

# Recording

The class will be recorded and the recordings made available via Canvas

To opt out: send a message in the Chat

# DS 4400

## Machine Learning and Data Mining I

Alina Oprea  
Associate Professor  
Khoury College of Computer Science  
Northeastern University

September 17 2020

# Announcements

- HW 1
  - ~~Will be~~ <sup>IS</sup> out today <sup>GRADESCOPE</sup>
  - Will be due on Monday, Sept. 28 <sup>MIDNIGHT</sup>
- Python tutorials
  - Numpy tutorial by Matthew Jagielski
    - Friday, Sept. 18, 1-2pm
  - Panda data frames tutorial by Alex Wang
    - Wed, Sept. 23, 5-6pm
  - Same Zoom links as office hours

# Recap

- ML is a subset of AI designing learning algorithms
- Learning tasks are *supervised* (e.g., classification and regression) or *unsupervised* (e.g., clustering)
  - Supervised learning uses labeled training data
- Learning the “best” model is challenging
  - Design algorithm to minimize the error
  - Bias-Variance tradeoff Model COMPLEXITY  $\uparrow \Rightarrow$  VAR  $\uparrow$  BIAS  $\downarrow$
  - Need to generalize on new, unseen test data
  - Occam’s razor (prefer simplest model with good performance) COMPLEXITY  $\downarrow \Rightarrow$  VAR  $\downarrow$  BIAS  $\uparrow$

# Outline

- Probability review
  - Conditional probabilities
  - Bayes Theorem
- Linear algebra review
  - Matrix and vector operations
  - Transpose, inverse
  - Rank of a matrix
- Covariance and correlation coefficient

# Probability review

# Probability Resources

- [Review notes](#) from Stanford's machine learning class
- Sam Roweis's [probability review](#)
- David Blei's [probability review](#)
- Books:
  - Sheldon Ross, A First course in probability

# Discrete Random Variables

- Let  $A$  denote a random variable FINITE
  - $A$  represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:
  - $A$  = I have a headache
  - $A$  = Sally will be the US president in 2020
- $P(A)$  is “the fraction of possible worlds in which  $A$  is true”



# Visualizing A

- Universe  $U$  is the event space of all possible worlds
  - Its area is 1
  - $P(U) = 1$

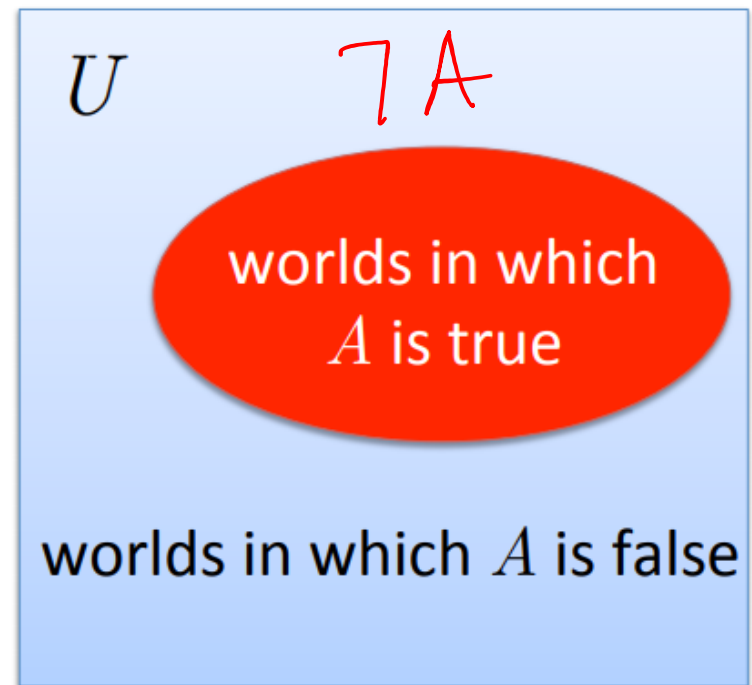
- $P(A) = \text{area of red oval}$

- Therefore:

