Welcome !! :)

Plan

Logistics

Probability Review

Problem Solving

Set Size Estimation

go/rads2024/

MTWR

↳ 10-12 lecture

↳ 2-3 problem solving

3-4 office hours

rwitter@midd

Please come!

Please no computers!

Recommend: Read notes, come to class, read notes again

Grades

Participation (14)

1 pt per day

Problem Set (56)

one problem per class

- 3 pts for solution
- 1 pt self-grade

Project (30)

Linear algebra ♡ probability
♡ algorithms

Randomized Algorithms

So. much. data. Every day...

↳ NASA creates 20 terabytes

↳ 8 billion google searches

↳ 300 million terabytes internet

Moore's Law says compute doubles

But we're hitting physical limit

Processing data requires speed

Randomized lets us work in sublinear time

↳ Estimate # unique items?

↳ Similarity search

↳ Process matrices

↳ Reconstruct measurements

Probability

X is a random variable

If we flip a coin

$$X = \begin{cases} 1 & \text{wp } 1/2 \\ 0 & \text{wp } 1/2 \end{cases}$$

If we roll a die

$$X = \begin{cases} 5 \cdot 1 & \text{wp } 1/6 \\ 5 \cdot 2 & \text{wp } 1/6 \\ 5 \cdot 3 & \vdots \\ 5 \cdot 4 & \\ 5 \cdot 5 & \\ 5 \cdot 6 & \end{cases}$$

$Pr(X = x)$ is the probability that X takes value x

$$Pr(\underset{\uparrow \text{die}}{X} = 4) = 1/6$$

$\mathbb{E}[X]$ is the expected value of X

$$\mathbb{E}[X] = \sum_x x \cdot Pr(X = x)$$

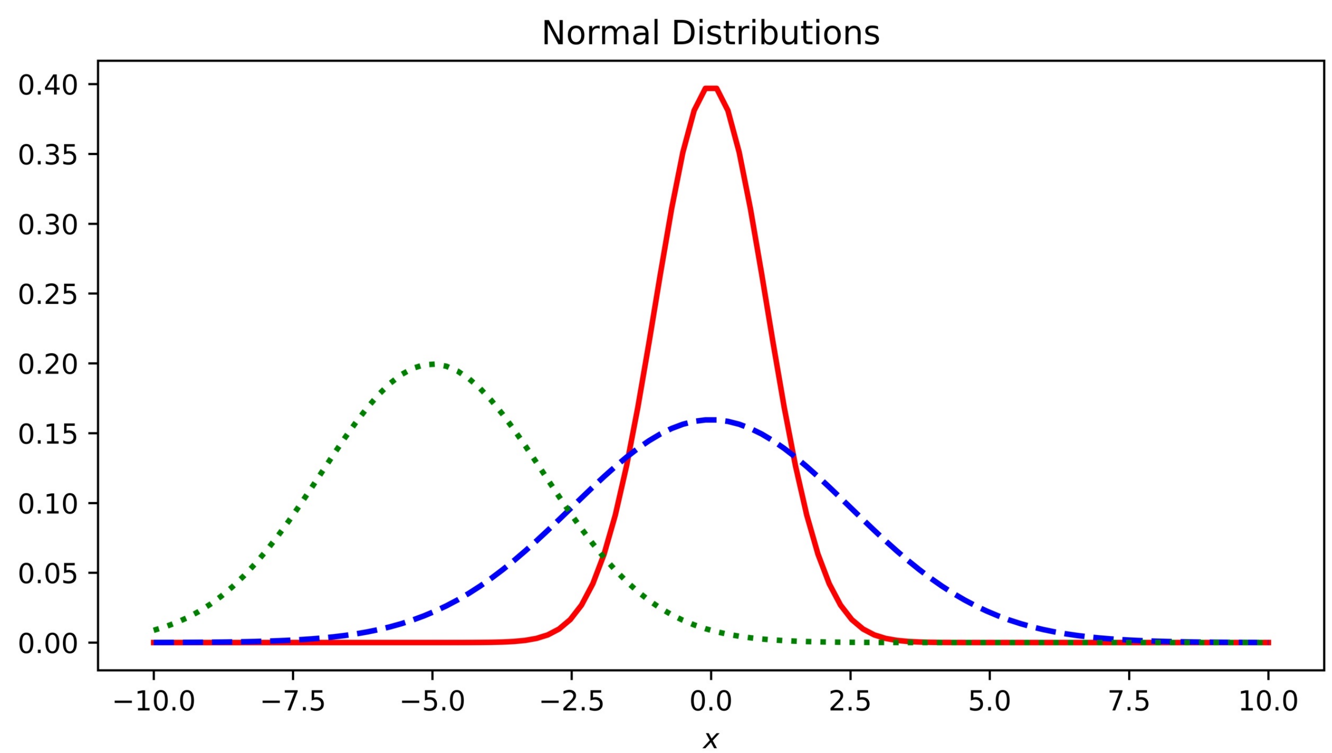
$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$= \frac{1}{6}(21) = 3.5$$

