Computer Science

CSC 480/580:Principles of Machine Learning

Probability review & HW0 review

Chicheng Zhang

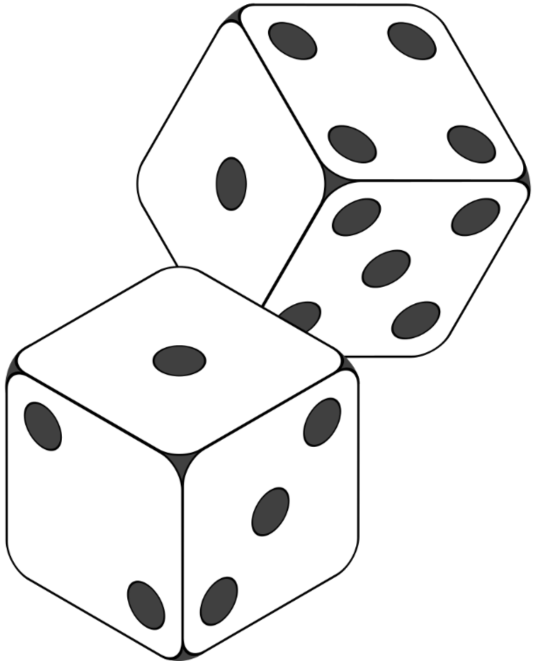
Administrivia

- Homework submission
  - Make sure questions are answered in PDF
  - Match pages to questions
  - Put code in PDF (relevant parts of code at least)
  - Doublecheck your submission

Outline

- Probability Refresher
- HW0 Review

Random Events and Probability



Suppose we roll two fair dice…

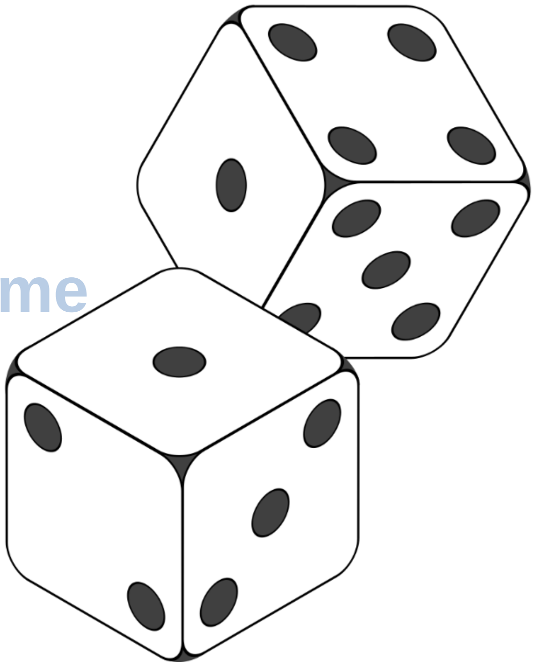
- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?
- What is the *probability* of rolling **odd** numbers?

…probability theory *gives a mathematical formalism to addressing such questions…*

Definition An **experiment** or **trial** is any process that can be repeated with well-defined outcomes. It is *random* if more than one outcome is possible.

Random Events and Probability

Definition An **outcome** is a possible result of an experiment or trial, and the collection of all possible outcomes is the **sample space** of the experiment,



Outcome

Example (1,1), (1,2), …, (6,1), (6,2), …, (6,6)

Sample Space

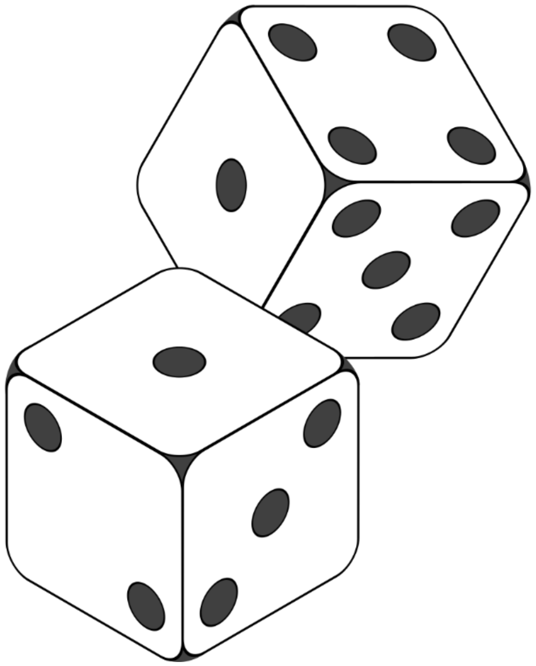
Definition An **event** is a *set* of outcomes (a subset of the sample space),