$$P(A) + P(\neg A) = 1$$

$$P(\neg A) = 1 - P(A)$$

$$P[A] = \frac{|A|}{|U|}$$



$$|A| + |\neg A| = |U|$$

# Working with Probabilities

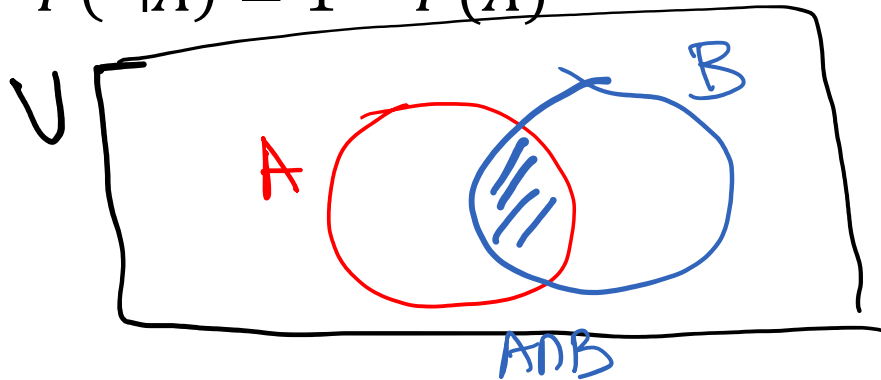
- $0 \leq P(A) \leq 1$
- $P(U) = 1; P(\Phi) = 0$
- $P(\neg A) = 1 - P(A)$

EMPTY SET

UNION OF TWO EVENTS

$$|A \cup B| = |A| + |B| - |A \cap B|$$

SET OPERATIONS



$$P[A \cup B] = P(A) + P(B) - \underbrace{P(A \cap B)}_{\geq 0}$$

$$\text{UNION BOUND: } \Rightarrow P[A \cup B] \leq P[A] + P[B]$$

# Examples discrete RV

- Bernoulli RV
  - X is modelling a coin toss
  - Output: 1 (head) or 0 (tail)
  - $P[X=1] = p$ ;  $P[X=0] = 1-p$
- Y is the number of points in a fair dice
  - $P[Y = k] = ?$  for  $k \in \{1, \dots, 6\}$ ?
  - $P[Y = \text{even}] = ?$

$$0 \leq p \leq 1$$

$$P[Y=1] = \frac{1}{6}, \dots, P[Y=6] = \frac{1}{6}$$
$$P[Y=\text{even}] = \frac{1}{2}; \quad P[Y=\text{odd}] = \frac{1}{2}$$

# Example discrete RV

- Z is the sum of two fair dice
  - What is  $P[Z = k]$  for  $k \in \{2, \dots, 12\}$ ?
  - What is  $k$  for which this probability is maximum?

$$P[Z=2] = \frac{1}{36} ; P[Z=3] = \frac{1}{18} = \frac{2}{36} ; P[Z=4] = \frac{3}{36}$$
$$P[Z=12] = \frac{1}{36}$$

...

$k=7$  is max

$$P[Z=7] = \frac{6}{36} = \frac{1}{6}$$

# Expectation and variance

Expectation for discrete random variable X

$$E[X] = \sum_v v \Pr[X = v]$$

↓ POSSIBLE VALUES

$$X \sim \begin{pmatrix} 1 & 2 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$E[X] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}$$

Properties

- •  $E[aX] = a E[X]$
- •  $E[X + Y] = E[X] + E[Y]$
- •  $E[f(X)] = \sum_v f(v) \Pr[X = v]$

X is RV; a is CONSTANT

$$E[X^2] = \sum_n n^2 \cdot \Pr[X=n]$$

Variance

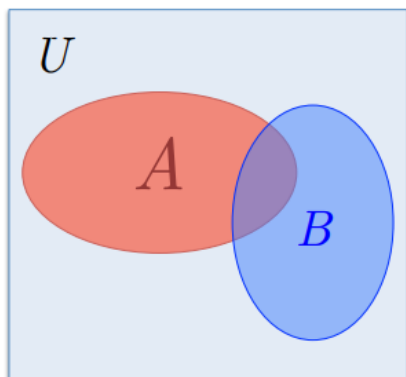
Def

$$\text{Var}[X] \triangleq E[(X - E(X))^2]$$

$$\begin{aligned} \text{Var}[X] &= E[(X - E(X))^2] = E[X^2 - 2XE(X) + E^2(X)] = \\ &= E[X^2] - E[2XE(X)] + E^2(X) = E[X^2] - 2E^2[X] + E^2[X] \\ &= E[X^2] - E^2[X] \end{aligned}$$

