67800: Probabilistic Methods in AI **Spring 2017/18**

Recitation 0: Background

Teaching Assistant: Eitan Richardson

0.1 Probability Theory

Recommended reading:

- PGM Book, Chapter 2 [2]
- Standford CS 229 Course / Probability notes [3]
- Notes in Hebrew by Gal Chechnik et al [1]

0.1.1 Basic Definitions

Sample space Ω – all possible outcomes e.g. single dice throw: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Event e – a subset of the sample space $e \subseteq \Omega$ e.g. odd throw result: $e = \{1, 3, 5\}$

Event space S – a set of all relevant events, including $\emptyset$ and Ω

Probability measure $P : S \to \mathbb{R}^+$

Basic properties:

- $\forall a \in S, P(a) \geq 0, \;\; P(\Omega) = 1, \;\; P(\Omega \setminus a) = 1 - P(a)$
- $a \subseteq b \implies P(a) \leq P(b)$
- $P(a \cap b) \leq min(P(a), P(b))$
- Union bound: $P(\bigcup_i a_i) \leq \sum_i P(a_i)$ (equal if $\{a_i\}$ are disjoint events)

Conditional probability:

$$P(a \mid b) = \frac{P(a \cap b)}{P(b)}$$

The chain rule:

$$P(\bigcap_i a_i) = P(a_1)P(a_2 \mid a_1) \cdots P(a_k \mid a_1 \cap \ldots \cap a_{k-1})$$

Bayes rule:

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

Independent events:

- $P \models (a \perp b) \iff P(a \cap b) = P(a)P(b)$, or equivalently: $P(a \mid b) = P(a)$
- Conditional independence: $P \models (a \perp b \mid c) \iff P(a \mid b \cap c) = P(a \mid c)$

[1] *Original LaTeX template courtesy of UC Berkeley.*

0.1.2 Random Variables

A random variable X is a function $X : \Omega \to \mathbb{R}$, (or to one of possible set of values).

Example 0.1 *Two dice roll*
$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$
$X((n_1, n_2)) = n_1 + n_2$ – *a random variable describing the roll sum.*

Probability distribution of a random variable:

$$P(X = x) = P_X = P(\{\omega \subseteq \Omega : X(\omega) = x\})$$

P_X is a new probability distribution function associated with the random variable X. It only records the probability of different values of X.

Discrete random variable:

$$\sum_{x \in Val(X)} P(X = x) = \sum_x P_X(x) = 1$$

0.1.2.1 Moments – Expectation and Variance

Expectation (for discrete variables): $\mathbb{E}_P[X] = \Sigma_x x P_X(x)$

Linearity of expectation:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \mathbb{E}[aX] = a\mathbb{E}[X]$$

Variance: $\mathrm{Var}_P[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad = \quad \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Proof:

$$\begin{aligned}
\mathrm{Var}_P[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2 \quad \text{(inner } \mathbb{E}[X] \text{ is considered as constant)} \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

■

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$$

X, Y are independent $\implies \mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$, $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Chebyshev inequality:

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

0.1.3 Multivariate Distributions

0.1.3.1 Joint and Marginal Probability

The explicit joint distribution (for two discrete random variables X and Y) is a table assigning a probability value for every combination of $Val(X) \times Val(Y)$, for example:

Example 0.2

$P(X,Y) =$

		X: 0	1	2	P_Y
Y	A	0.3	0.3	0	0.6
	B	0.2	0.1	0.1	0.4
	P_X	0.5	0.4	0.1	

☞ *There are* $(|Val(X)| \times |Val(Y)| - 1)$ *degrees of freedom in the general joint probability table.*

The joint probability should be consistent with the marginal probabilities (sums of rows or columns):

$$\Sigma_x P_{X,Y}(x,y) = P_Y(y) \quad \text{and} \quad \Sigma_y P_{X,Y}(x,y) = P_X(x)$$

0.1.3.2 Conditional Probability of Random Variables

$P(X \mid Y = y)$ is the *conditional distribution* over the outcomes defined by X *given* the knowledge that $Y = y$.

$P(X \mid Y)$ assigns a probability distribution over X for each value of Y:

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

In **Example 0.2**, $P(X \mid Y = A) = \begin{bmatrix} 0.5 & 0.5 & 0 \end{bmatrix}$.