Example Event Roll at least a single 1

{(1,1), (1,2), (1,3), …, (1,6), …, (6,1)}

Random Variables

(Informally) A random variable maps outcomes to numeric values.



Example X is the *sum of two dice* with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

Example Flip a coin and let random variable Y represent the outcome,

$$Y \in \{\text{Heads}, \text{Tails}\}$$

Random Variables and Probability

Capitol letters represent random variables

Lowercase letters are realized *values*

$$X = x$$

$X = x$ is the **event** that X takes the value x

Example Let X be the random variable (RV) representing the sum of two dice with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

X=5 is the *event* that the dice sum to 5.

$$\{X = 5\} = \{(1,4), (2,3), (3,2), (4,1)\}$$

Probability Mass Function

A function $p(X)$ is a **probability mass function (PMF)** of a discrete random variable X, if the following conditions hold:

(a) It is nonnegative for all values in the support,

$$p(X = x) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x p(X = x) = 1$$

Intuition Probability mass is conserved, just as in physical mass. Reducing probability mass of one event must increase probability mass of other events so that the definition holds...

Probability Mass Function

Example Let X be the outcome of a single fair die. It has the PMF,

$$p(X = x) = \frac{1}{6} \qquad \text{for } x = 1, \ldots, 6$$

Uniform Distribution

Example We can often represent the PMF as a vector. Let S be an RV that is the *sum of two fair dice*. The PMF is then,

Observe that S does not follow a uniform distribution

$$p(S) = \begin{pmatrix} p(S=2) \\ p(S=3) \\ p(S=4) \\ \vdots \\ p(S=12) \end{pmatrix} = \begin{pmatrix} 1/36 \\ 1/18 \\ 1/2 \\ \vdots \\ 1/36 \end{pmatrix}$$

PMF Notation

- We will use $p(X)$ to refer to the probability mass *function* of the RV X
- We use $p(X=x)$ to refer to the probability of the *outcome* $X=x$ (also called an “event”)
- We will often use $p(x)$ as shorthand for $p(X=x)$

Joint Probability

Definition Two (discrete) RVs X and Y have a *joint PMF* denoted by $p(X, Y)$ and the probability of the event X=x and Y=y denoted by $p(X = x, Y = y)$ where,

(a) It is nonnegative for all values in the support,

$$p(X = x, Y = y) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x \sum_y p(X = x, Y = y) = 1$$

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

All values are nonnegative

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

The sum over all values is 1:
0.04 + 0.36 + 0.30 + 0.30 = 1

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y	
		0	1
X	0	0.04	0.36
	1	0.30	0.30

P(X=1, Y=0) = 0.30

Fundamental Rules of Probability

Given two RVs X and Y the **conditional distribution** is:

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\sum_x p(X{=}x,Y)}$$

Multiply both sides by $p(Y)$ to obtain the **probability chain rule**:

$$p(X, Y) = p(Y)p(X \mid Y)$$

The probability chain rule extends to N RVs $X_1, X_2, \ldots, X_N$:

$$p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2 \mid X_1) \ldots p(X_N \mid X_{N-1}, \ldots, X_1)$$

$$= p(X_1) \prod_{i=2}^{N} p(X_i \mid X_{i-1}, \ldots, X_1)$$

Chain rule valid for any ordering

Fundamental Rules of Probability

Law of total probability

$$p(Y) = \sum_x p(Y, X = x)$$

- p(Y) is a **marginal** distribution
- This is called **marginalization**

Proof

$$\sum_x p(Y, X = x) = \sum_x p(Y) p(X = x \mid Y) \quad \text{(chain rule)}$$

$$= p(Y) \sum_x p(X = x \mid Y) \quad \text{(distributive property)}$$

$$= p(Y) \quad \text{(PMF sums to 1)}$$

Generalization for conditionals:

$$p(Y \mid Z) = \sum_x p(Y, X = x \mid Z)$$

Tabular Method

Let X, Y be binary RVs with the joint probability table

For X, Y that can take K values, use K-by-K probability table.

