Session 10

I. Announcements [5 minutes]

- **Homework 4 is online and is due November 4th**
  - Get started early and get ahead of the game.
- **Homework 5 is after that and is a programming assignment.**

Cheating

The written assignments are to be done individually, the project assignments in pairs.

Discussion of assignments among students is permitted and encouraged, but solutions and programs may not be copied. I would recommend NOT mixing discussion with writing up of solutions or code.

Overview

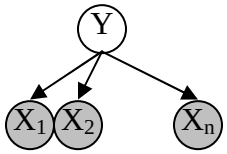
Last session we covered an introduction to Bayes Nets and probabilistic inference. Today's session is a review session focusing on the topics you feel least comfortable with. The following topics are fair game for discussion and are relevant to homework questions:

1. Separation in Bayes Nets
   a. undirected graph separation
   b. separation in Bayes Nets
2. Constructing Bayes Nets
   a. Building in the Causal direction.
3. Probabilistic Inference on Bayesian Nets
   a. joint probability and marginalization.
   b. variable elimination
   c. junction tree
4. Approximate Inference on Bayes Nets
   a. Simple sampling (estimating pi)
   b. The Markov Chain
   c. MCMC
      i. stationary distributions
      ii. Metropolis-Hastings?
      iii. Gibbs sampler
   d. particle filtering?
5. An introduction to temporal structure
   a. Markov Assumption
   b. Markov model, structure, HMM
   c. Kalman Filter
   d. Dynamic Bayes Net

II. Introduction to Bayes Nets

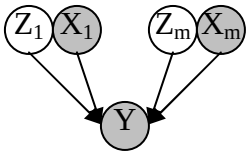
Structure of Bayes Nets

- 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.
  - Naïve Bayes Model



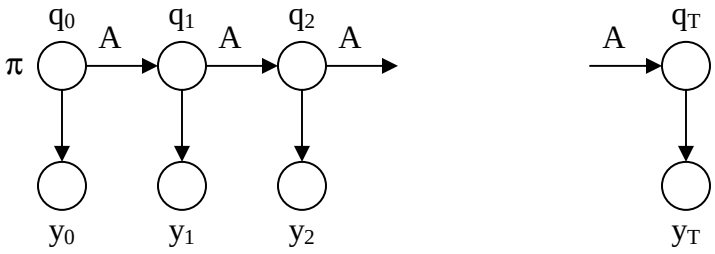
(a) conditionally independent features

  - Noisy Or Model



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

  - Hidden Markov Model



- 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:

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

Chain Rule of Probability Theory – In general,

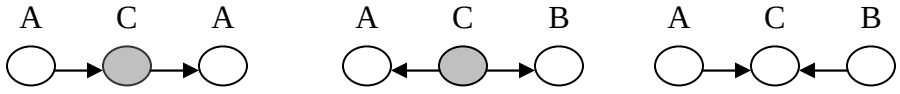
$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

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.

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

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 $P(X_F \mid X_E)$ for a graph $G = (\nu, \varepsilon)$ 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 $R = \nu \setminus (F \cup E)$. Must be marginalized!
- **marginal** $P(x_F, x_E) = \sum_{x_R} P(x_F, x_E, x_R)$
- **prior** $P(x_E) = \sum_{x_F} P(x_F, x_E)$
- **conditional** $P(x_F \mid x_E) = \dfrac{P(x_F, x_E)}{P(x_E)}$
- 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** $\delta(x_i, \overline{x}_i)$ - potential that is 1 if $x_i = \overline{x}_i$; 0 otherwise: Kronecker delta function.
  - evidence potentials transform evaluations into sums:
$$g(\overline{x}_i) = \sum_{x_i} g(x_i)\,\delta(x_i, \overline{x}_i)$$
- Continuous Random Variables:
  - **discretization** – dividing variable's possible values into intervals.
  - **parameterization** – describing the variable's distribution by a finite set of parameters.
  - **hybrid BN** – a BN containing both discrete and continuous variables.
  - conditional distributions for continuous variables:
    - discrete parents' values are enumerated.
    - continuous parents' must be summarized in a distribution, for instance, the linear Gaussian distribution where mean varies linearly with parents' value and std dev is fixed: $\mu = ax + b$.
    - linear Gaussian has joint distribution is multivariate Gaussian over all variables. These are combined with discrete variables in conditional Gaussians.
  - conditional distributions for discrete variables with continuous parents.