The chain rule:

$$P(X1, \ldots, X_k) = P(X_1)P(X_2 \mid X_1) \cdots P(X_k \mid X_1, \ldots, X_{k-1})$$

Bayes rule:

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

0.1.4 Conditional Probability Distributions and Noisy Or

In some cases (as we will see extensively in this course), instead of defining the *joint probability distribution*, we decompose it and use *conditional probability distributions* (CPDs). For example, a joint probability distribution over three RV's can be decomposed per the chain rule to $P(X,Y,Z) = P(X)P(Y|X)P(Z|X,Y)$, and if we know that $P \models X \perp Y$, we have: $P(X,Y,Z) = P(X)P(Y)P(Z \mid X,Y)$.

In the case of discrete RVs, the CPD $P(Z \mid X,Y)$ can be defined explicitly by a table. Notice that each row sums to 1 (so there are 4 free parameters). Notice the notation $P(z^0) = P(Z=0) = P_Z(0)$.

X	Y	$P(z^0 \mid X,Y)$	$P(z^1 \mid X,Y)$
0	0	1	0
0	1	0.5	0.5
1	0	0.2	0.8
1	1	0.1	0.9

☞ *The explicit CPD table grows exponentially with the number of parameters conditioned-on.*

In many cases, we don't need to model a complex interaction between the different causes (combinations of parent values) – we might want the CPD to represent some sort of a probabilistic *OR* model that scales linearly with additional variables. The *Noisy Or* model provides this kind of *independence of causal influence*:

Definition 0.3 *Noisy Or*
A binary RV Y depends on k binary variables $X_1, \ldots, X_k$ in a noisy-or *model if:*

$$P(Y = 0 \mid x_1, \ldots, x_k) = (1 - \lambda_0) \prod_{i=1}^{k} (1 - \lambda_i)^{x_i}$$

λ_0 is the *leak* parameter, allowing a positive probability for $Y = 1$ even of all X_i are 0.
λ_i are the *noise parameters* defining the amount by which $X_i = 1$ reduces the probability that $Y = 0$.
When $\lambda_0 = 0$ and all noise parameters λ_i equal 1, the model behaves like a deterministic OR.

The CPD we defined above actually matches a Noisy-Or model with parameters $\lambda_0 = 0$, $\lambda_Y = 0.5$, $\lambda_X = 0.8$.

Example 0.4 *Noisy-or and* explaining away
We have a joint distribution of three binary RVs defined by $P(X, Y, Z) = P(X)P(Y)P(Z \mid X, Y)$, where $P(Z \mid X, Y)$ is defined by a noisy-or model. We need to show that the model satisfies the explaining away *property: $P(x^1|z^1) \geq P(x^1|y^1, z^1)$.*

Proof: We will prove for the case of $\lambda_0 = 0$ (although the claim is true in the general case)

$$P(x^0 \mid y^1, z^1) = \frac{P(y^1 \mid x^0, z^1)P(x^0 \mid z^1)}{P(y^1 \mid z^1)}$$

(this is an extension of the Bayes rule, we will prove later)

$$\begin{aligned} P(y^1 \mid x^0, z^1) &= 1 - P(y^0 \mid x^0, z^1) \\ &= 1 - \frac{P(z^1|x^0, y^0)P(y^0 \mid x^0)}{P(z^1 \mid x^0)} \\ &= 1 \quad \text{since } P(z^1|x^0, y^0) = 0 \end{aligned}$$

Substituting this into the first expression we get, $P(x^0 \mid y^1, z^1) = \frac{P(x^0|z^1)}{P(y^1|z^1)} \geq P(x^0 \mid z^1)$ (since $P(y^1 \mid z^1) \leq 1$).
Thus $P(x^1 \mid y^1, z^1) = 1 - P(x^0 \mid y^1, z^1) \leq 1 - P(x^0 \mid z^1) = P(x^1|z^1)$.

☞ *Note that we did not use the fact that the CPD is a noisy-or, but only that $P(z^1|x^0, y^0) = 0$.*

■

Claim 0.5

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

Proof: Per the chain rule:

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

■

0.1.5 Independence in Random Variables

$\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ (in bold) are <u>sets</u> of random variables.