$\mathrm{Var}(X)$ is the variance of a random variable: how much it varies from its expectation

positive

$$\mathrm{Var}(X) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$



$$\mathbb{E}[cX] = \sum_x x \cdot c \Pr(X = x)$$

$$= c \sum_x x \cdot \Pr(X = x)$$

$$= c \cdot \mathbb{E}[X]$$

$$\mathrm{Var}(c \cdot X) = \mathbb{E}\left[\left(cX - \mathbb{E}[cX]\right)^2\right]$$

$$= \mathbb{E}\left[\left(cX - c\,\mathbb{E}[X]\right)^2\right]$$

$$= \mathbb{E}\left[c^2 \left(X - \mathbb{E}[X]\right)^2\right]$$

$$= c^2 \,\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$

$$= c^2 \,\mathrm{Var}(X)$$

Events defined on rvs

random variables

A = event that die is 1 or 2

B = event that die is odd

Does A give information about B?

$$Pr(A \cap B) = Pr(A) \cdot Pr(B \mid A)$$

$$\frac{1}{6} = \frac{2}{6} \cdot \frac{1}{2}$$

$$Pr(B \mid A) \triangleq \frac{Pr(A \cap B)}{Pr(A)}$$

A, B independant iff

$$Pr(B \mid A) = Pr(B)$$

$$\Leftrightarrow Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

X, Y are independent iff

$$Pr(X = x \mid Y = y) = Pr(X = x))$$

for all x, y

Linearity of Expectation

$\mathbb{E}[X+Y]$

always?
sometimes?
never?

$$= \sum_x \sum_y (x+y) \Pr(X=x \cap Y=y)$$

$$= \sum_x x \sum_y \Pr(X=x \cap Y=y) + \sum_y y \sum_x \Pr(X=x \cap Y=y)$$

$$= \sum_x x \Pr(X=x) \overbrace{\sum_y \Pr(Y=y \mid X=x)}^{1} + \sum_y y \cdot \Pr(Y=y) \cdot 1$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[XY] \stackrel{?}{=} \mathbb{E}[X]\,\mathbb{E}[Y]$$

always?
Sometimes?
never?

$$\mathbb{E}[XY] = \sum_{x,y} xy \Pr(X=x \wedge Y=y)$$

indep

$$= \sum_{x} \sum_{y} xy \Pr(X=x) \Pr(Y=y)$$

$$= \sum_{x} x \Pr(X=x) \sum_{y} y \Pr(Y=y)$$

$$= \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$X = \begin{cases} +1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases} \qquad Y = X$$

$$\mathbb{E}[X] = 1 \cdot 1/2 + -1 \cdot 1/2 = 0$$

$$\mathbb{E}[Y] = 0$$

$$\mathbb{E}[X \cdot Y] = 1/2 \cdot 1 \cdot 1 + 1/2 \cdot -1 \cdot -1$$

$$= 1$$

$$\mathrm{Var}(X) \stackrel{?}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\mathrm{Var}(X) \stackrel{\text{def}}{=} \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ when X, Y are independent

$$X = \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$$

$$Y = -X$$

$$\mathrm{Var}(X+Y) = \mathbb{E}[([X+Y] - \mathbb{E}[X+Y])^2]$$

$$= \mathbb{E}[((X-\mathbb{E}[X]) + (Y-\mathbb{E}[Y]))^2]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2 + 2(X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) + (Y-\mathbb{E}[Y])^2]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 0$$

$$\mathbb{E}[XY - \mathbb{E}[X]\cdot Y + \mathbb{E}[X]\cdot\mathbb{E}[Y] - X\mathbb{E}[Y]]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y] + \mathbb{E}[X]\cdot\mathbb{E}[Y] - \mathbb{E}[Y]\cdot\mathbb{E}[X]$$

$$\overset{\text{indep}}{=} \mathbb{E}[X]\mathbb{E}[Y]$$

$$X = \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$$

$$Y = -X$$

$$\mathbb{E}[X] = 1 \cdot 1/2 + -1 \cdot 1/2 = 0$$

$$\mathbb{E}[Y] = \mathbb{E}[-1 \cdot X] = -1 \cdot \mathbb{E}[X] = 0$$

$$Var(X) = \mathbb{E}[(X-0)^2] = 1/2 \cdot 1^2 + 1/2 \cdot (-1)^2$$
$$= 1$$

$$Var(Y) = Var(-1 \cdot X) = (-1)^2 \, Var(X)$$
$$= Var(X)$$

$$X + Y = \begin{cases} 1 + -1 & \text{wp } 1/2 \\ -1 + 1 & \text{wp } 1/2 \end{cases}$$
$$= 0 \quad \text{wp } 1$$

