Lecture 2: January 14, 2015

cs 573: Probabilistic Reasoning

Professor Nevatia

Spring 2015

Review

- Course requirements and Grading
  - Assignments (7-8, 1-2 may be programming, no projects), 30% weight
  - Exams 1 (between 7^{th} and 9^{th} weeks), Exam 2 (April 29, class time), 30% weight each
  - Class attendance, 10% (does not apply for DEN students)
- Enrollment: sign in, if in class
- Course content: as listed in Lec 1 slides (subject to listed *caveats*)
- Assignment #1 to be posted today, due Jan 26
- Last lecture: Intro to probability
  - Distribution function
  - Joint probability, conditional probability
- Today's objective
  - Independences
  - Continuous distributions
  - Graph terminology
  - Many, many definitions…

⊥

Independence

- Event α is independent of event β in P, denoted as $P \models \alpha \perp \beta$,
  if $P(\alpha \mid \beta) = P(\alpha)$, or if $P(\beta) = 0$
- Follows that $P(\alpha \cap \beta) = P(\alpha) P(\beta)$
- Examples: toss two coins; coin toss and weather…
- Full independence is rare, *conditional independence* where two events are independent, given a third event
- Conditional independence
  - P (USC | UCLA, GradeA) = P(USC | GradeA)
    (USC means admitted to USC, similar for UCLA)
  - P (Congestion | Flu, Hayfever, Season) = P (Congestion | Flu, Hayfever)
- Event α is independent of event β in P, given event γ, denoted as $P \models (\alpha \perp \beta \mid \gamma)$ if $P(\alpha \mid \beta \cap \gamma) = P(\alpha \mid \gamma)$, or if $P(\beta \mid \gamma) = 0$
- Follows that $P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma) P(\beta \mid \gamma)$

Conditional Independence Properties

- Conditional independence of variables
  - **Defn: X** is *cond indep* of **Y** given **Z**, in distribution P, if P satisfies $(x \perp y \mid z)$ for all possible values of x, y and z
- Proposition: P satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ *iff* $P(\mathbf{X},\mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})\, P(\mathbf{Y} \mid \mathbf{Z})$
- Properties of conditional independence, equations (2.7) thru (2.11)
  - Given without proof
    - **Symmetry:**
      $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{Y} \perp \boldsymbol{X} \mid \boldsymbol{Z})$.
    - **Decomposition:**
      $(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$.
    - **Weak union:**
      $(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W})$.
    - **Contraction:**
      $(\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}, \boldsymbol{Y}) \,\&\, (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z})$..
    - **Intersection:** For positive distributions, and for mutually disjoint sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}$:
      $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W}) \,\&\, (\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}, \boldsymbol{Y}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z})$.

Queries

- Probability Query
  - Given some evidence, find probability of desired variables
  - Evidence consists of instantiation $\boldsymbol{e}$ of a set of variables $\boldsymbol{E}$
  - Compute $P(\mathbf{Y} | \boldsymbol{E} = \boldsymbol{e})$, where $\mathbf{Y}$ is set of query variables
    - Marginal over $\mathbf{Y}$ conditioned on $\boldsymbol{e}$; also called *posterior* distribution
    - There may be additional variables that we don't care about
- MAP Query
  - Maximum *a posteriori* assignment or *most probable explanation* (MPE)
  - MAP $(\mathbf{W} | \boldsymbol{e}) = \arg\max_w P(\mathbf{w}, \mathbf{e})$, $\mathbf{W} = \boldsymbol{X} - \mathbf{E}$
  - Note that maximal joint assignment is not same as maxima of individual assignments, example on next page

MAP Example (Ex 2.4)

a^0	a^1
0.4	0.6

A	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Note: right table gives conditional, not joint, probabilities

$$\mathrm{MAP}(A) = a^1$$

However, $\mathrm{MAP}(A, B) = (a^0, b^1)$:

$$\arg\max_{a,b} P(a,b) \neq (\arg\max_a P(a), \arg\max_b P(b))$$

Marginal MAP Query

- We only care about the assignment of a subset of the variables
  - Disease diagnosis: full MAP query would compute joint distribution of diseases and symptoms, we may only be interested in disease probabilities
  - MAP $(\mathbf{Y} | \boldsymbol{e}) = \arg\max_y P(\mathbf{y}, \mathbf{e})$
  - Let $Z = \boldsymbol{X} - \mathbf{Y} - \mathbf{E}$
  - MAP $(\mathbf{Y} | \boldsymbol{e}) = \arg\max_y \sum_z P(\mathbf{Y}, \mathbf{Z} | \boldsymbol{e})$
  - Note: marginal MAP query can not be computed directly from a MAP query

Continuous Spaces

- Random variables may take continuous values, say in range [0,1]. Example: max temperature tomorrow
  - P (X = *x*) is = 0 in such cases
- Probability *density* function (pdf), say $p(x)$
- Cumulative distribution function

