Discrete Probability

Throughout the study of graphical models, we will make use of some basic facts about discrete probability distributions. Here, we review the most important definitions and examples that will be useful for this study.

A **sample space**, which we will denote by Ω, is a set of possible outcomes of a random process. As an example, $\Omega = \{H, T\}$ could represent the two possible outcomes (heads and tails) of a coin toss. Each element ω in the sample space is associated with a nonnegative number $p(\omega) \in [0, 1]$ corresponding to the **probability** that the outcome ω occurs. We require that these probabilities sum to one.

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

might have that $p(H) = .6$ and $p(T) = .4$.

An **event** is a subset of the sample space. The probability of an event is equal to the sum of the probabilities of the outcomes contained in that event. Let's consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ of all possible outcomes of a fair die roll. $A = \{1, 5, 6\} \subseteq \Omega$ would correspond to the event that the die came up as either a one, a five, or a six. The probability of A, denoted $p(A)$, is then equal to $p(1) + p(5) + p(6) = 1/2$ for a fair die.

Independent Events

Two events $A, B \subseteq \Omega$ are said to be **independent** if

$$p(A \cap B) = p(A)p(B).$$

That is, the occurence of event A does not affect the probability that event B occurs and vice versa.

> all possible outcomes resulting from rolling a fair die.
> If $A = \{1, 2, 5\}$ and $B = \{3, 4, 6\}$ are A and B indpendent?

> Let Ω be the set of all possible outcomes resulting from rolling two fair dice.
> If
> $A = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$
> and
> $B = \{(1,6), (2,6), (3,6), (4,6), (5,6), (6,6)\}$
> are A and B indpendent?

Notationally, we will often write $A \perp B$ to mean that A is independent of B. As independence is a symmetric property, $A \perp B$ is equivalent to $B \perp A$.

Conditional Probability

The **conditional probability** of an event A given an event B such that $p(B) > 0$ is given by Bayes' Rule.

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

new sample space $\Omega' = B$ and constructing a probability distribution over this sample space where $p_{\Omega'}(\omega') = \frac{p(\omega')}{p(B)}$ for each $\omega' \in \Omega'$. We can check that this corresponds to a proper probability distribution over Ω' by verifying that $\sum_{\omega' \in \Omega'} p(\omega')/p(B) = 1$. The conditional probability of A given B is then the probability of the event $A \cap B$ in the sample space Ω' with respect to $p_{\Omega'}$.

> If A and B are independent events such that $p(B) > 0$, show that $p(A|B) = p(A)$.

> Show that $\sum_{\omega \in \Omega} p(\omega|B) = 1$ whenever $p(B) > 0$.

The Chain Rule

Bayes' rule, for the events A and B with $p(B) > 0$, can be equivlaently expressed as

$$p(A \cap B) = p(A|B) \cdot p(B).$$

the probability that all of these events occur has a similar expression.

$$p\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} p\left(A_i \middle| \bigcap_{j=1}^{i-1} A_j\right)$$

This expresion is known as the chain rule and can be derived by repeated application of Bayes' rule (consider applying Bayes' rule to the events A_1 and $\cap_{i=2}^{n} A_i$).

Conditional Independence

Two events A and B are said to be conditionally independent given a third event C such that $p(C) > 0$ if

$$p(A \cap B|C) = p(A|C)p(B|C)$$

or, equivalently, $p(A|B, C) = p(A|C)$ and $p(B|A, C) = p(B|C)$. Conditional independence is usually denoted by writing $A \perp B|C$.

Conditioning does not preserve independence. Two dependent events conditioned on

events can become dependent after conditioning on a third event. This seems somewhat surprising, but consider the following example. Let $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ be the possible outcomes of two independent, fair coin tosses. Let A be the event that the first coin comes up heads. Let B be the event that the second coin comes up heads. Finally, let C be the event that both coins come up heads or both coins come up tails. We can easily check that A and B are independent, but A and B are not conditionally independent given C.

Random Variables

A **discrete random variable**, X, is a function from the state space Ω into a discrete space D. For each $x \in D$,

$$p(X = x) = p(\{\omega \in \Omega | X(\omega) = x\})$$

is the probability that the random variable X takes the value x. Note that

ion the state

$A_x \triangleq \{\omega \in \Omega | X(\omega) = x\}$ for each $x \in D$. As a result,

$$\sum_{x \in D} p(X = x) = \sum_{x \in D} p(A_x) = 1.$$

We say that $p(X)$ is the probaility distribution corresponding to the random variable X.

Random variables are useful when describing the outcome of a random process as they allow us to carve up the sample space into meaningful events. For example, let Ω be the set of all possible pairs of outcomes obtained by rolling two fair dice, that is $p(1, 2) = 1/36$ is the probability that the first die is showing a one and the second die is showing a two. Define the random variable $X(\omega)$ to be the sum of the numbers showing on the two dice.