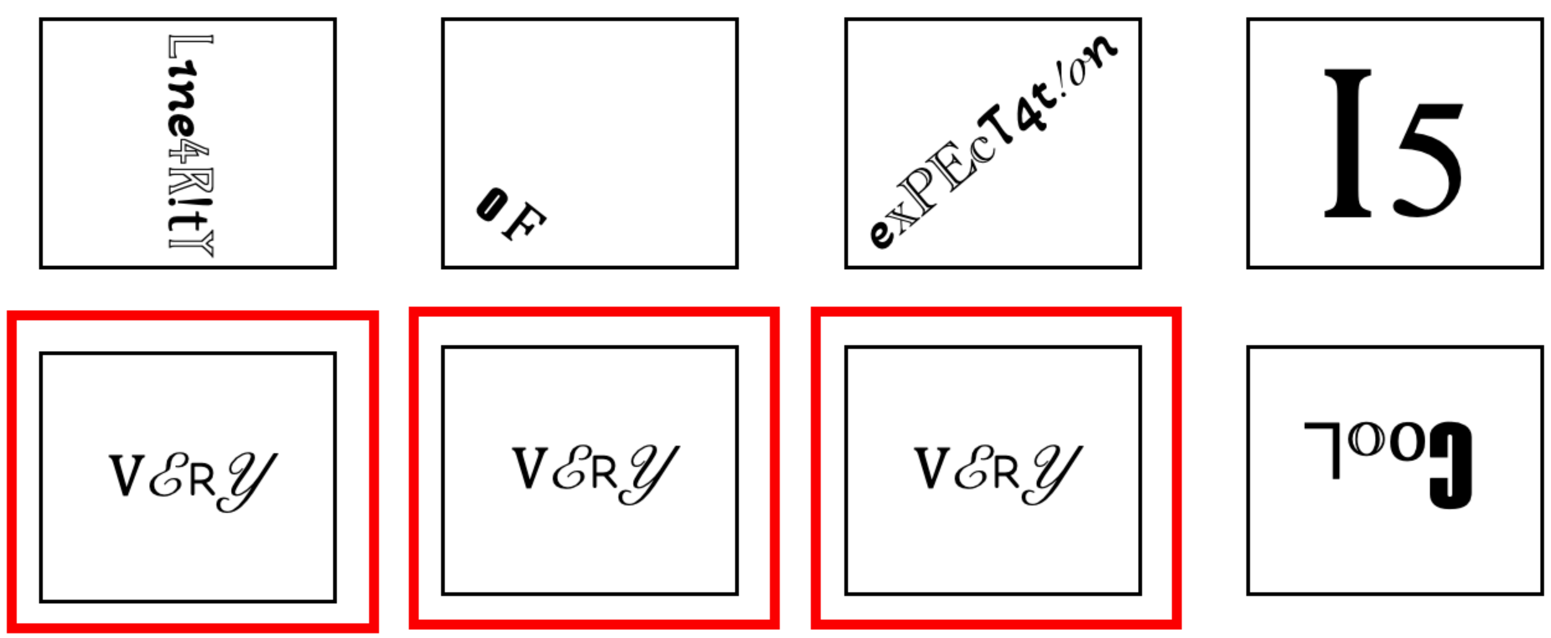
$$Var(X+Y) = 0 \neq 2 = Var(X) + Var(Y)$$