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx$$

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

$$\int_{Val(X)} p(x)dx = 1$$

Distributions

- Uniform distribution

A variable X has a uniform distribution *over* $[a, b]$, *denoted* $X \sim \mathrm{Unif}[a, b]$ *if it has the PDF*

$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise.} \end{cases}$$

- Gaussian (Normal) distribution

A random variable X has a Gaussian distribution *with* mean μ *and* variance σ^2, *denoted* $X \sim \mathcal{N}(\mu; \sigma^2)$, *if it has the PDF*

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Joint Density Functions

Let P be a joint distribution over continuous random variables $X_1, \ldots, X_n$. A function $p(x_1, \ldots, x_n)$ is a joint density function *of $X_1, \ldots, X_n$ if*

- $p(x_1, \ldots, x_n) \geq 0$ *for all values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$.*
- *p is an integrable function.*
- *For any choice of $a_1, \ldots, a_n$, and $b_1, \ldots, b_n$,*

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_1 \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

∎

Marginalize joint density function

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy.$$

Conditional Density Functions

Let $p(x, y)$ be the joint density of X and Y. The conditional density function *of Y given X is defined as*

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

When $p(x) = 0$, the conditional density is undefined. ∎

$$p(x, y) = p(x)p(y \mid x)$$

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{p(y)}$$

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be sets of continuous random variables with joint density $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. We say that $\boldsymbol{X}$ is conditionally independent *of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ if*

$p(\boldsymbol{x} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ such that $p(\boldsymbol{z}) > 0$. ∎

Expectation

- Expectation of *X* (expected or mean value) under the distribution *P*, is given by: $E_P [X] = \Sigma_x$ x . P(x)
  - Example of a dice roll
- If *X* is continuous: $E_P [X] = \int$ x . P(x)
- Expectation of *f* (x) is $E_P [f(X)] = \Sigma_x f(x)$. P(x)
- E [a . X + b] = a E[X] + b
- E[X + Y]= E[X] + E[Y]
- If X and Y are independent E [X.Y]= E[X]. E[Y]
- $E_P [X|y] = \Sigma_x$ x . P(x|y)

Variance

- *Variance* of *X*, given distribution *P*, is given by:
  $Var_P[X] = E_P[(X - E_P[X])^2] = E_P[X^2] - E_P[X]^2$

Standard Deviation $\sigma_X = \sqrt{\boldsymbol{Var}[X]}$

If X and Y are independent, then

$$\boldsymbol{Var}[X+Y] = \boldsymbol{Var}[X] + \boldsymbol{Var}[Y]$$

Let X be a random variable with Gaussian distribution $N(\mu, \sigma^2)$, then $\boldsymbol{E}[X] = \mu$ and $\boldsymbol{Var}[X] = \sigma^2$.

(Chebyshev inequality):

$$P(|X - \boldsymbol{E}_P[X]| \geq t) \leq \frac{\boldsymbol{Var}_P[X]}{t^2}$$

$$P(|X - \boldsymbol{E}_P[X]| \geq k\sigma_X) \leq \frac{1}{k^2}.$$

Entropy of a Distribution (Appendix A.1)

- Definition: **Entropy** of X given distribution P(x)
  $H_P(X) = E_P(\log 1/P(x)) = \sum_x P(x) \log(1/P(x))$
  $= -\sum_x P(x) \log P(x)$
- Consider a fair coin, H_p will be .5log .5 + .5 log .5 = 1 (log base 2)
- What if a coin always comes up heads: entropy = 0 (no need to transmit the result)
- If the coin is unfair, P(heads) = .9; entropy will be lower than for a fair coin.
- Entropy can be related to coding/information theory
  - How many bits needed to transmit the data in an optimal code
- Another interpretation is how much information do we get from a result, or how much uncertainty is introduced by a distribution
  - Consider uniform *vs* highly peaked or bi-modal distributions

More Entropy Definitions

- Joint Entropy
  $H_P(X_1, X_2 \ldots X_n) = E_P[\log 1/P(X_1, X_2 \ldots X_n)]$

- Conditional Entropy
  $H_P(X|Y) = E_P[\log 1/P(X|Y)] = H_P(X,Y) - H_P(Y)$
  - Additional cost of encoding X, when we already know Y

- Entropy Chain Rule
  $H_P(X_1, X_2 \ldots X_n) = H_P(X_1) + H_P(X_2|X_1) + H_P(X_n| X_1 \ldots, \ldots, X_{n-1})$

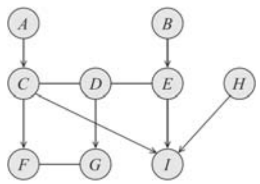
Graph Terminology

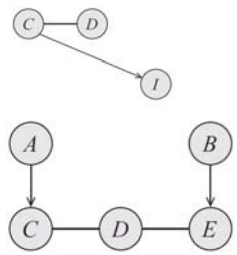
- Set of nodes and edges
- *Nodes*: $\mathcal{X} = \{X_1, \ldots, X_n\}$.
- *Edges*: *directed edge* $X_i \rightarrow X_j$ or an *undirected edge* $X_i — X_j$
- Unspecified edge type denoted by $X_i \rightleftharpoons X_j$
- Only one type of edge between two nodes (though graph may be mixed)
- *Directed graph*: all edges are directed
- *Undirected graph*: all edges are undirected
- *Parent, Child* relations (directed edge)
- *Neighbor* (undirected edge)
- *Degree* of a node: number of edges that the node participates in
- *Indegree* of a node: X number of directed edges pointing to X
- *Degree of a graph*: maximal degree of a node in the graph