$(\mathbf{X} \perp \mathbf{Y}) \iff P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$: $\mathbf{X}$ and $\mathbf{Y}$ are (marginally) independent.

In **Example 0.2**, $P(X = 1, Y = A) = 0.3 \neq 0.4 \times 0.6 = P(X = 1)P(Y = A) \Rightarrow$ not independent.

$(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \iff P(\mathbf{X}, \mathbf{Y} \mid Z) = P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Y} \mid \mathbf{Z})$: $\mathbf{X}$ and $\mathbf{Y}$ are *conditionally independent* given $\mathbf{Z}$.

Additional properties:

- Symmetry: $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \implies (\mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z})$
- Decomposition: $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
- Weak union: $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \implies (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$
- Contraction: $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z}, \mathbf{Y})$ and $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \implies (\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$

Proof: (Decomposition)
By the definition of conditional independence we have: $P(X, Y, W \mid Z) = P(X \mid Z) \cdot P(Y, W \mid Z)$

$$\begin{aligned} P(X, Y \mid Z) &= \sum_w P(X, Y, w \mid Z) \\ &= P(X \mid Z) \sum_w P(Y, w \mid Z) \\ &= P(X \mid Z) P(Y \mid Z) \end{aligned}$$

■

0.1.6 Queries

Once we built a probability distribution, we can use it to answer some questions.

The posterior distribution given some evidence: $P(\mathbf{Y} \mid \mathbf{E} = e)$

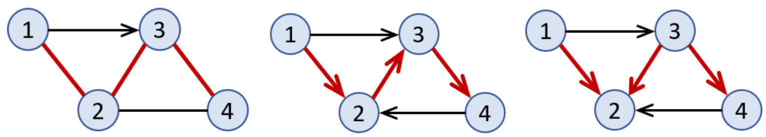
Let $\boldsymbol{\chi}$ be the set of all random variables, $\mathbf{E}$ the observed variables (evidence), $\mathbf{Y}$ the set of variables we are interested in and $\mathbf{Z} = \boldsymbol{\chi} - \mathbf{Y} - \mathbf{E}$, all other variables.

The *marginal MAP query* of $\mathbf{Y}$ given $\mathbf{E}$ is:

$$MAP(\mathbf{Y} \mid \mathbf{E} = e) = \arg\max_y \sum_z P(Y = y, Z = z \mid E = e)$$

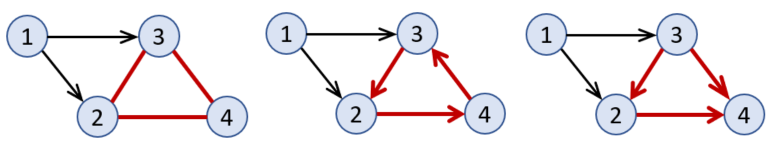
0.2 Graphs

0.2.1 Paths, Trails, Cycles and Loops



(a) (b) (c)

Figure 0.1: (a) undirected path (b) directed path (c) trail



(a) (b) (c)

Figure 0.2: (a) undirected cycle (b) directed cycle (c) loop

0.2.2 Trees and Forests

Some definitions:

- *DAG* – directed graph that contains no cycles
- *Singly-connected* – contains no cycles or loops
- Singly-connected undirected graph = *forest*
- Singly-connected undirected graph that is also connected = *tree*
- Singly-connected directed graph = *polytree*
- Directed graph with at most one *parent* per node = *forest*
- A connected directed forest = *tree*

