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 out today
  - Will be due on Monday, Sept. 28
- 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
  - Need to generalize on new, unseen test data
  - Occam's razor (prefer simplest model with good performance)

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

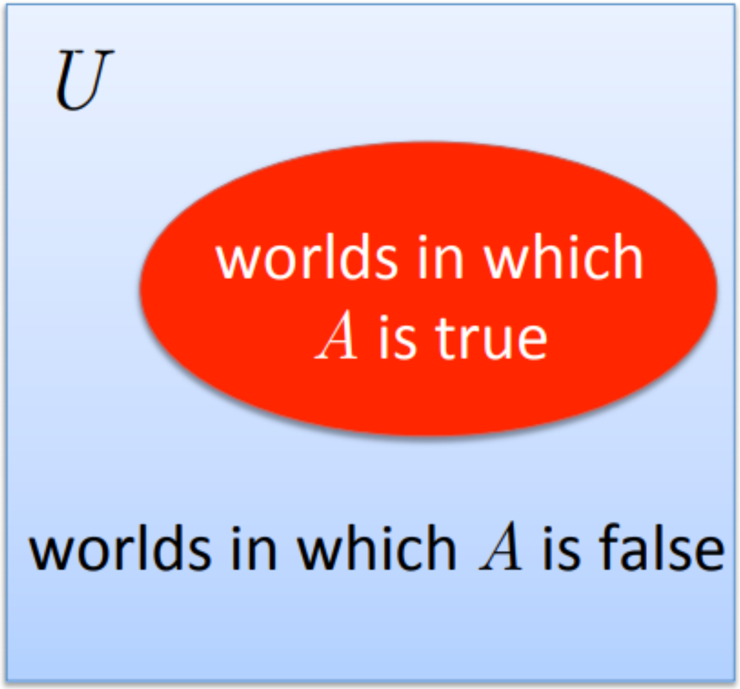
- Review notes from Stanford's machine learning class
- Sam Roweis's probability review
- David Blei's probability review
- Books:
  - Sheldon Ross, A First course in probability

Discrete Random Variables

- Let A denote a random variable
  - 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
- $\mathrm{P}(A)$ 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
  - $\mathrm{P}(U) = 1$



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

- Therefore:

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

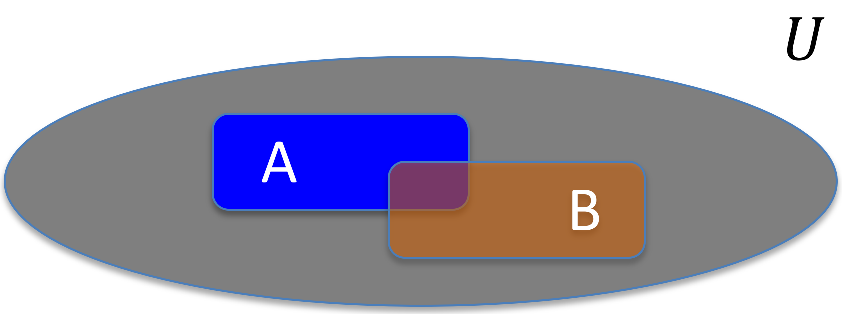
$$P(\neg A) = 1 - P(A)$$

Working with Probabilities

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

Working with Probabilities

- $0 \leq P(A) \leq 1$
- $P(U) = 1; P(\Phi) = 0$
- $P(\neg A) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



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
- Y is the number of points in a fair dice
  - P[Y = k]= ? for $k \in \{1, \ldots, 6\}$?
  - P[Y = even] = ?

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?

Expectation and variance

Expectation for discrete random variable X

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

Properties

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

Variance $$Var[X] \triangleq E[(X - E(X))^2]$$

Expectation and variance

Expectation for discrete random variable X

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

Properties

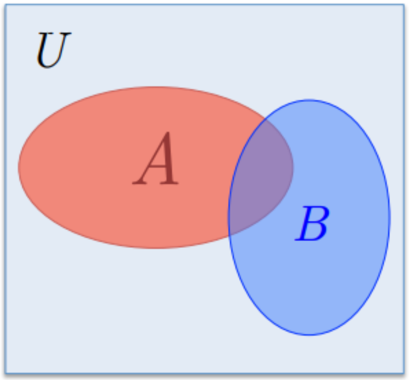
- $E[aX] = a\, E[X]$
- $E[X + Y] = E[X] + E[Y]$
- $E[f(X)] = \sum_v f(v) Pr[X = v]$