Approximate Inference in Bayesian Networks

- Monte Carlo algorithms – algorithms that approximate a desired quantity through random sampling.
- Direct Sampling
- Rejection Sampling
- Likelihood Weighting
- Markov chain Monte Carlo (MCMC) – a sampling technique that settles into a *dynamic equilibrium* such that the long-term fraction of time spent in each state is exactly its posterior probability given certain conditions.
    - **Markov chain** – a structure that defines the probability of transitioning from the "current" state to the "next" state.
        - *transition probability* $q(\mathbf{x} \to \mathbf{x}')$ - the probability that the process transitions from state **x** to state **x'**.
        - ***ergodic*** – essentially every state much be reachable from every other and there can be no strictly periodic cycles.
        - *state distribution* $\pi_t(\mathbf{x})$ - the probability of being in state **x** at the *t*-th step of the Markov chain.
    - ***stationary distribution*** – a state distribution such that $\pi_t = \pi_{t+1}$

$$\forall \mathbf{x}' \qquad \pi(\mathbf{x}') = \sum_{\mathbf{x}} \pi(\mathbf{x}) q(\mathbf{x} \to \mathbf{x}')$$

        - This distribution is unique if the chain is *ergodic*.
        - A distribution is stationary if it satisfies the <u>*detailed balance equation:*</u>

$$\forall \mathbf{x}, \mathbf{x}' \qquad \pi(\mathbf{x}) q(\mathbf{x} \to \mathbf{x}') = \pi(\mathbf{x}') q(\mathbf{x}' \to \mathbf{x})$$

    -

15: Probabilistic Reasoning Over Time

Modeling Uncertainty over Time

- Setting
  - X_t - a set of unobserved state variables at time *t*.
  - E_t - a set of observable evidence variables for time *t*.
  - *a:b* – denotes an interval from *a* to *b*.
- **Stationary Process** – process of change that is governed by laws that do not change over time.
- **Markov Assumption** – current state depends only on a *finite* history of previous states. Processes satisfying this assumption are *Markov Processes (Chains).*
  - **transition model** – law describing how state changes over time.

    $$P(X_t \mid X_{0:t-1}) = P(X_t \mid X_\alpha) \text{ where } \alpha \subseteq \{1 \ldots t-1\}$$
  - **first-order Markov Process** – current state is solely dependent on the previous state
    - transition model: $P(X_t \mid X_{t-1})$
- We assume the evidence variables at time *t* depend only on the current state.
  - **sensor model** – law describing how the evidence depends on the state.

    $$P(E_t \mid X_{0:t}, E_{0:t-1}) = P(E_t \mid X_t)$$
- prior probability for the initial state: $P(X_0)$
- complete joint

$$P(X_{0:T}, E_{1:T}) = P(X_0)\prod_{t=1}^{T} P(X_t \mid X_{t-1})P(E_t \mid X_t)$$

- Ways to deal with inaccurate Markov modeling:
  1. Increase the order of the Markov process
  2. Increase the set of state variables

Filter (monitoring) – the task of computing the *belief state* – the posterior distribution of the current state given all evidence; $P(X_T \mid e_{1:T})$.

- Recursive estimation – forward chaining.

$$P(X_t \mid e_{1:t}) \propto P(e_t \mid X_t)\sum_{X_{t-1}} P(X_t \mid X_{t-1})\underbrace{P(X_{t-1} \mid e_{1:t-1})}_{\text{recursive estimate}}$$

$$f_{1:t} \propto FORWARD(f_{1:t-1}, e_t)$$

- When the state variables are discrete, this update is constant in space and time.
- *Likelihood* $P(e_{1:T})$ can be calculated by a likelihood message: $l_{1:t} = P(X_t, e_{1:t})$:

$$L_{1:T} = \sum_{X_T} l_{1:T}(X_T, e_{1:T})$$

Prediction – task of computing the posterior distribution over a *future* state, given all evidence; $P(X_{T+k} \mid e_{1:T})$ where $k > 0$.

- This is equivalent to filtering without new evidence. Hence, we can easily derive the following update:

$$P(X_{T+k} \mid e_{1:T}) = \sum_{X_{t+k}} P(X_{T+k} \mid X_{T+k-1}) \underbrace{P(X_{T+k-1} \mid e_{1:T})}_{\text{recursive estimate}}$$

- **stationary distribution** – The fixed point of the Markov process that is approached upon successive applications of the transition model.
  - **mixing time** – the amount of time required to reach stationarity.
  - Prediction is doomed to failure for future times more than a small fraction of the mixing time.

Smoothing (hindsight) – task of computing posterior distribution for a *past* state, given all evidence; $P(X_k \mid e_{1:T})$ where $0 \le k < T$.

- Accounting for hindsight is done with an additional backwards message:

$$P(X_k \mid e_{1:T}) \propto \underbrace{P(X_k \mid e_{1:k})}_{f_{1:k}} \underbrace{P(e_{k+1:T} \mid X_k)}_{b_{k+1:T}}$$

$$b_{k+1:T} = \sum_{X_{k+1}} P(e_{k+1} \mid X_{k+1}) P(X_{k+1} \mid X_k) b_{k+2:T}$$

- The time and space needed for each backward message are constant.
- Thus, the process of smoothing with respect to $e_{1:T}$ is *O(t)*.
- Thus, to smooth the whole sequence naively, requires $O(t^2)$.
- using dynamic programming the cost is only *O(t)* by recording results of forward filtering over the entire sequence while running the backward algorithm from *T* to 1 and use the smoothed message at each time step ➔ **forward-backward algo**.
  - space is now $O(|f|t)$
- In on-line setting, smoothed estimates must be computed for earlier time slices as new observations are added:
  - **fixed-lag smoothing** – smoothing is done for the time slice *d* steps behind the current time *T*.