# Conditional Probability

- $P(A \mid B)$  = Fraction of worlds in which  $B$  is true that also have  $A$  true



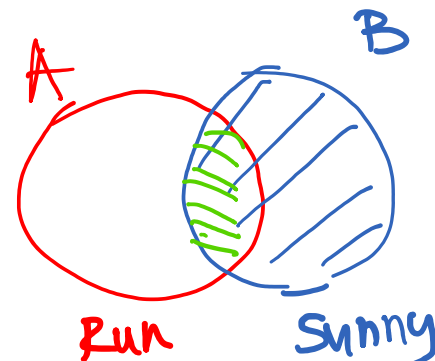
What if we already know that  $B$  is true?

That knowledge changes the probability of  $A$

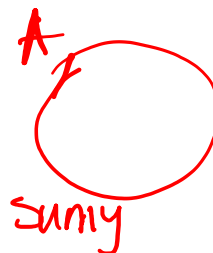
- Because we know we're in a world where  $B$  is true

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A \mid B) \times P(B)$$



$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad \text{DEF}$$



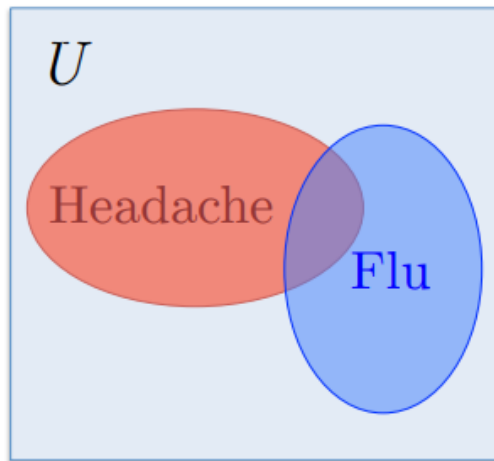
Events  $A$  and  $B$  are **independent** if  $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$

$$\underline{P[A|B]} = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = \underline{P[A]}$$

# Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

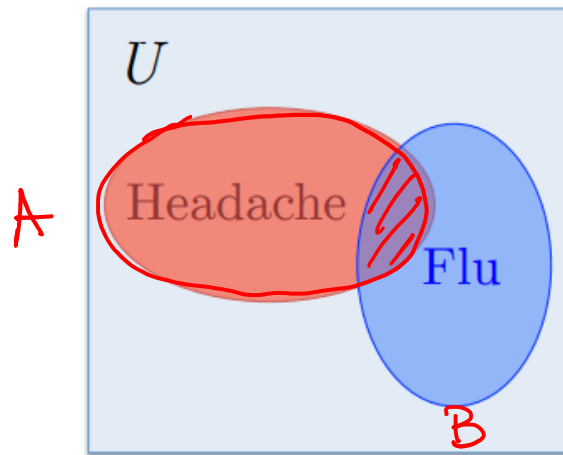
$$P(\text{headache} \mid \text{flu}) = 1/2$$

} GIVEN

“Headaches are rare and flu is rarer, but if you’re coming down with the flu there’s a 50-50 chance you’ll have a headache.”

# Inference from Conditional Probability

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$
$$P(A \wedge B) = P(A | B) \times P(B)$$



$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} | \text{flu}) = 1/2$$

KNOW

$$\rightarrow P(A|B) = \frac{1}{2}$$

One day you wake up with a headache.  
You think: “Drat! 50% of flus are  
associated with headaches so I must have  
a 50-50 chance of coming down with flu.”

Is this reasoning good?

$$P(F|H)$$



# Inference from Conditional Probability

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

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

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A | B) \times P(B)$$

GIVEN {

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} | \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} | \text{headache}) = ?$$

$$\text{If } P(A) = P(B) \Rightarrow \underline{P(A|B)} = \underline{P(B|A)}$$

# Exercises

$$E[X] = p; \text{Var}[X] = E[X^2] - E^2[X] = p - p^2 = p(1-p)$$

1. Compute Expectation and Variance for a Bernoulli RV

$$E[X^2] = \sum n^2 P(X=n)$$

$$- P[X = 1] = p; P[X = 0] = 1 - p$$

2. Conditional probabilities

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A | B) \times P(B)$$

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} | \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} | \text{headache}) = ? \quad \frac{P(H \wedge F)}{P(H)} = \frac{\frac{1}{80}}{\frac{1}{10}} = \frac{1}{8}$$



