Intro to probability: outcomes

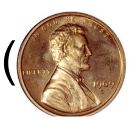
- An **outcome space** specifies the possible outcomes that we would like to reason about, e.g.

Ω = {  ,  } Coin toss

Ω = { , , , , ,  } Die toss

- We specify a **probability** p(**x**) for each outcome **x** such that

$$p(x) \geq 0, \qquad \sum_{x \in \Omega} p(x) = 1$$

E.g., p() = .6

p() = .4

Intro to probability: events

- An **event** is a subset of the outcome space, e.g.

E = {  **}**   Even die tosses

O = {  **}**   Odd die tosses

- The **probability** of an event is given by the sum of the probabilities of the outcomes it contains,

$$p(E) = \sum_{x \in E} p(x)$$

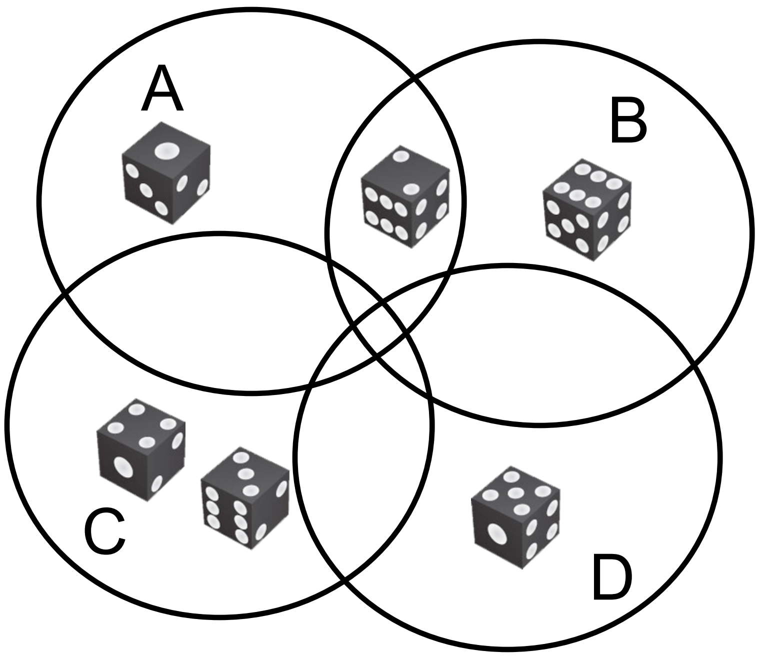
E.g., p(E) = p() + p() + p()

= 1/2, if fair die

Intro to probability: union bound

- P(A or B or C or D or …)

  $\leq$ P(A) + P(B) + P(C) + P(D) + …



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$\leq p(A) + p(B)$$

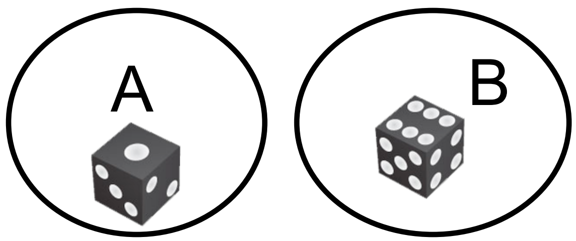
Q: When is this a tight bound?

A: For disjoint events
(i.e., non-overlapping circles)

Intro to probability: independence

- Two events A and B are **independent** if

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



Are these events independent?

No! $p(A \cap B) = 0$

$$p(A)p(B) = \left(\frac{1}{6}\right)^2$$

Intro to probability: independence

- Two events A and B are **independent** if

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

Analogy: outcome space defines all possible sequences of e-mails in training set

- Suppose our outcome space had two different die:

$$\Omega = \{$$  $$\}$$

2 die tosses

$6^2 = 36$ outcomes

and the probability of each outcome is defined as

p() = $a_1\, b_1$ p() = $a_1\, b_2$ $\cdots$

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

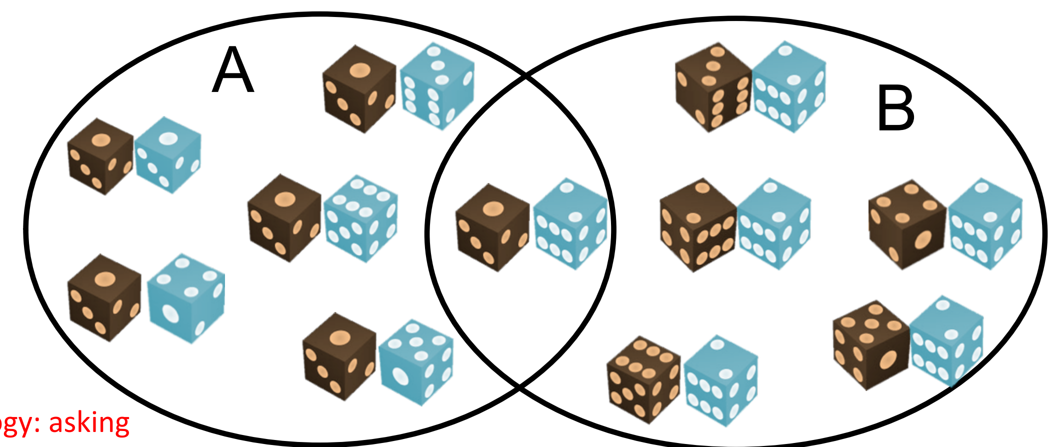
$$\sum_{j=1}^{6} b_j = 1$$

Intro to probability: independence

- Two events A and B are **independent** if

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

- Are these events independent?



