DS 4400

Machine Learning and Data Mining I
Spring 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

January 26 2021

Today's Outline

- Announcements
  - HW 1 is out, due on Friday, Feb 5
  - First numpy tutorial by Prabal M.
    - Thu, Jan 28, 5-6pm, Zoom link for office hours
- Probability review
  - Conditional probabilities
  - Bayes Theorem
- Linear algebra review
  - Matrix and vector operations
  - Transpose, inverse
  - Rank of a matrix

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
  - A represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:

  A = It will snow tomorrow

  B = The patient will recover

- $\mathrm{P}(A)$ 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
  - $\mathrm{P}(U) = 1$

A = "snow"

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

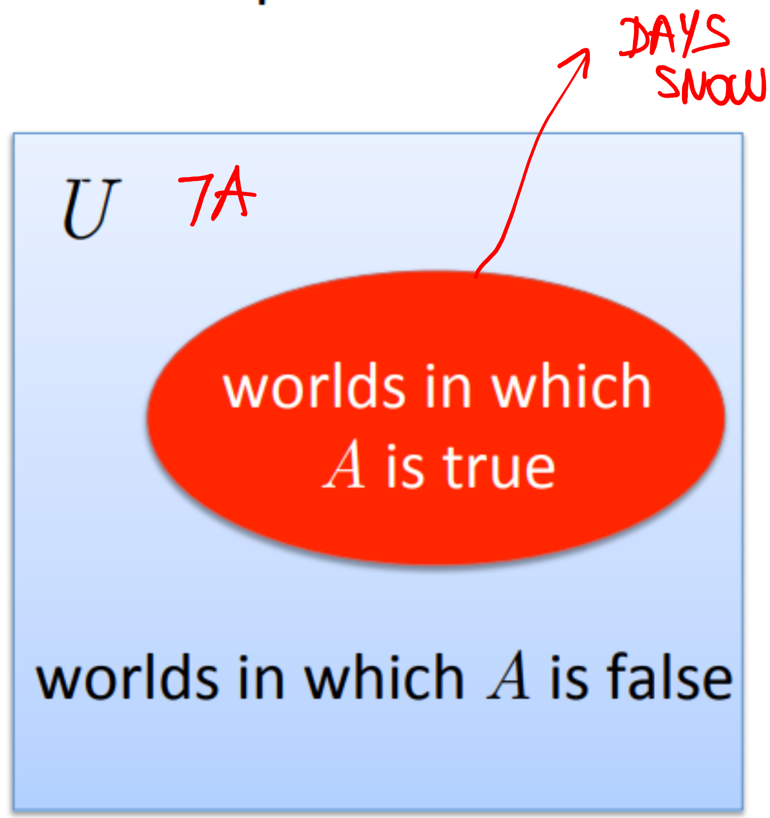
$P[A] = \frac{|A|}{|U|} = \frac{20}{365}$

- Therefore:

$$\begin{cases} P(A) + P(\neg A) = 1 \\ P(\neg A) = 1 - P(A) \end{cases}$$

$|A| + |\neg A| = |U| = 365$

$P(A) = \frac{|A|}{|U|}$; $P(\neg A) = \frac{|\neg A|}{|U|}$; $P(A) + P(\neg A) = 1$

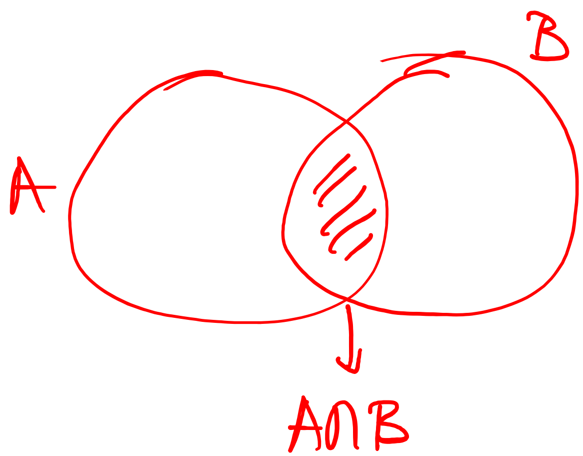


Working with Probabilities

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

A = SKI

B = SNOW



$P[A \cup B] = P(A) + P(B) - P(A \cap B)$

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

UNION BOUND:

$P[A \cup B] \leq P(A) + P(B)$

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 ; $0 < p < 1$
- Y is the number of points in a fair dice
  - $k \in \{1, \dots, 6\}$, P[Y = k]= $\frac{1}{6}$
  - P[Y = even] = $\frac{1}{2}$

  $P[Y=odd] = \frac{1}{2}$

Example discrete RV

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

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

$P[Z=12] = \frac{1}{36}$

. . . . .

