DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 11 2018

Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Supervised learning - 5 weeks
  - Linear regression
  - Classification (logistic regression, LDA, kNN, decision trees, random forest, SVM, Naïve Bayes)
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 1.5 weeks
  - Back-propagation, gradient descent
  - NN architectures
- Unsupervised learning – 2.5 weeks
  - Dimensionality reduction (PCA)
  - Clustering (k-means, hierarchical)
- Adversarial ML – 1 week
  - Security of ML at testing and training time

Grading

- Assignments – 20%
  - 4-5 assignments based on studied material in class, including programming exercises
  - Language: R or Python; Jupyter notebooks
- Final project – 25%
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Presentation at end of class (10 min) and report
- Exams – 50%
  - Midterm – 25%
  - Final exam – 25%
- Class participation – 5%
  - Participate in class discussion and on Piazza

Supervised Learning

Training

Data → Pre-processing → Feature extraction → Learning model

Data: Labeled (Typically)

Pre-processing: Normalization

Feature extraction: Selection

Learning model: Classification, Regression

Testing

New data → Learning model → Predictions

New data: Unlabeled

Predictions:

- Healthy / Sick — Classification
- Price / Risk score — Regression

Supervised Learning: Overview

Hypothesis space

Functions $\mathcal{F}$

$f : \mathcal{X} \rightarrow \mathcal{Y}$

Training data

$\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

Training

LEARNING

find $\hat{f} \in \mathcal{F}$
s.t. $y_i \approx \hat{f}(x_i)$

Learning machine

New data

PREDICTION

$y = \hat{f}(x)$

x

Testing

$\hat{f}$ model

Review

- 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
  - Select hypothesis space and loss function
  - Design algorithm to min loss function (error on training)
  - Bias-Variance tradeoff
  - Need to generalize on new, unseen test data
  - Occam's razor (prefer simplest model with good performance)

Outline

- Probability review
  - Random variables
  - Expectation, Variance, CDF, PDF
  - Example distributions
  - Independence and conditional independence
  - Bayes' Theorem
- Linear algebra review
  - Matrix, vectors
  - Inner products
  - Norms
  - Distance

Probability review

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

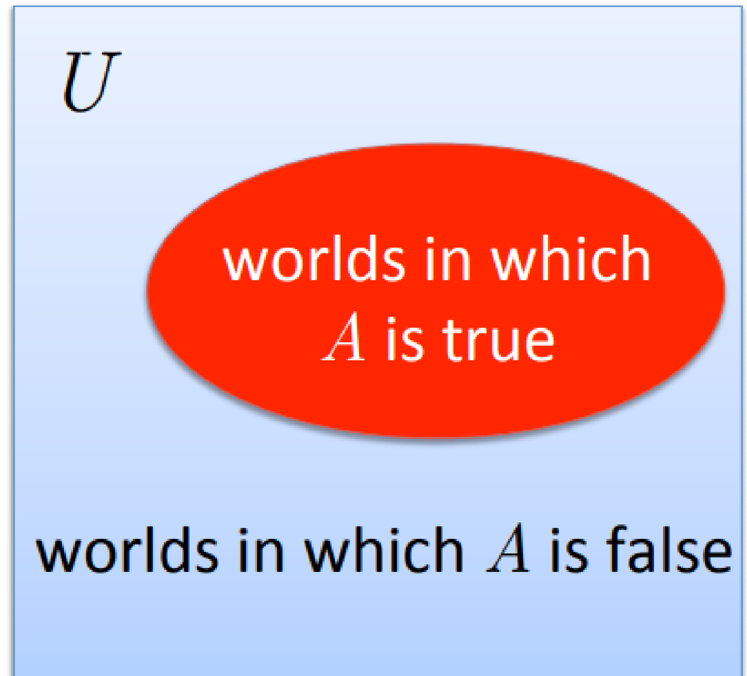
- Universe U is the event space of all possible worlds
  - Its area is 1
  - $\mathrm{P}(U) = 1$

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

- Therefore:

