STAT 535 Lecture 2

Independence and conditional independence

©Marina Meilă
mmp@stat.washington.edu

1 Conditional probability, total probability, Bayes' rule

Definition of *conditional distribution* of A given B.

$$P_{A|B}(a|b) \;=\; \frac{P_{AB}(a,b)}{P_B(b)}$$

whenever $P_B(b) \neq 0$ (and $P_{A|B}(a|b)$ undefined otherwise).

Total probability formula. When A, B are events, and $\bar{B}$ is the complement of B

$$P(A) \;=\; P(A,B) + P(A,\bar{B}) \;=\; P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

When A, B are random variables, the above gives the *marginal distribution* of A.

$$P_A(a) \;=\; \sum_{b\in\Omega(B)} P_{AB}(a,b) \;=\; \sum_{b\in\Omega(B)} P_{A|B}(a|b)P_B(b)$$

Bayes' rule

$$P_{A|B} \;=\; \frac{P_A P_{B|A}}{P_B}$$

The chain rule Any multivariate distribution over n variables $X_1, X_2, \ldots X_n$ can be decomposed as:

$$P_{X_1,X_2,\ldots X_n} \;=\; P_{X_1} P_{X_2|X_1} P_{X_3|X_1X_2} P_{X_4|X_1X_2X_3} \ldots P_{X_n|X_1\ldots X_{n-1}}$$

2 Probabilistic independence

$$\boxed{A \perp B \iff P_{AB} = P_A.P_B}$$

We read $A \perp B$ as "A independent of B". An equivalent definition of independence is:

$$\boxed{A \perp B \iff P_{A|B} = P_A}$$

The above notation are shorthand for

$$\begin{aligned} \text{for all}\, a \in \Omega(A),\, b \in \Omega(b),\;\; P_{A|B}(a|b) &= P_A(a) \text{ whenever } P_B \neq 0 \\ \frac{P_{AB}(a,b)}{P_B(b)} &= P_A(a) \text{ whenever } P_B \neq 0 \\ P_{AB}(a,b) &= P_A(a)P_B(b) \end{aligned}$$

Intuitively, probabilistic independence means that knowing B does not bring any additional information about A (i.e. doesn't change what we already believe about A). Indeed, the *mutual information*[1] of two independent variables is zero.

Exercise 1 [The exercises in this section are optional and will not influence your grade. Do them only if you have never done them before.]

Prove that the two definitions of independence are equivalent.

Exercise 2 A, B are two real variables. Assume that A, B are jointly Gaussian. Give a necessary and sufficient condition for $A \perp B$ in this case.

Conditional independence

A richer and more useful concept is the *conditional independence* between sets of random variables.

[1] An information theoretical quantity that measures how much information random variable A has about random variable B.

$$\boxed{A \perp B \mid C \iff P_{AB|C} = P_{A|C}.P_{B|C} \iff P_{A|BC} = P_{A|C}}$$

Once C is known, knowing B brings no additional information about A. Here are a few more equivalent ways to express conditional independence. The expressions below must hold for every a, b, c for which the respective conditional probabilities are defined.

$$\begin{aligned}
P_{ABC}(a, b|c) &= P_{A|C}(a|c)P_{B|C}(b|c) \\
\frac{P_{ABC}(a, b, c)}{P_C(c)} &= \frac{P_{AC}(a, c)P_{BC}(b, c)}{P_C(c)^2} \\
P_{ABC}(a, b, c) &= \frac{P_{A|C}(a|c)P_{B|C}(b|c)}{P_C(c)} \\
\frac{P_{ABC}(a, b, c)}{P_{BC}(b, c)} &= \frac{P_{A|C}(a, c)}{P_C(c)} \\
P_{A|BC}(a|b, c) &= P_{A|C}(a|c) \\
P_{ABC}(a, b, c) &= P_C(c)P_{A|C}(a|c)P_{B|C}(b|c)
\end{aligned}$$

Since independence between two variables implies the joint probability distribution factors, it implies far fewer parameters are necessary to represent the joint distribution ($r_A + r_B$ rather than $r_A r_B$), and other important simplifications such as $A \perp B \Rightarrow E[AB] = E[A]E[B]$ [**Exercise 3** Prove this relationship.]

These definitions extend to sets of variables:

$$AB \perp CD \equiv P_{ABCD} = P_{AB}P_{CD}$$

and

$$AB \perp CD|EF \equiv P_{ABCD|EF} = P_{AB|EF}P_{CD|EF}.$$

Furthermore, independence of larger sets of variables implies independence of subsets (for fixed conditions):

$$AB \perp CD|EF \Rightarrow A \perp CD|EF, A \perp C|EF, \ldots$$

.

[**Exercise 4** Prove this.]

Notice that $A \perp B$ does *not* imply $A \perp B|C$, nor the other way around. [**Exercise 5** Find examples of each case.]

Two different factorizations of the joint probability $p(a, b, c)$ when $A \perp B|C$ are a good basis for understanding the two primary types of graphical models we will study: *undirected graphs*, also known as *Markov random fields (MRFs)*; and *directed graphs*, also known as *Bayes' nets (*BNs*)* and *belief networks.*

One factorization,

$$P_{ABC}(a, b, c) \;=\; \frac{P_{AC}(a, c)P_{BC}(b, c)}{P_C(c)} = \phi_{AC}(a, c)\phi_{BC}(b, c)$$

is into a product of *potential* functions defined over subsets of variables with a term in the denominator that compensates for "double-counting" of variables in the intersection. From this factorization an *undirected* graphical representation of the factorization can be constructed by adding an edge between any two variables that cooccur in a potential.

The other factorization,

$$P_{ABC}(a, b, c) = P_C(c)P_{A|C}(a|c)P_{B|C}(b|c),$$

is into a product of conditional probability distributions that imposes a partial order on the variables (e.g C comes before A, B). From this factorization a *directed* graphical model can be constructed by adding directed edges that match the "causality" implied by the conditional distributions. In particular, if the factorization involves $P_{X|YZ}$ then directed edges are added from Y and Z to X.

3 The (semi)-graphoid axioms

For any distribution P_V over a set of variables V the following properties of independence hold. Let X, Y, Z, W be disjoint subsets of discrete variables from V.

- **[S]** $X \perp Y \,|\, Z \;\Rightarrow\; Y \perp X \,|\, Z$ *(Symmetry)*
- **[D]** $X \perp YW \,|\, Z \;\Rightarrow\; X \perp Y \,|\, Z$ *(Decomposition)*
- **[WU]** $X \perp YW \,|\, Z \;\Rightarrow\; X \perp Y \,|\, WZ$ *(Weak union)*
- **[C]** $X \perp Y \,|\, Z$ and $X \perp W \,|\, YZ \;\Rightarrow\; X \perp YW \,|\, Z$ *(Contraction)*
- **[I]** If P is strictly positive for all instantiations of the variables,
  $X \perp Y \,|\, WZ$ and $X \perp W \,|\, YZ \;\Rightarrow\; X \perp YW \,|\, Z$ *(Intersection)*

Properties S, D, WU, C are called the **semi-graphoid axioms**. The semi-graphoid axioms together with property [I] are called the **graphoid axioms**.