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

Expectation

Expectation for discrete random variable X

$$\boxed{E[X] = \sum_{v} v Pr[X = v]}$$

OVER ALL POSSIBLE VALUES

Bernoulli: P[X=1] =p; P[X=0] = 1-p

$E[X] = 1 \cdot P[X=1] + 0 \cdot P[X=0] = p$

Expectation and variance

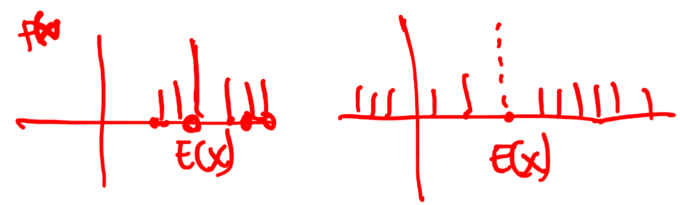
Expectation for discrete random variable X

$$\boxed{E[X] = \sum_v v Pr[X = v]}$$

$\rightarrow$ $E[X^2] = \sum_v v^2 P[X = v]$

Properties

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



$\rightarrow$ Variance: $\quad \mathrm{Var}[X] = E\left[\left(X - E(X)\right)^2\right]$

$$\mathrm{Var}(x) = E\left[x^2 - 2XE(x) + E^2(x)\right] = E[x^2] - \underbrace{E[2XE(x)]}_{=\,2E(x)\cdot E(x) = 2E^2(x)} + \underbrace{E[E^2(x)]}_{E^2(x)}$$

$\rightarrow$ $\mathrm{Var}(x) = E[x^2] - E^2(x)$

Variance of Bernoulli

- Variance: $\mathrm{Var}[X] = E(X^2) - E^2(X)$

Bernoulli: P[X=1] =p; P[X=0] = 1-p ; $0 < p < 1$

$E(X) = p$

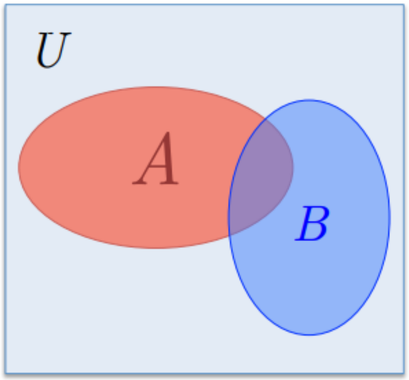
$E(X^2) = p$

$\mathrm{Var}(X) = p - p^2 = p(1-p)$

$p = \frac{1}{2}$, UNIFORM , MAX VARIANCE

Conditional Probability

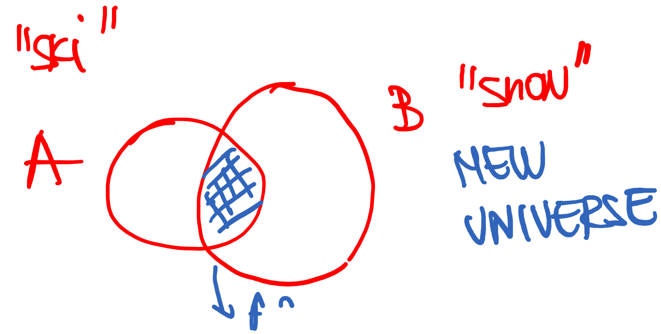
- $P(A \mid B)$ = Fraction of worlds in which B is true that also have A true



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



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

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

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

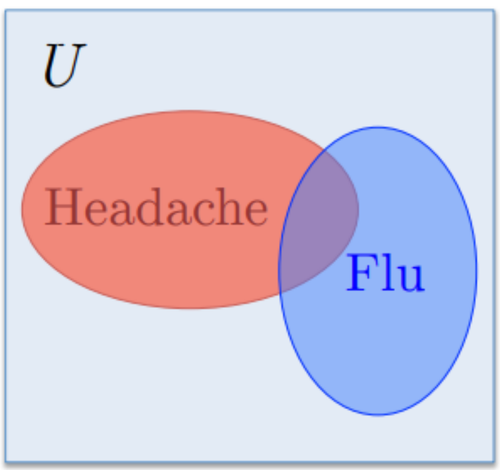
Events A and B are **independent** if Pr[A ∩ B] = Pr[A] · Pr[B]

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

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(headache) = 1/10
P(flu) = 1/40
P(headache | 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

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

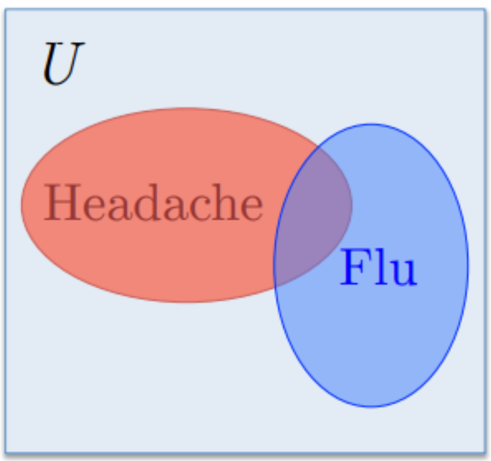
$$P(A \wedge B) = P(A \mid B) \times P(B)$$

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

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

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

only if $P(A) = P(B)$



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 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

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

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

P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

Want to solve for:
$\rightarrow$ P(headache ∧ flu) = ?
$\rightarrow$ P(flu | headache) = ?