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

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



Axioms of Probability

Kolmogorov showed that three simple axioms lead to the rules of probability theory

- de Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

1. All probabilities are between 0 and 1:

$$0 \leq \mathrm{P}(A) \leq 1$$

2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:

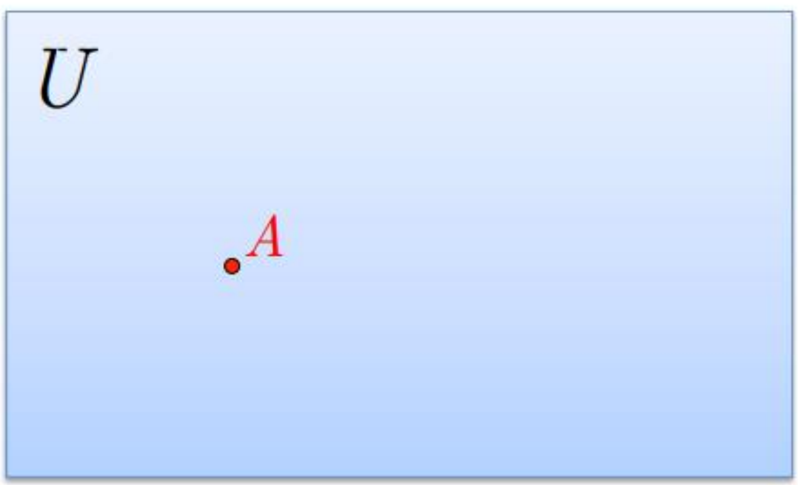
$$\mathrm{P}(\mathrm{true}) = 1 \; ; \quad \mathrm{P}(\mathrm{false}) = 0$$

3. The probability of a disjunction is given by:

$$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$$

Interpreting the Axioms

- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\text{true}) = 1$
- $\mathrm{P}(\text{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