Variance $\quad Var[X] \triangleq E[(X - E(X))^2]$

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

Conditional Probability

- $\mathrm{P}(A \mid B)$ = 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

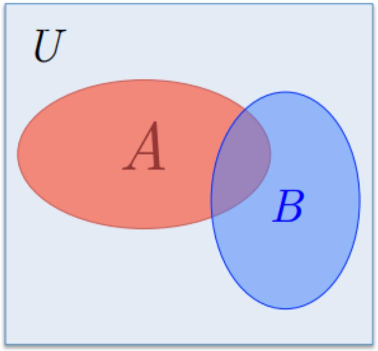
$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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

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

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

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

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

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

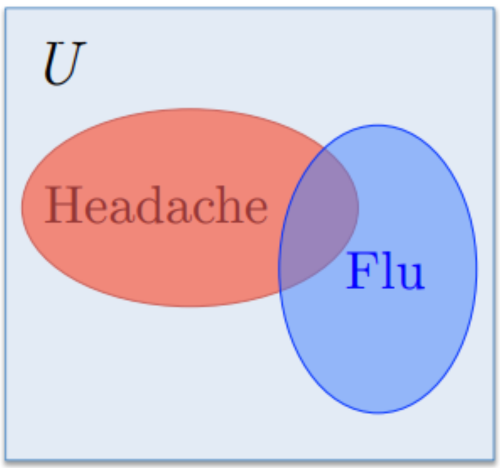
If A and B are independent

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

Inference from Conditional Probability

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

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



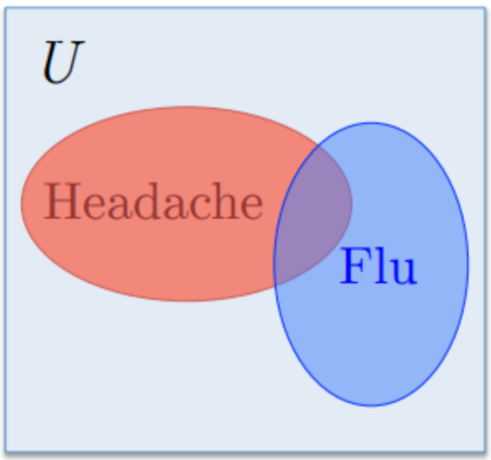
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

"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(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?

Inference from Conditional Probability

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

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

$P(\text{headache}) = 1/10$
$P(\text{flu}) = 1/40$
$P(\text{headache} \mid \text{flu}) = 1/2$

Want to solve for:

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

Exercises

1. Compute Expectation and Variance for a Bernoulli RV
   - $P[X = 1] = p; P[X = 0] = 1 - p$
2. 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(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

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

Inference from Conditional Probability

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

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

$P(\text{headache}) = 1/10$
$P(\text{flu}) = 1/40$
$P(\text{headache} \mid \text{flu}) = 1/2$

Want to solve for:
$P(\text{headache} \wedge \text{flu}) = ?$
$P(\text{flu} \mid \text{headache}) = ?$

$P(\text{headache} \wedge \text{flu}) = P(\text{headache} \mid \text{flu}) \times P(\text{flu})$
$= 1/2 \times 1/40 = 0.0125$

$P(\text{flu} \mid \text{headache}) = P(\text{headache} \wedge \text{flu}) / P(\text{headache})$
$= 0.0125 / 0.1 = 0.125$

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

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**

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

(Super Easy) Derivation:

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

these are the same

Just set equal...

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

and solve...

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…

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

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

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

EXAMPLE

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…

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

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

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

EXAMPLE

A: Month of the Year

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

Marginalization

- We can also show that:

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

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

- This is called **marginalization** over A

EXAMPLE

Marginalization

- We can also show that:

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

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

- This is called **marginalization** over A

EXAMPLE A: Month of the Year; B: Tomorrow is sunny

$$P(Sunny) = \sum_{i=1}^{12} P(Sunny \mid A = Month\ i) P(A = Month\ i)$$

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:

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

  - A row vector:

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

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

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

Vector operations

- Addition component by component

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

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

- Subtraction is also done component by component

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

  - Can add and subtract row or column vectors of same dimension
- Dot product
  - Only works for row and column vector of same size

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

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

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.

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

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

Matrix transpose

Transpose: You can think of it as

- "flipping" the rows and columns

OR

- "reflecting" vector/matrix on line

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

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

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

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

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

Linear independence

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

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

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

(c_1, c_2)=(0,0), i.e. the columns are linearly independent.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linearly dependent

$$x_3 = -2x_1 + x_2$$

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)

- 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

System of linear equations

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

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 $x = A^{-1}b$