**BREAKOUT  
ROOMS**

$$P(H|F) \cdot P(F) = \frac{1}{2} \cdot \frac{1}{40} = \frac{1}{80}$$

# Bayes' Rule

$$P(A | B) = \frac{P(B | A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

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

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

$$P(A \cap B) = \underbrace{P(A|B)P(B) = P(B|A)P(A)}_{\text{BAYES}}$$



**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418

# Multi-Value Random Variable

- Suppose  $A$  can take on more than 2 values
- $A$  is a *random variable with arity  $k$*  if it can take on exactly one value out of  $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

$$1 = \sum_{i=1}^k P(A = v_i)$$

$$\begin{aligned} P(A = \text{Jan}) \vee A = \text{Feb}) \\ = P(A = \text{Jan}) + P(A = \text{Feb}) \end{aligned}$$

$A = \text{Month of Year}$

EXAMPLE  $P[A = \text{Jan}] = \frac{31}{365} \approx \frac{1}{12}$

# Marginalization

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_k])$$

$$P(B) = \sum_{i=1}^k P(B \wedge A = v_i) = \sum_{i=1}^k P(B | A = v_i) P(A = v_i)$$

*A binary*

- This is called **marginalization** over  $A$   $P(B) = P(B \cap A) + P(B \cap \neg A)$

EXAMPLE

$B = \text{Sunny}$   
 $A = \text{Month}$

$$P(\text{Sunny}) = \sum_{i=1}^{12} P(\text{Sunny} \cap A = \text{Month } i)$$
$$= P(\text{Sunny} | A = \text{Month } i) \cdot P(A = \text{Month } i)$$

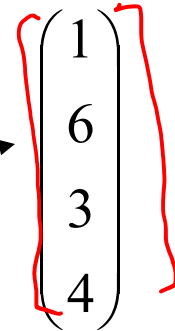
# Linear algebra review

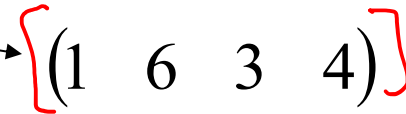
# Resources

- Zico Kolter, [Linear algebra review](#)
- Sam Roweis's [linear algebra review](#)
- Books:
  - O. Bretscher, Linear Algebra with Applications

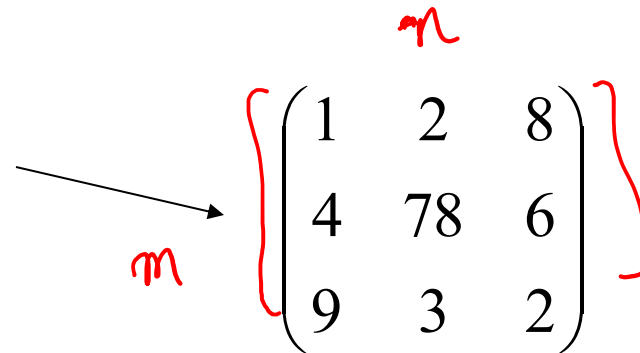
# Vectors and matrices

- **Vector** in  $\mathbb{R}^n$  is an ordered set of  $n$  real numbers.
  - e.g.  $v = (1, 6, 3, 4)$  is in  $\mathbb{R}^4$
  - A column vector:
  - A row vector:


$$\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$$


$$(1 \ 6 \ 3 \ 4)$$

- $m$ -by- $n$  **matrix** is an object in  $\mathbb{R}^{m \times n}$  with  $m$  rows and  $n$  columns, each entry filled with a (typically) real number:


$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$



# Vector operations

- Addition component by component

$$[a_1, a_2, \dots, a_n] + [b_1, b_2, \dots, b_n] = [a_1 + b_1, \dots, a_n + b_n]$$

$$[1, -2, 5] + [0, 3, 7] = [1, 1, 12]$$

- Subtraction is also done component by component

$$[a_1, a_2, \dots, a_n] - [b_1, b_2, \dots, b_n] = [a_1 - b_1, \dots, a_n - b_n]$$

– Can add and subtract row or column vectors of same dimension

- Dot product

– Only works for row and column vector of same size

$$[a_1, a_2, \dots, a_n] \cdot \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = [a_1 b_1 + \dots + a_n b_n]$$

$$[1, -2, 5] \cdot \begin{bmatrix} 0 \\ 3 \\ 7 \end{bmatrix} = 1 \cdot 0 + (-2) \cdot 3 + 5 \cdot 7 = -6 + 12 = 6$$

# Matrix multiplication

We will use upper case letters for matrices. The elements are referred by  $A_{i,j}$ .