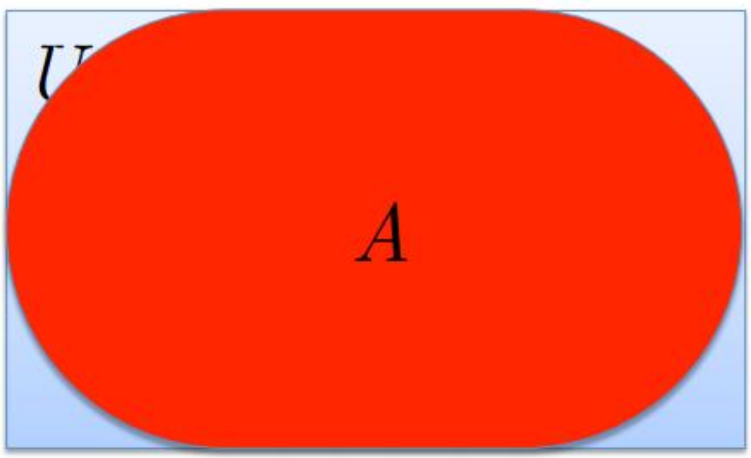


The area of A can't get any smaller than 0

A zero area would mean no world could ever have A true

Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

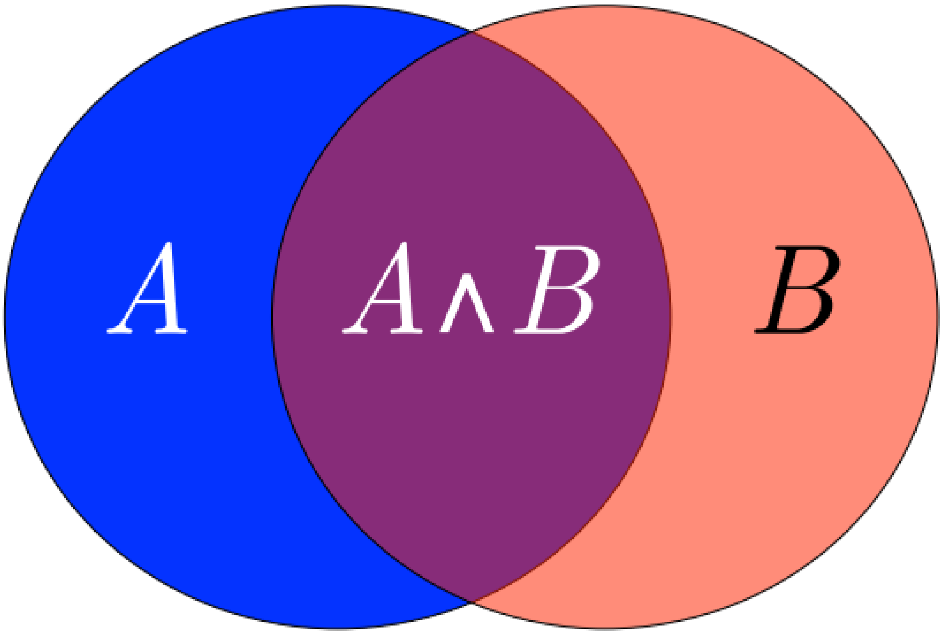


The area of A can't get any bigger than 1

An area of 1 would mean A is true in all possible worlds

Interpreting the Axioms

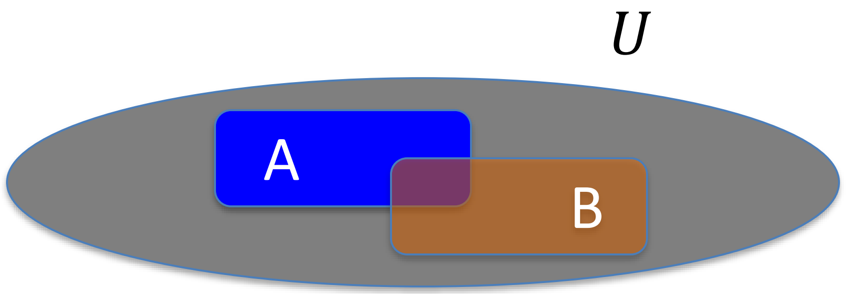
- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\text{true}) = 1$
- $\mathrm{P}(\text{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$



The union bound

- For events A and B

$$P[\, A \cup B \,] \le P[A] + P[B]$$



Axiom: $P[\, A \cup B \,] = P[A] + P[B] - P[A \cap B]$

If $A \cap B = \Phi$, then $P[\, A \cup B \,] = P[A] + P[B]$

Example:

$A_1 = \{$ all x in $\{0,1\}^n$ s.t $lsb_2(x)=11$ $\}$; $A_2 = \{$ all x in $\{0,1\}^n$ s.t. $msb_2(x)=11$ $\}$

$P[\, lsb_2(x)=11$ or $msb_2(x)=11 \,] = P[A_1 \cup A_2] \le ¼+¼ = ½$

Negation Theorem

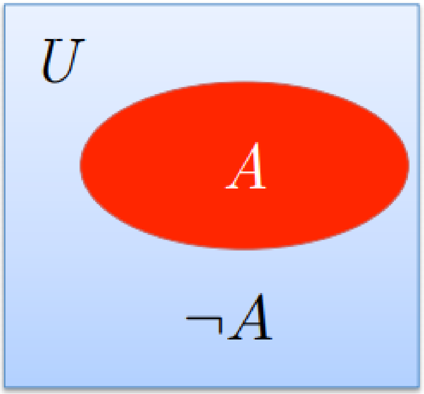
$0 \leq \mathrm{P}(A) \leq 1$

$\mathrm{P}(\mathrm{true}) = 1; \quad \mathrm{P}(\mathrm{false}) = 0$

$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

From these we can prove:

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



Marginalization

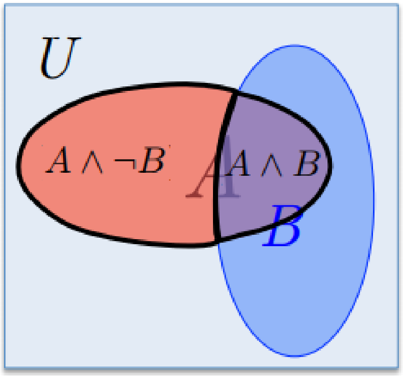
.