Exercises

$E(x^2) = \frac{1}{6}\cdot 1^2 + \frac{1}{6} 2^2 + \dots + \frac{1}{6}\cdot 6^2 = \frac{91}{6}$

- Compute Expectation and Variance for dice rolling random variable X
  - P[X=k] =1/6 for $k \in \{1, \dots, 6\}$

$E(x) = \frac{1}{6}(1+2+\dots+6) = 3.5$

$Var(x) = E(x^2) - E^2(x) = \frac{35}{12}$

- Conditional probabilities

BREAKOUT ROOMS

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

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

$P(H \mid F) \cdot P(F) = \frac{1}{80}$

P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

Want to solve for:
P(headache ∧ flu) = ?
P(flu | headache) = ?

$\frac{P(F \cap H)}{P(H)} = \frac{\frac{1}{80}}{\frac{1}{10}} = \frac{1}{8}$

Inference from Conditional Probability

Indep

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

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

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

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

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

P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2 $\neq P(H)$

Want to solve for:
P(headache ∧ flu) = ?
P(flu | headache) $\neq$? $P(\text{Flu})$

P(headache ∧ flu) = P(headache | flu) x P(flu) $= \frac{1}{80}$
= 1/2 x 1/40 = 0.0125

P(flu | headache) = P(headache ∧ flu) / P(headache)
= 0.0125 / 0.1 = 0.125

Bayes Theorem

Bayes' Rule

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

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

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

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

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) P(A)$$

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 $\{v_1, v_2, \ldots, v_k\}$
- Thus…

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

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

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

A = Month of Year

$P(A = \text{Jan}) = \frac{31}{365}$; $P(A = \text{Feb}) = \frac{28}{365}$

Marginalization

- We can also show that:

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

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

- This is called **marginalization** over A

A = binary

$P(B) = P(B \cap A) + P(B \cap \neg A)$

$= P(B|A)\,P(A) +$

$P(B|\neg A) \cdot P(\neg A)$

EXAMPLE

A = Month of Year

B = Sunny

$$P(\text{Sunny}) = \sum_{i=1}^{12} \underbrace{P(\text{Sunny} \wedge A = \text{Month } i)}_{= P(\text{Sunny} \mid A = \text{Month } i)\, P(A = \text{Month } i)}$$

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 R^n is an ordered set of n real numbers.
  – e.g. v = (1,6,3,4) is in R^4
  – A column vector:
  – A row vector:

COLUMN

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

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

ROW

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

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

Vector operations

- Addition component by component

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

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

- 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 (INNER PRODUCT)
  - Only works for row and column vector of same size

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

REAL NUMBER

$$[1, -2, 5] \cdot \begin{bmatrix} 0 \\ 3 \\ 7 \end{bmatrix} = 0 + 6 + 35 = 29$$

Matrix Operations

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

$-$

Addition

Subtraction

Matrix multiplication

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

- Matrix product:

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

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

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

e.g.

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

ROW 1 COLUMN 1

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

ROW 1 COLUMN 2

Matrix transpose

Transpose: You can think of it as
- "flipping" the rows and columns

OR
- "reflecting" vector/matrix on line

A: (m x n)   m ≠ p
B: (n x p)
A^T: (n x m)
B^T: (p x n)

PROPERTIES:
1) $(A^T)^T = A$
2) $(A+B)^T = A^T + B^T$
3) $(AB)^T = B^T A^T$

SIZE: (p x m)

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

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

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

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
  - rank(A)<= min(m,n)

- Examples

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

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)

A: size n x n, A^{-1}: size n x n

$$I = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

n x n

- Inverse of a square matrix exists only if the matrix is full rank

In General

$AB \neq BA$

- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$

Diagonal matrices

$$D = \begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & & d_n \end{pmatrix}$$

$$D \cdot D^{-1} = D^{-1} \cdot D = I$$

Identity

If $d_1 \neq 0, \ldots, d_n \neq 0$

$$D^{-1} = \begin{pmatrix} \frac{1}{d_1} & 0 & \cdots & & 0 \\ 0 & \frac{1}{d_2} & \cdots & & 0 \\ \vdots & & & & \\ 0 & & \cdots & & \frac{1}{d_n} \end{pmatrix}$$

System of linear equations

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

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 $\boxed{x = A^{-1}b}$