		Y: y_1	Y: y_2	$P(x)$
X	x_1	0.04	0.36	0.4 ← $P(x_1)$
X	x_2	0.30	0.30	0.6 ← $P(x_2)$
$P(y)$		0.34 ← $P(y_1)$	0.66 ← $P(y_2)$	

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
[i.e., sum down columns]

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
$P(x_2)=P(x_2,y_1)+P(x_2,y_2)$
[i.e., sum across rows]

Tabular Method

We don't care about event $Y=y_2$

Y

X		y_1	y_2
	x_1	0.04	Censored!
	x_2	0.30	
		0.34	

0.34 ← $P(y_1)$

$P(x|y_1)=?$

Tabular Method

		$Y=y_1$
X	x_1	0.04
	x_2	0.30

0.34

$P(y_1)$

→

0.04 / 0.34

0.30 / 0.34

$P(x|y_1)$

These sum to one:
A conditional probability distribution is still a probability distribution

Intuition Check

<u>Question:</u> Roll two dice and let their outcomes be $X_1, X_2 \in \{1, \ldots, 6\}$ *for die 1 and die 2, respectively. Recall the definition of conditional probability,*

$$p(X_1 \mid X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Which of the following are true?

a) $p(X_1 = 1 | X_2 = 1) > p(X_1 = 1)$

b) $p(X_1 = 1 | X_2 = 1) = p(X_1 = 1)$ Outcome of die 2 doesn't *affect* die 1

c) $p(X_1 = 1 | X_2 = 1) < p(X_1 = 1)$

Intuition Check

Question: *Let* $X_1 \in \{1, \dots, 6\}$ *be outcome of die 1, as before. Now let* $X_3 \in \{2, 3, \dots, 12\}$ *be the* sum of both dice. Which of the following are true?

a) $p(X_1 = 1 | X_3 = 3) > p(X_1 = 1)$

b) $p(X_1 = 1 | X_3 = 3) = p(X_1 = 1)$

c) $p(X_1 = 1 | X_3 = 3) < p(X_1 = 1)$

Only 2 ways to get $X_3 = 3$ *, each with equal probability:*

$(X_1 = 1, X_2 = 2)$ *or* $(X_1 = 2, X_2 = 1)$

so

$$p(X_1 = 1 \mid X_3 = 3) = \frac{1}{2} > \frac{1}{6} = p(X_1 = 1)$$

Dependence of RVs

Intuition…

Consider $P(B|A)$ where you want to bet on B

Should you pay to know A?

In general you would pay something for A if it changed your belief about B. In other words if,

$$P(B|A) \neq P(B)$$

Independence of RVs

Definition *Two random variables X and Y are <u>independent</u> if and only if,*

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for all values x and y, and we say $X \perp Y$.

- Shorthand: $p(X,Y) = p(X)\,p(Y)$
- Equivalent definition of independence: $p(X \mid Y) = p(X)$

Shorthand notation
Implies for all *x, y*

Definition RVs $X_1, X_2, \ldots, X_N$ are <u>mutually independent</u> if and only if,

$$p(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} p(X_i = x_i)$$

Independence of RVs

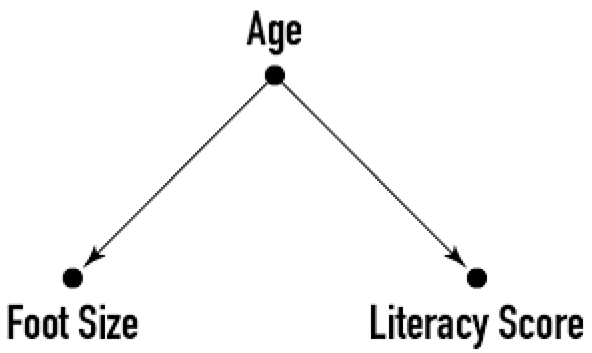
Definition *Two random variables* X *and* Y *are* *<u>conditionally independent</u>* *given* Z *if and only if,*

$$p(X = x, Y = y \mid Z = z) = p(X = x \mid Z = z)p(Y = y \mid Z = z)$$

for all values x, y, and z, and we say that $X \perp Y \mid Z$.

