Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Recitation 1

1 Probability Review

1.1 Bayes' Rule

Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbb{P}(A_i) > 0$, for all i. Then, for any event B s.t. $\mathbb{P}(B) > 0$, we have

$$\begin{aligned}\mathbb{P}(A_i \mid B) &= \frac{\mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \ldots + \mathbb{P}(B \mid A_n)\mathbb{P}(A_n)}.\end{aligned}$$

Example: Medical Diagnosis

Suppose there is a deadly rare disease, and 0.1% – one out of every 1000 people – have this disease. The current medical test for this disease is 99% accurate, i.e. if someone has it, the test will be positive 99% of the time; and if someone doesn't have it, the test will come up negative 99% of the time.

Bob took the test and got positive results. How worried should he be? More precisely, what is the probability that Bob actually has the disease given that his test result is positive?

Let variable D $\in \{0, 1\}$ denote whether or not Bob has the disease, and let variable T $\in \{0, 1\}$ denote whether or not the test result is positive.

$$\begin{aligned}\mathbb{P}(D = 1 \mid T = 1) &= \frac{\mathbb{P}(T = 1 \mid D = 1)\mathbb{P}(D = 1)}{\mathbb{P}(T = 1)} \\ &= \frac{\mathbb{P}(T = 1 \mid D = 1)\mathbb{P}(D = 1)}{\sum_{d \in \{0,1\}} \mathbb{P}(T = 1 \mid D = d)\mathbb{P}(D = d)} \\ &= \frac{\mathbb{P}(T = 1 \mid D = 1)\mathbb{P}(D = 1)}{\mathbb{P}(T = 1 \mid D = 0)\mathbb{P}(D = 0) + \mathbb{P}(T = 1 \mid D = 1)\mathbb{P}(D = 1)} \\ &= \frac{(0.99)(0.001)}{(0.01)(0.999) + (0.99)(0.001)} \\ &= 0.0902\end{aligned}$$

In other words, although the test came back positive, the probability that Bob actually has the disease is still less than 10%! However, notice this conditional probability is considerably higher than the prior (0.1%), as expected.

1.2 Independence

Two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

In addition, two events A and B are conditionally independent on an event C if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C).$$

In general, events $A_1, \ldots, A_n$ are independent if

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{S}} A_i\right) = \prod_{i \in \mathcal{S}} \mathbb{P}(A_i), \quad \text{for every subset } \mathcal{S} \text{ of } \{1, \ldots, n\}.$$

Example: $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ **not enough for independence of A, B, C**

Consider two independent rolls of a fair die, and the following events:

$$\begin{aligned} A &= \{\text{1st roll is 1, 2, or 3}\}, \\ B &= \{\text{1st roll is 3, 4, or 5}\}, \\ C &= \{\text{The sum of the two rolls is 9}\}. \end{aligned}$$

We have

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{36} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

Yet

$$\begin{aligned} \mathbb{P}(A \cap B) &= \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B) \\ \mathbb{P}(A \cap C) &= \frac{1}{36} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbb{P}(A)\mathbb{P}(C) \\ \mathbb{P}(B \cap C) &= \frac{1}{12} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbb{P}(B)\mathbb{P}(C). \end{aligned}$$

The intuition behind the independence of a collection of events is analogous to the case of two events. Independence means that the occurrence or non-occurrence of **any number** of the events from that collection carries no information on the remaining events or their complements. For example, if the events A_1, A_2, A_3, A_4 are independent, one obtains relations such as

$$\mathbb{P}(A_1 \cup A_2 \mid A_3 \cap A_4) = \mathbb{P}(A_1 \cup A_2)$$

or

$$\mathbb{P}(A_1 \cup \bar{A}_2 \mid \bar{A}_3 \cap A_4) = \mathbb{P}(A_1 \cup \bar{A}_2).$$

Example: Pairwise independence does NOT imply global independence

Consider two independent fair coin tosses, and the following events:

$$H_1 = \{\text{1st toss is a Head}\}$$
$$H_2 = \{\text{2nd toss is a Head}\}$$
$$D = \{\text{The two tosses have different results}\}.$$