Subgraphs and Cliques

Let $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and let $\boldsymbol{X} \subset \mathcal{X}$. We define the induced subgraph *$\mathcal{K}[\boldsymbol{X}]$ to be the graph $(\boldsymbol{X}, \mathcal{E}')$ where $\mathcal{E}'$ are all the edges $X \rightleftharpoons Y \in \mathcal{E}$ such that $X, Y \in \boldsymbol{X}$.* ∎

A subgraph over $\boldsymbol{X}$ is complete *if every two nodes in $\boldsymbol{X}$ are connected by some edge. The set $\boldsymbol{X}$ is often called a* clique*; we say that a clique $\boldsymbol{X}$ is* maximal *if for any superset of nodes $\boldsymbol{Y} \supset \boldsymbol{X}$, $\boldsymbol{Y}$ is not a clique.* ∎





More Graph Definitions

- Path
- Trail
- Connected Graph
- Ancestor/descendant
- Topological ordering

Paths and Trails

We say that $X_1, \ldots, X_k$ form a path *in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ if, for every $i = 1, \ldots, k-1$, we have that either $X_i \to X_{i+1}$ or $X_i — X_{i+1}$. A path is* directed *if, for at least one i, we have $X_i \to X_{i+1}$.* ∎

We say that $X_1, \ldots, X_k$ form a trail *in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ if, for every $i = 1, \ldots, k-1$, we have that $X_i \rightleftharpoons X_{i+1}$.* ∎

In the graph $\mathcal{K}$ of figure 2.3, A, C, D, E, I is a path, and hence also a trail. On the other hand, A, C, F, G, D is a trail, which is not a path.

A graph is connected *if for every X_i, X_j there is a trail between X_i and X_j.* ∎

Orderings

We say that X is an ancestor *of Y in $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and that Y is a* descendant *of X, if there exists a directed path $X_1, \ldots, X_k$ with $X_1 = X$ and $X_k = Y$. We use* $\mathrm{Descendants}_X$ *to denote X's descendants,* $\mathrm{Ancestors}_X$ *to denote X's ancestors, and* $\mathrm{NonDescendants}_X$ *to denote the set of nodes in $\mathcal{X} - \mathrm{Descendants}_X$.* ∎

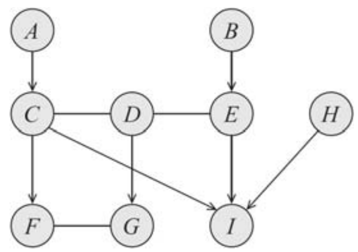
In our example graph $\mathcal{K}$, we have that F, G, I are descendants of C. The ancestors of C are A, via the path A, C, and B, via the path B, E, D, C.

A final useful notion is that of an ordering of the nodes in a directed graph that is consistent with the directionality its edges.

Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a graph. An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering *relative to $\mathcal{K}$ if, whenever we have $X_i \to X_j \in \mathcal{E}$, then $i < j$.* ∎

Cycles

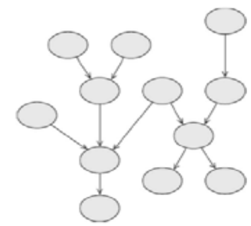
- Cycles: a cycle is a path $X_1, X_2 \ldots X_k$ where $X_1 = X_k$
- Directed acyclic graphs (DAGs)
- Partially directed acyclic graph (PDAG): some edges may be undirected. Chain components consist of subgraphs connected by undirected edges; chains are connected by directed edges. Six chain components in example: {A}, {B}, {C,D,E}, {F,G}, {H},



Loops

- Loops: a loop is a trail $X_1, X_2 \ldots X_k$ where $X_1 = X_k$
- Singly connected: no loops
  - Leaf node: only one adjacent node
  - Polytree: singly connected, directed
  - Forest: undirected- singly connected

    Directed- Each node has at most one parent
- Tree: connected direct forest



Polytree Example

Chordal Graphs

- Define a chord and a chordal graph

Let $X_1 — X_2 — \cdots — X_k — X_1$ be a loop in the graph; a chord *in the loop is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. An undirected graph $\mathcal{H}$ is said to be* chordal *if any loop $X_1 — X_2 — \cdots — X_k — X_1$ for $k \geq 4$ has a chord.* ∎

- In a chordal graph, longest minimal loop is a triangle: also called a triangulated graph
- Directed Chordal graph:

A graph $\mathcal{K}$ is said to be chordal *if its underlying undirected graph is chordal.*

Next Class

- Read sections 3.1,3.2 and 3.3 of the KF book