Most Likely Explanation – task of finding the sequence of states most likely to have generated a sequence of observations; $\arg\max_{x_{1:t}} P(x_{1:t} \mid e_{1:t})$.

- most likely sequence must consider joint probabilities over all time steps.
- *there is a recursive relationship between most likely paths to each state X_{t+1} and the most likely paths to each state X_t.*
- Recursive formulation:

$$\max_{X_{1:t-1}} P(X_{1:t} \mid e_{1:t}) \propto \underbrace{P(e_t \mid X_t)}_{\text{observation}} \max_{X_{t-1}} \left[\underbrace{P(X_t \mid X_{t-1})}_{\text{transition}} \underbrace{\max_{X_{1:t-2}} P(X_{1:t-1} \mid e_{1:t-1})}_{\text{previous message}} \right]$$

  - messages: $m_{1:t} = \max_{X_{1:t-1}} P(X_{1:t} \mid e_{1:t})$
  - summation over X_t replaced by a maximization.
- Pointers are used to retrieve the most-likely explanation
- Viterbi algorithm has a space and time requirement of *O(t).*

Learning – task of learning the transition and sensor models from observed data. This process leverages inference through EM.

Hidden Markov Models (HMM) – a temporal probabilistic model in which the state of the process is described by a *single discrete* random variable and transitions obey the Markov assumption.

- transition model: $T_{ij} = P(X_t = j \mid X_{t-1} = i)$
- observation model: $(\mathbf{O_t})_{i,i} = P(e_t \mid X_t = i)$
  - *forward* message - $\mathbf{f}_{1:t+1} \propto \mathbf{O}_{t+1}\mathbf{T}^T\mathbf{f}_{1:t}$
  - *backward* message - $\mathbf{b}_{k+1:t} \propto \mathbf{T}\mathbf{O}_{k+1}\mathbf{b}_{k+2:t}$
  - time complexity of forward-backward becomes $O(S^2 t)$ where S is the number of hidden states and space complexity is $O(St)$.

Kalman Filters – a temporal probabilistic model for continuous state spaces under the Markov assumption and using linear Gaussian distributions to model the states. A Kalman filter can model any system of continuous state variables with noisy measurements.

- a *multivariate Gaussian* distribution can be specified completely by its mean $\boldsymbol{\mu}$ and its covariance matrix $\boldsymbol{\Sigma}$.
- In general, filtering with continuous or hybrid spaces generate state distributions whose representations grow without bound, but the Gaussian distribution is "well-behaved" since it has the following properties:
    1. If the current distribution $P(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ is Gaussian and the transition model $P(\mathbf{X}_{t+1} \mid \mathbf{x}_t)$ is linear Gaussian, then the predicted distribution of the next step is:
    $$P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) = \int_{\mathbf{x}_t} P(\mathbf{X}_{t+1} \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) \, d\mathbf{x}_t$$
    2. If the predicted distribution is Gaussian and the observation (sensor) model is linear Gaussian, then conditioning on new evidence yields the updated distribution:
    $$P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) \propto P(\mathbf{e}_{1:t+1} \mid \mathbf{X}_{t+1}) P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t})$$
- General formulation:
    $$P(\mathbf{x}_{t+1} \mid \mathbf{x}_t) = N(\mathbf{F}\mathbf{x}_t, \boldsymbol{\Sigma}_x)(\mathbf{x}_{t+1})$$
        - $\mathbf{F}$ and $\boldsymbol{\Sigma}_x$ describe the linear transition model & noise.
    $$P(\mathbf{z}_t \mid \mathbf{x}_t) = N(\mathbf{H}\mathbf{x}_t, \boldsymbol{\Sigma}_z)(\mathbf{z}_t)$$
        - $\mathbf{H}$ and $\boldsymbol{\Sigma}_z$ describe the linear sensor model & noise.
- Updates:
    $$\boldsymbol{\mu}_{t+1} = \mathbf{F}\boldsymbol{\mu}_t + \mathbf{K}_{t+1}(\mathbf{z}_{t+1} - \mathbf{H}\mathbf{F}\boldsymbol{\mu}_t)$$
    $$\boldsymbol{\Sigma}_{t+1} = (\mathbf{I} - \mathbf{K}_{t+1})(\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)$$
    - Kalman gain $K_{t+1} = (\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)\mathbf{H}^T \left(\mathbf{H}(\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)\mathbf{H}^T + \boldsymbol{\Sigma}_z\right)^{-1}$
        - A measure of "how seriously to take the new observation" relative to the prediction.
    - predicted state at t+1 is $\mathbf{F}\boldsymbol{\mu}_t$, predicted observation is $\mathbf{H}\mathbf{F}\boldsymbol{\mu}_t$, and error of predicted observation is $(\mathbf{z}_{t+1} - \mathbf{H}\mathbf{F}\boldsymbol{\mu}_t)$.
- Extended Kalman Filter (EKF) – allows for limited nonlinearity in the model by modeling the system *locally* as linear in $\mathbf{x}_t$ in the region of $\mathbf{x}_t = \boldsymbol{\mu}_t$.
- Switching Kalman Filter –

Dynamic Bayesian Networks

Speech Recognition