Shorthand notation
Implies for all *x, y, z*

- Shorthand: $p(X, Y \mid Z) = p(X \mid Z)\, p(Y \mid Z)$
- Equivalent defn of conditional independence: $p(X \mid Y, Z) = p(X \mid Z)$



Outline

- Probability Refresher
- HW0 Review

Problem 1

1 Vectors and Matrices

Consider the matrix X and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$X = \begin{pmatrix} 1 & -1 \\ -2 & 2 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

1. Compute $\mathbf{y}^\top X \mathbf{z}$.

2. Is X invertible? If so, give the inverse, and if no, explain why not.

- Can we verify that X is invertible, without calculating its determinant?

Problem 2

$$y = e^{x^2} + \tan(z)x^{6z} - \ln\left(\frac{8x+16}{x^4}\right)$$

- How to easily compute the derivative of the third term with respect to x?
- Observation: $\ln\left(\frac{8x+16}{x^4}\right) = \ln(8) + \ln(x+2) - 4\ln x$

Problem 3

- Sequence of coin flip S = (0,1,1,0,0,1,1)

- $F(p)$: Probability of observing this sequence, assuming that the coin has bias p
  - $(1-p) * p * p * (1-p) * (1-p) * p * p$
  - $= p^4(1-p)^3$

- Should it have binomial coefficient $\binom{7}{4}$?

- How to compute the maximizer of $F(p)$?
  - Find a point p such that $F'(p) = 0$. Are we done?

Problem 3 (cont'd)

$$P(A = 1 \vee B = 1)$$

- What outcomes does this event contain?

$$P(A = 1 | B = 0)$$

- What steps shall we take to compute this?

a	b	$P(A = a, B = b)$
0	0	0.4
0	1	0.3
1	0	0.2
1	1	0.1

Problem 4

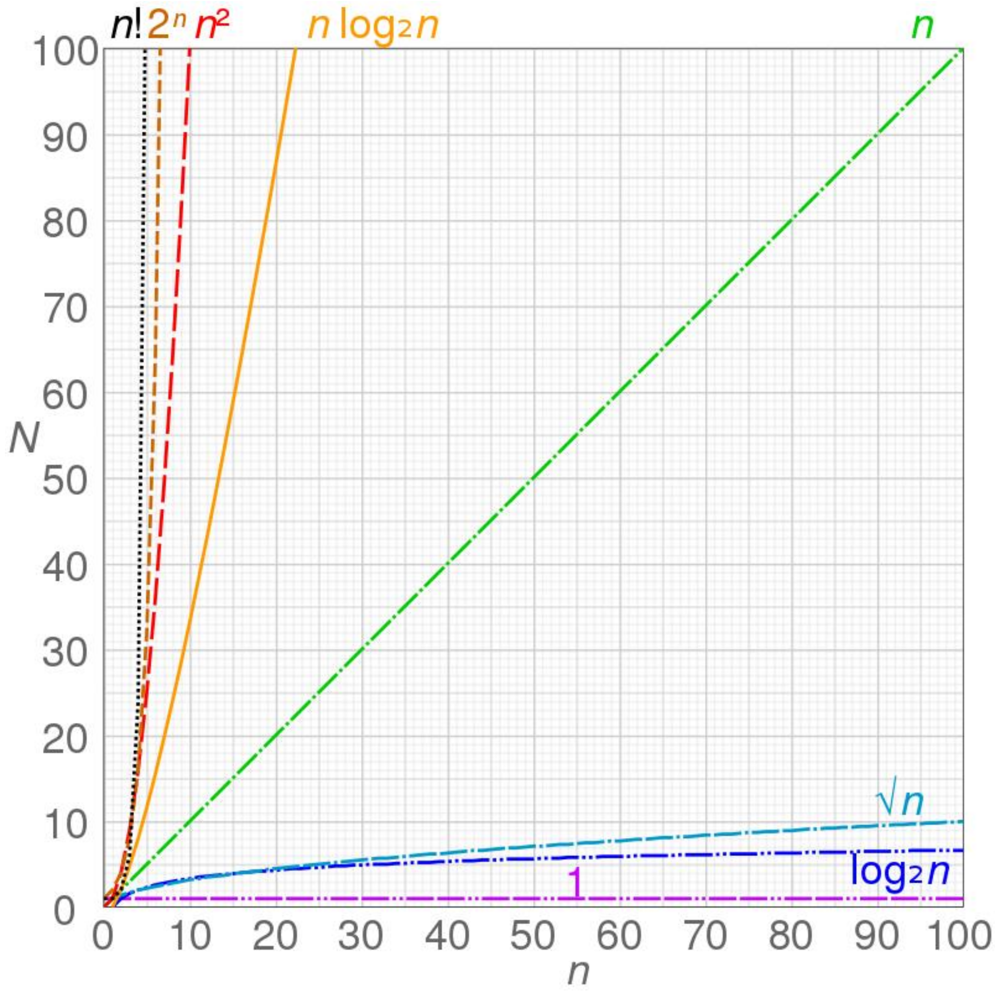
- Intuition: $f(n) = O(g(n))$ if f grows no faster than g (as n grows), up to constant factors