- **Matrix product:**

$$A \in \mathbb{R}^{m \times n} \quad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11} + b_{11} & \cdot \\ \cdot & a_{22} + b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

DOT  
PRODUCT

# Matrix transpose

**Transpose:** You can think of it as

– “flipping” the rows and columns

OR

– “reflecting” vector/matrix on line

e.g.  $\begin{bmatrix} a \\ b \end{bmatrix}^T = \begin{bmatrix} a & b \end{bmatrix}$

$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

A is a **symmetric matrix** if  $A = A^T$

# Linear independence

- A set of vectors is **linearly independent** if none of them can be written as a linear combination of the others.

- Vectors  $x_1, \dots, x_k$  are linearly independent if  $c_1x_1 + \dots + c_kx_k = 0$  implies  $c_1 = \dots = c_k = 0$
- Otherwise they are linearly dependent

*vector*  
*LINEAR COMBINATION*

$c_1, c_2 \in \mathbb{R}$

$$c_1x_1 + c_2x_2 = 0$$

$$x_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \quad x_2 = \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$$

$$c_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} c_1 + 0c_2 = 0 \\ 2c_1 + 3c_2 = 0 \\ c_1 + 3c_2 = 0 \end{cases} \Rightarrow \begin{aligned} c_1 &= 0 \\ c_2 &= 0 \end{aligned}$$

# Linear independence

- A set of vectors is **linearly independent** if none of them can be written as a linear combination of the others.
- Vectors  $x_1, \dots, x_k$  are linearly independent if  $c_1x_1 + \dots + c_kx_k = 0$  implies  $c_1 = \dots = c_k = 0$
- Otherwise they are **linearly dependent**

LINEARLY DEPENDENT!

$$c_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + c_2 \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} + c_3 \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{cases} c_1 + 4c_2 + 2c_3 = 0 & (1) \\ 2c_1 + c_2 - 3c_3 = 0 & (2) \\ 3c_1 + 5c_2 - c_3 = 0 & (3) \end{cases}$$

$$7c_1 + 14c_2 = 0, \quad c_1 = -2c_2$$

$$\begin{cases} -6c_2 + 5c_2 - c_3 = 0 \\ c_2 = -c_3; \quad c_3 = -c_2 \end{cases}$$

$$x_3 = -2x_1 + x_2$$

$$\begin{bmatrix} -4 & 2 & -2 \end{bmatrix}$$

$$c_1 = -2, \quad c_2 = 1, \quad c_3 = -1$$

# Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is
  - The maximal number of linearly independent columns
  - The maximal number of linearly independent rows

- If A is n by m, then  $A = m \left[ \begin{array}{c} \overbrace{\hspace{2cm}}^n \end{array} \right]$   $m \leq n$ 
  - rank(A)  $\leq \min(m, n)$

- Examples

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

RANK = 2

$$\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$$

RANK = 1

LINEARLY  
INDEPENDENT

$$\begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

RANK = 2

# Inverse of a matrix

- Inverse of a square matrix  $A$ , denoted by  $A^{-1}$  is the *unique* matrix s.t.

- $AA^{-1} = A^{-1}A = I$  (identity matrix)

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

INVERSE:  $A^{-1}$

- Inverse of a square matrix exists only if the matrix is **full rank**
- If  $A^{-1}$  and  $B^{-1}$  exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$

# Diagonal matrices

$$D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix} \quad d_1 \neq 0, \dots, d_n \neq 0$$

$$D^{-1} = \begin{bmatrix} \frac{1}{d_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{d_n} \end{bmatrix}$$

$$DD^{-1} = I$$



# System of linear equations

$$\begin{array}{rclcrcl} 4x_1 & - & 5x_2 & = & -13 \\ -2x_1 & + & 3x_2 & = & 9. \end{array}$$

Matrix formulation

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

If A has an inverse, solution is  $x = A^{-1}b$