joint probability distribution over a set of random variables via the independence relationships between those variables, thus concisely encapsulating a family of probability of distributions that respect those independence assumptions.

- <u>Nodes</u> – correspond in a 1-1 relationship with the variables in the distribution.
- <u>Edges</u> – represent dependence between a pair of random variables. The interpretation of this dependence depends on whether or not the graph is directed.

Directed Graphical Models – A Directed Acyclic Graph that represents the joint probability over a set of random variables. The directed structure can be interpreted as causality in constructing the models, although some philosophical thought brings this interpretation into dispute. Directed Graphical Models have a structure that represents the conditional independence assumptions made in the model.

In a DAGM, the joint probability distribution can be defined as

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid X_{\pi_i}\right)$$

where π_i is the set of parent nodes of the node X_i.

> **PROOF**: since it's a DAG, we can do a constructive proof over the topological ordering using the chain rule.
>
> > topological ordering – an ordering I of the variables in a DAG such that all ancestors of node i appear before i in the ordering.
>
> For a DAG, we can always order the nodes topologically; without loss of generality assume the following is topological: $X_1, X_2, \ldots, X_n$
>
> By the chain rule: $p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, X_2, \ldots, X_{i-1})$
>
> But, for any $p(X_i \mid X_1, X_2, \ldots, X_{i-1})$, we are conditioning on all of X_i's ancestors, which is equivalent to only conditioning on it's parents.
>
> $$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid X_{\pi_i}\right) \qquad \textbf{QED}$$

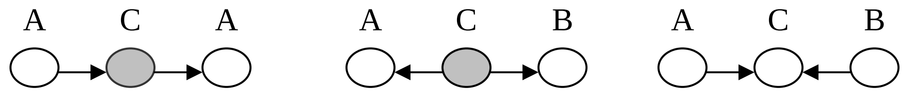
- Variables missing in the local conditional probability functions given by the chain rule over a topological ordering of the variables correspond exactly to the missing edges in the underlying graph. Thus, in defining the local functions of a variable, one is defining the probability of that variable conditioned on its parents.
- Let a_i be the ancestors of node i. The following is true of any DAG:

  $$X_i \perp\!\!\!\perp X_{a_i \setminus \pi_i} \mid X_{\pi_i}$$

  That is, given the parents of a node, that node is independent of all earlier nodes in a topological ordering. More generally, it can be shown that given the parents of a node, that node is independent of all nodes not connected to its descendant nodes in the DAG.
- Conditional Independence corresponds to the notion of d-separation in a directed graph. Slightly different than what you are accustomed to in graph separability.
- A node is conditionally independent of all other nodes in network given its *Markov Blanket* (parents, children, and children's parents).

d-separation – two nodes X and Y in a directed graph are d-separated if every path between X and Y is blocked.

- A path between X and Y is blocked if it has any of the following 3 cases for any 3 nodes along the path.
  - head-to-tail with intermediary observed: $A \perp\!\!\!\perp B \mid C$
  - tail-to-tail with intermediary observed: $A \perp\!\!\!\perp B \mid C$
  - head to head with neither the intermediary nor any of its descendants observed: $A \perp\!\!\!\perp B \mid \varnothing$



Bayes Ball Algorithm – an algorithm for determining reachability under a particular definition of separation. In particular, it determines if there exists a path from set X_A to set X_B given that the X_C are "specified."

1. Place a ball in all nodes of X_A.
2. For each ball in the graph, explore each direct path the ball could use to move through some neighboring node; this includes return paths where a node serves as both origin and destination. If the path is valid according to the rules of separation, place a ball at the destination.
3. Upon termination, if a ball is in a member of X_B, the set is reachable; return true. Otherwise return false.

Probabilistic Inference – the computation of $P\left(X_F \mid X_E\right)$ for a graph $G = \left(\nu, \varepsilon\right)$ where $F, E \subseteq \nu$ index sets such that $F \cap E = \varnothing$; disjoint.

- **query nodes**: X_F; we want to obtain the conditional probability of these.
- **evidence nodes**: variables begin conditioned on, X_E
- **remaining nodes**: X_R where $R = \nu \setminus \left(F \cup E\right)$. Must be marginalized!
- **marginal** $P\left(x_F, x_E\right) = \sum_{x_R} P\left(x_F, x_E, x_R\right)$
- **prior** $P\left(x_E\right) = \sum_{x_F} P\left(x_F, x_E\right)$
- **conditional** $P\left(x_F \mid x_E\right) = \dfrac{P\left(x_F, x_E\right)}{P\left(x_E\right)}$
- Notes:
  - Using the distributive law, factors irrelevant to a summation can be brought outside of it. By associative law, the order of sums can also be swapped.
  - Each summation introduces a new factor that has the marginalized variable removed but incorporates all other variables used in that product.
  - Determining the optimal ordering of sums that minimizes size of intermediate terms is, in general, NP-hard.

- **Conditioning –** the act of basing the probability of the query nodes on specific values of the evidence nodes.
  - **evidence potential** $\delta\left(x_i, \overline{x}_i\right)$ **-** potential that is 1 if $x_i = \overline{x}_i$; 0 otherwise: Kronecker delta function.
  - evidence potentials transform evaluations into sums:

$$g\left(\overline{x}_i\right) = \sum_{x_i} g\left(x_i\right)\delta\left(x_i, \overline{x}_i\right)$$

DO BAYES NET EXAMPLE