> What is $p(X = 2)$?

> What is $p(X = 8)$?

joint probability distribution over all possible assignments of a collection of random variables $X_1, \ldots, X_n$ denoted $p(X_1 = x_1 \cap \ldots \cap X_n = x_n)$. The joint probability distribution is often written as $p(X_1 = x_1, \ldots, X_n = x_n)$ or more compactly as $p(x_1, \ldots, x_n)$ when it is understood that the random variable X_i is assigned the value x_i. As random variables partition the sample space into events, all of the results discussed above apply in this context as well: Bayes' rule, conditioning, the chain rule, etc. Two random variables X_1 and X_2 are independent if

$$p(X_1 = x_1, X_2 = x_2) = p(X_1 = x_1)p(X_2 = x_2)$$

for *all* possible assignments x_1 and x_2. By using Bayes' rule, we can formuate the analogous definition of conditional independence.

Computational Issues

From a computational perspective, two

nary concern:

distribution and performing statistical inference. First, in order to use probability distributions as part of a computation of any kind, they will need to be represented. The most general way to represent a joint probability distribution is as a mapping from each assignment of the random variables to probabilities. This can be represented as a table with one entry for each of the possible values of each of the random variables in the joint distribution. In practice, this can be computationally prohibitive. For example, consider a joint probability distribution over n random variables $X_1, \ldots, X_n$ each of which takes values in the set D. In the worst case, the corresponding table would necessarily have $|D|^n - 1$ entries, one for each of the different possible assignments to the random variables minus one (the probability of the last assignment is equal to one minus the sum of the rest of the probabilities). Fortunately, the

al case.

> the joint probability distribution of n independent random variables each taking values in the set D in the worst case?

> Consider a collection of random variables $X_1, \ldots, X_n$ each taking values in the set D such that for $i > 2$, X_i is independent of $X_1, \ldots, X_{i-2}$ given X_{i-1}. How large of table is needed to store this joint probability distribution in the worst case?

Given a joint probability distribution we will also be interested in performing **statistical inference**. One common type of statistical inference will involve computing **marginal distributions**, i.e., computing $p(x)$ from $p(x, y)$. Marginal distributions are formed by summing out over all possible values of a subset of the random variables in a joint probability distribution while the others remain fixed. For example, consider a joint probability distribution over n random

the set D. To compute the marginal distribution over the variable X_1, we would compute $p(x_1) = \sum_{x_2,\ldots,x_n \in D} p(x_1, \ldots, x_n)$. As before, this computation could require $|D|^{n-1}$ operations in the worst case. However, if the joint distribution satisfies certain properties, the computational cost can be significantly reduced. Again, the simplest example is given by a joint probability distribution over independent random variables.

Genearlly speaking, more independence means easier computation and less storage. We want models of probability distributions that somehow make the underlying independence assumptions explicit so that we can take advantage of them (checking for all of the possible independence relationships is itself a computationally challenging problem). Handling, either exactly or approximately, the above computational issues will be the primary thrust of this book. In the coming chapters, we will

connection between conditional independence and the complexity of representation and inference. This, in turn, will lead us to a useful discussion about how to design efficient algorithms for exact and approximate statistical inference in practice.

A Motivating Example

We conclude this chapter with an example of the kinds of probability distributions that we will encounter in our study of graphical models. An undirected graph $G = (V, E)$ is a collection of vertices and edges (pairs of vertices). An independent set in a graph G is a subset $S \subseteq V$ of the vertices such that no two vertices in S are joined by an edge of G. Let Ω be the set of all subsets of the vertex set V, and let p be the uniform probability distribution over independent sets of G. For each vertex $v \in V$, define the random variable $X_v(\omega)$ to be equal to 1 if $v \in \omega$ and 0 otherwise. X_v is then equal

set in G contains the vertex v. If $V = \{v_1, \ldots, v_n\}$, the joint probability distribution $p(x_{v_1}, \ldots, x_{v_n})$ is nonzero if and only if $S = \{v \in V | x_v = 1\}$ is an independent set in G.

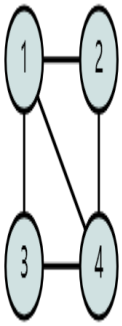

Figure 1. A graph with vertex set $V = \{1, 2, 3, 4\}$ and edge set $E = \{(1, 2), (2, 4), (1, 3), (3, 4), (1, 4)\}$.

> Consider the undirected graph in Figure 1. What is $p(X_1 = 1)$ under the uniform distribution on indepedent sets described above?

These types of probability distrbutions over combinatorial structures arise naturally in machine learning, artificial intelligence, statistical physics, computer vision, and many other application areas.