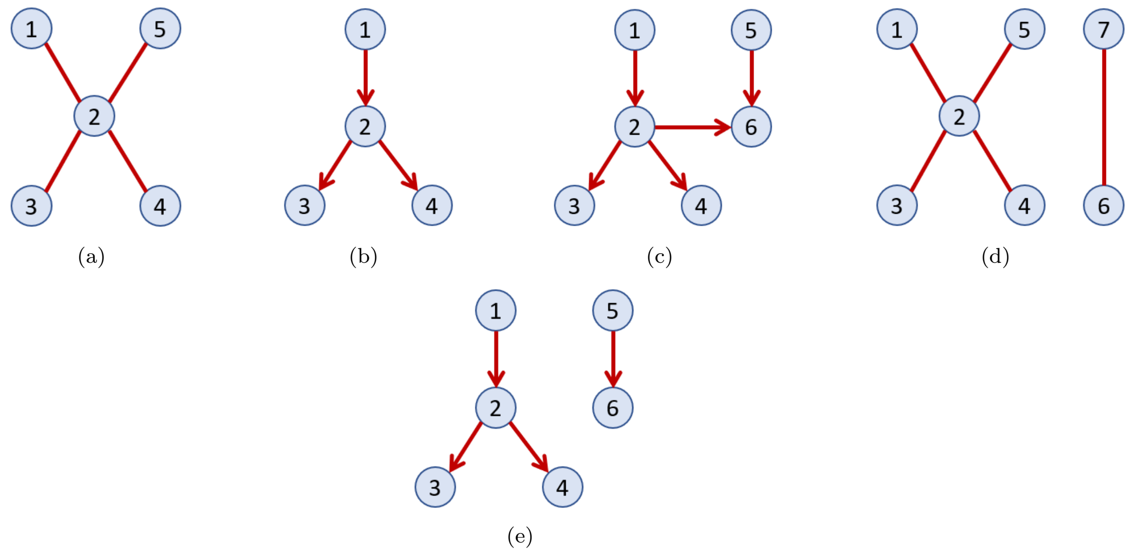


Figure 0.3: (a) undirected tree (b) directed tree (c) polytree (single-connected directed graph) (d) undirected forest (e) directed forest

0.2.3 Topological Ordering

Definition 0.6 *Topological Ordering*
An ordering of the nodes $X_1, \ldots, X_n$ in a graph $\mathcal{G} = (\chi, \mathcal{E})$ is a topological ordering *if whenever $X_i \to X_j \in \mathcal{E}$ than $i < j$.*

Given a directed graph, there might be several valid topological ordering of the nodes. For example:

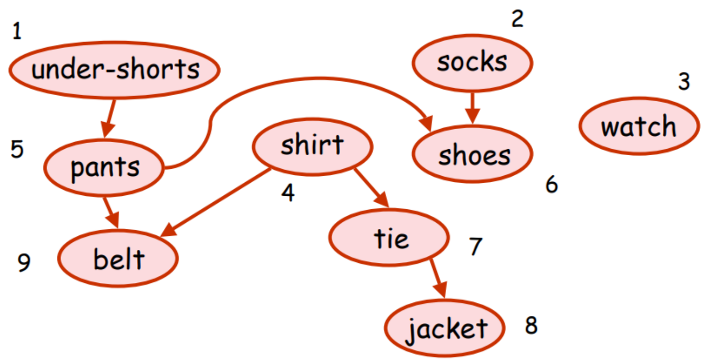


Figure 0.4: Directed graph and a valid topological ordering

Claim 0.7 *A DAG has at least one topological order.*

There are several algorithms for finding a topological order (for example, using DFS). We will discuss the following algorithm:

Algorithm 1 Topological Sort

procedure TOPOLOGICALSORT(DAG $\mathcal{G}$)
    $res \leftarrow \emptyset$
    while $\mathcal{G}$ is not empty **do**
        $v \leftarrow$ any node in $\mathcal{G}$ with zero in-degree (no parents)
        Add v to res
        Remove v and its edges from $\mathcal{G}$
    return res

Proof: Correctness of Algorithm 1
We need to show that whenever we add a node v to res, we do not break the topological order i.e. there is no edge in $\mathcal{G}$ from a node not in res yet to v.

This is true since:

- If v had no parents originally, there cannot be such an edge.
- If v has parents that were removed, they are already in res.

The algorithm cannot get stuck since every DAG has at least one node with zero in-degree (if every node has a parent, we can continue traveling upwards until we reach a visited node). ■

References

[1] Gal Chechnik. Background on probability theory. `https://moodle.cs.huji.ac.il/cs12/file.php/67800/chap1.pdf`.

[2] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[3] Andrew Ng. CS 229 machine learning course materials. *Stanford University*, 2016. `http://cs229.stanford.edu/section/cs229-prob.pdf`.