Yes! $p(A \cap B) =$ p( )

$p(A)p(B) =$ p( ) p( )

Analogy: asking about first e-mail in training set

p(A) = p( )

$$= \sum_{j=1}^{6} a_1 b_j = a_1 \sum_{j=1}^{6} b_j = a_1$$

p(B) = p( ) $= b_2$

Analogy: asking about second e-mail in training set

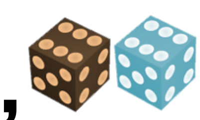
Intro to probability: discrete random variables

- A **random variable** X is a mapping $X : \Omega \rightarrow D$
  - D is some set (e.g., the integers)
  - Induces a partition of all outcomes Ω
- For some $x \in D$, we say

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

"probability that variable X assumes state x"

- Notation: $\text{Val}(X)$ = set D of all values assumed by X (will interchangeably call these the "values" or "states" of variable X)

$\mathbf{\Omega = \{}$ , , , $\cdots$,  $\mathbf{\}}$ 2 die tosses

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Random variables have domains
  - R in {true, false}   (sometimes write as {+r, ¬r})
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$
- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice
- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = a_1 b_1 p() = a_1 b_2 …

X_1 and X_2 NOT identically distributed

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

$$\sum_{j=1}^{6} b_j = 1$$

Ω = { , , , …,  **}** 2 die tosses

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$

- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice

- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = a_1 a_1 p() = a_1 a_2 $\cdots$

X_1 and X_2 identically distributed

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

$\Omega = \{$ , , , $\cdots$,  $\}$ 2 die tosses

Intro to probability: discrete random variables

- X=x is simply an event, so can apply union bound, etc.
- Two random variables **X** and **Y** are **independent** if:

$$p(X = x, Y = y) = p(X = x)p(Y = y) \quad \forall x \in \mathrm{Val}(X), y \in \mathrm{Val}(Y)$$

Joint probability. Formally, given by the event $X = x \cap Y = y$

- The **expectation** of **X** is defined as: $E[X] = \sum_{x \in \mathrm{Val}(X)} p(X = x)x$

- If **X** is binary valued, i.e. x is either 0 or 1, then:

$$\begin{aligned} E[X] &= p(X = 0) \cdot 0 + p(X = 1) \cdot 1 \\ &= p(X = 1) \end{aligned}$$

- Linearity of expectations: $E\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} E[X_i]$

PAC bound and Bias-Variance tradeoff

for all h, with probability at least 1-δ:

$$\text{error}_{true}(h) \leq \underbrace{\text{error}_D(h)}_{\text{“bias”}} + \underbrace{\sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}}_{\text{“variance”}}$$

- For large |H|
  - low bias (assuming we can find a good h)
  - high variance (because bound is looser)
- For small |H|
  - high bias (is there a good h?)
  - low variance (tighter bound)

Probability Distributions

- Discrete random variables have distributions

$P(T)$

T	P
warm	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A discrete distribution is a TABLE of probabilities of values
- The probability of a state (lower case) is a single number

$$P(W = rain) = 0.1 \qquad P(rain) = 0.1$$

- Must have:

$$\forall x \, P(x) \geq 0 \qquad \sum_x P(x) = 1$$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - How many assignments if *n* variables with domain sizes *d*?
  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- For all but the smallest distributions, impractical to write out or estimate
  - Instead, we make additional assumptions about the distribution

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_w P(t,w)$$

$$P(w) = \sum_t P(t,w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

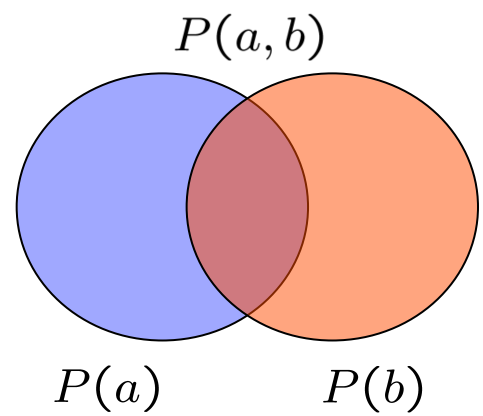
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=r|T=c) = ???$$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \Leftrightarrow \quad P(x,y) = P(x|y)P(y)$$

- Example:

$P(W)$

W	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Leftrightarrow$

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Let's us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many practical systems (e.g. ASR, MT)

- In the running for most important ML equation!