- $\ln(n)$ vs. $\log_2 n$ -- the latter grows faster -- $\log_2 n = \ln n \cdot \log_2 e$
  - Does this imply that $\log_2 n \neq O(\ln n)$?

- Note: $O(f(n)) = O(g(n))$ does not parse

Problem 4

- https://en.wikipedia.org/wiki/Big_O_notation



Problem 5

- 5.1: in counterexample constructions, need to specify the probability of each outcome P({1}), P({2}), etc
  - If using uniform distribution, this needs to be declared explicitly
- 5.2: binomial distribution vs. multinomial distribution
  - Flip n coins vs. flip n 6-sided dice
- 5.3 $\mathrm{Var}(3X) = 3^2\,\mathrm{Var}(X)$
  - Intuition: variance measures the average *squared deviation* of a r.v. around its mean

Problem 5 (cont'd)

- 5.4.3(b)

(b) Suppose I rolled two dice independently, and I tell you that the sum of the outcomes of the two dice are an even number (but I do not tell you the outcomes of the two dice). Given this information, is the outcome of the second die independent of the outcome of the first die? Prove or disprove.

- How to formalize the argument using math language?

No. Let X_1 and X_2 denote the outcome of the first and the second die, respectively. Let E denote the event that the sum of the two outcomes are an even number.
Random variables X_1 and X_2 are said to be independent given E, if and only if for any i, j, k, $\mathbb{P}(X_2 = i \mid X_1 = j, E) = \mathbb{P}(X_2 = i \mid X_1 = k, E)$.
However, in the above two-dice example, $\mathbb{P}(X_2 = 1 \mid X_1 = 1, E) = \frac{1}{3} \neq 0 = \mathbb{P}(X_2 = 1 \mid X_1 = 1, E)$.

Problem 6

Prove: The smallest Euclidean distance from the origin to some point $\mathbf{x}$ in the hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ is $\frac{|b|}{\|\mathbf{w}\|_2}$. You may assume $\mathbf{w} \neq 0$.

- Idea:
  - First, guess a point x_0 on the hyperplane
  - Second, prove that x_0 has the smallest distance to the origin, among all points on the hyperplane

Problem 7

3. Make a scatterplot by drawing 100 samples from a mixture distribution $0.3 \cdot \mathcal{N}\left((1,0)^\top, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right) +$
$0.7 \cdot \mathcal{N}\left((-1,0)^\top, \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}\right)$.

- Definition: see e.g. https://en.wikipedia.org/wiki/Mixture_distribution

- One candidate solution: draw 30 samples from the first normal distn, and 70 samples from the second normal distn

- Is this the right approach? What if we are asked to draw 1 sample?