Set Size Estimation

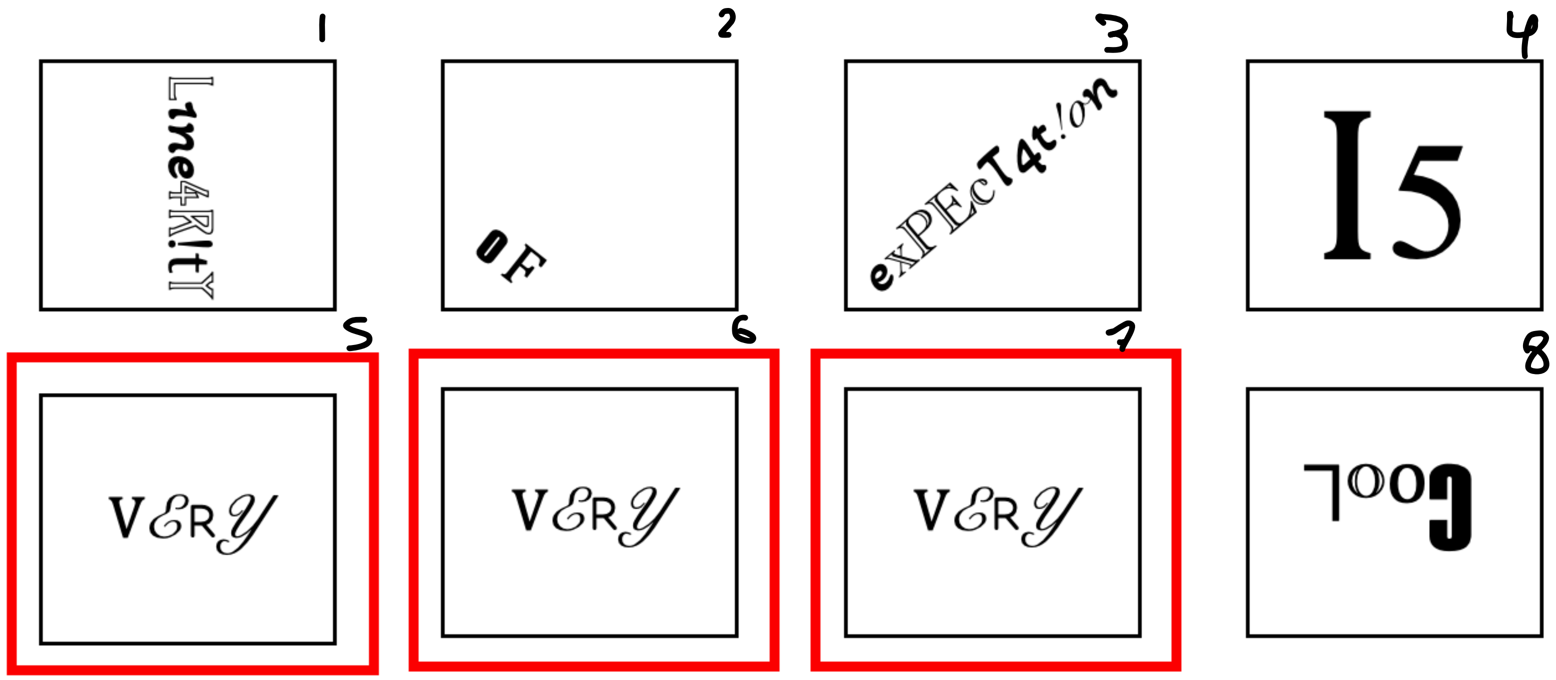
↳ internet traffic

↳ ecology

↳ social networks



naive: we keep making calls until we see 1 million CAPTCHAs



$$D_{i,j} = \begin{cases} 1 & \text{if ith and jth item are same} \\ 0 & \text{else} \end{cases}$$

$$D_{1,2} = 0$$

$$D_{5,7} = 1$$

\# samples

$$D = \sum_{i=1}^{m} \sum_{j=i+1}^{m} D_{i,j}$$

$$E[D_{i,j}] = 1 \cdot Pr(i^{th}, j^{th} \text{ same}) + 0 \cdot$$

$$= Pr(i^{th}, j^{th} \text{ same})$$

$$= \frac{1}{n}$$

n = # unique items

$$E[D] = \sum_{i=1}^{m} \sum_{j=i+1}^{m} E[D_{i,j}]$$

$$= \frac{1}{n} \sum_{i=1}^{m} \sum_{j=i+1}^{m} 1$$

$$= \frac{m(m-1)}{2 \cdot n}$$

$$\mathbb{E}[D] = \frac{m(m-1)}{2 \cdot n}$$

Suppose we see $D=10$

when $m = 1000$,

$n = 1{,}000{,}000$

$$\mathbb{E}[D] = \frac{1000 \cdot 999}{2 \cdot 1{,}000{,}000} = .4995$$

$$\boxed{\begin{array}{l} \underline{\text{Markov's Inequality}} \\ X \text{ non-negative r.v.} \\ t > 0 \\ \Pr(X \geq t) \leq \dfrac{\mathbb{E}[X]}{t} \end{array}}$$

$$\Pr(D \geq 10) \leq \frac{\mathbb{E}[D]}{10} = \frac{.4995}{10}$$

$$= .04995$$

$$\mathbb{E}[X] = \sum_{x} x \Pr(X=x)$$

$$= \sum_{\substack{x \\ x \geq t}} x \Pr(X=x) + \sum_{\substack{x \\ x < t}} x \Pr(X=x)$$

$$\geq \sum_{\substack{x \\ x \geq t}} t \Pr(X=x) + 0$$

$$\geq t \sum_{\substack{x \\ x \geq t}} \Pr(X=x) = t \Pr(X \geq t)$$

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$