The events H_1, H_2 are independent by definition. To see that H_1 and D are independent, we note that

$$\mathbb{P}(D \mid H_1) = \frac{\mathbb{P}(D \cap H_1)}{P(H_1)} = \frac{1/4}{1/2} = \frac{1}{2} = \mathbb{P}(D).$$

H_2 and D are independent following similar argument.
On the other hand, we have that:

$$\mathbb{P}(D \cap H_1 \cap H_2) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(D)\mathbb{P}(H_1)\mathbb{P}(H_2).$$

Thus H_1, H_2 and D are not globally independent.

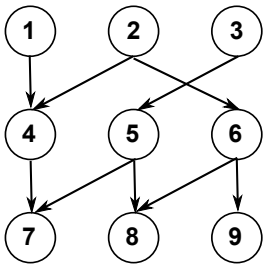
2 Directed Acyclic Graphs

Consider the following two ways of associating graphical models with distributions:
(1) Given a graphical model, what is the set of distributions that factorize according to the graph structure?
(2) Given a graphical model, what is the set of distributions that satisfy all the conditional independences the graph has?

It turns out that these two sets are exactly the same[1]. In other words, the two characterizations (through factorization and conditional independences) are equivalent. For directed graphs, there is a natural factorization into a product of conditional probabilities but determining conditional independences require relatively complicated rules such as Bayes ball algorithm. Next week, we will see that for undirected graphs, in contrast, conditional independences can be directly read off the graph while factorization requires additional thinking.

[1] A rigorous proof of this statement can be found in standard textbooks, for example, *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller and Nir Friedman.

Example Given the directed graph below, use Bayes Ball algorithm to determine:



(1) Are Node 3 and Node 9 independent?
(2) Are Node 3 and Node 9 independent given that Node 7 is observed?
(3) Find all i, such that Node 3 and Node 9 are independent given that Node i is observed.

Solution:
(1) Yes
(2) No
(3) 1,2,4,5,6

3 Order of growth: Big-O notation

We say that $f(n) = O(g(n))$ ('$f(n)$ is big-O of $g(n)$'), if and only if there exist constants $C > 0$ and n_0 such that

$$|f(n)| \leq C|g(n)| \qquad \text{for all } n > n_0.$$

Examples: $n^2 + n + 1 = O(n^2)$, $n = O(n^2)$, $2^n + n^{42} = O(2^n)$.
Remark: Although $n = O(n^2)$ is technically correct because big-O is an upper bound, typically people give the smallest upper bound as this is more informative, e.g., we would say $n = O(n)$ rather than $n = O(n^2)$.

For completeness, we include the definitions of other asymptotic notations here:

1. We say that $f(n) = o(g(n))$ ('$f(n)$ is little-o of $g(n)$'), if and only if $\forall \epsilon > 0$, $\exists\ n_0$, s.t. $\forall\ n > n_0$
$$|f(n)| \leq \epsilon|g(n)| \qquad \text{for all } n > n_0.$$
2. We say that $f(n) = \Omega(g(n))$ ('$f(n)$ is Big-Omega of $g(n)$'), if and only if there exist constants $C > 0$ and n_0 such that
$$|f(n)| \geq C|g(n)| \qquad \text{for all } n > n_0.$$
3. We say that $f(n) = \omega(g(n))$ ('$f(n)$ is little-omega of $g(n)$'), if and only if $\forall K > 0$, $\exists\ n_0$, s.t. $\forall\ n > n_0$
$$|f(n)| \geq K|g(n)| \qquad \text{for all } n > n_0.$$
4. We say that $f(n) = \Theta(g(n))$ ('$f(n)$ is Big-Theta of $g(n)$'), if and only if there exist constants $C_1 > 0$, $C_2 > 0$ and n_0 such that
$$C_1|g(n)| \leq |f(n)| \leq C_2|g(n)| \qquad \text{for all } n > n_0.$$

6.438 Algorithms for Inference

Fall 2014