$0 \leq \mathrm{P}(A) \leq 1$
$\mathrm{P}(\mathrm{True}) = 1; \quad \mathrm{P}(\mathrm{False}) = 0$
$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

From these we can prove:

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

How?



Random Variables (Discrete)

Def: a random variable X is a function $X: U \longrightarrow V$

Def: A discrete random variable takes a finite number of values: $|V|$ is finite

Example: X is modeling a coin toss with output 1 (heads) or 0 (tail)

$\Pr[X=1] = p$, $\Pr[X=0] = 1-p$

Bernoulli Random Variable

We write $X \longleftarrow U$ to denote a **<u>uniform random variable</u> (discrete)** over U

for all $u \in U$: $\Pr[\, X = u \,] = 1/|U|$

Example: If p=1/2; then X is a uniform coin toss

Probability Mass Function (PMF): $p(u) = \Pr[X = u]$

Example

1. X is the number of heads in a sequence of n coin tosses

What is the probability $\mathrm{P}[X = k]$?

$\mathrm{P}[X = k] = \binom{n}{k}\ p^k(1-p)^{n-k}$ Binomial Random Variable

2. X is the sum of two fair dice

What is the probability $\mathrm{P}[X = k]$ for $k \in \{2, \dots, 12\}$?

P[X=2]=1/36; P[X=3]=2/36; P[X=4]= 3/36

For what k is $\mathrm{P}[X = k]$ highest?

Example discrete RVs

- $X \sim Bernoulli(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim Binomial(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim Geometric(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

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…

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

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

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

Multi-Value Random Variable

- We can also show that:

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

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

- This is called **marginalization** over A

Continuous Random Variables

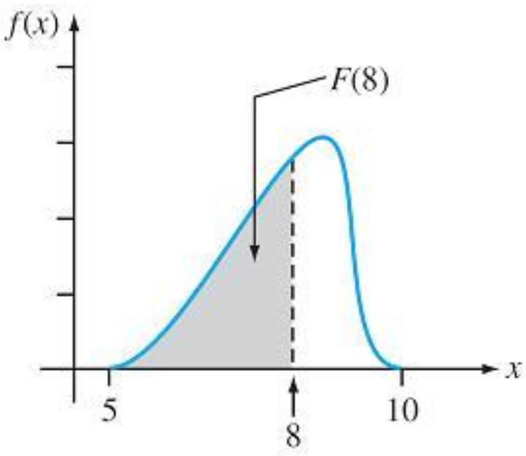
- X:U⟶V is continuous RV if it takes infinite number of values
- The **cumulative distribution function CDF** *F: R* ⟶ {0,1} for *X* is defined for every value *x* by:

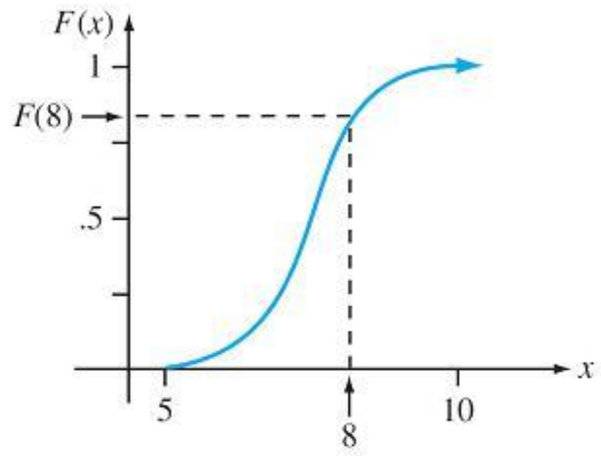
$$F(x) = \Pr(X \leq x)$$

- The **probability distribution function PDF** *f*(*x*) for *X is*

$$f(x) = dF(x)/dx$$

Increasing





Example continuous RV

- $X \sim Uniform(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

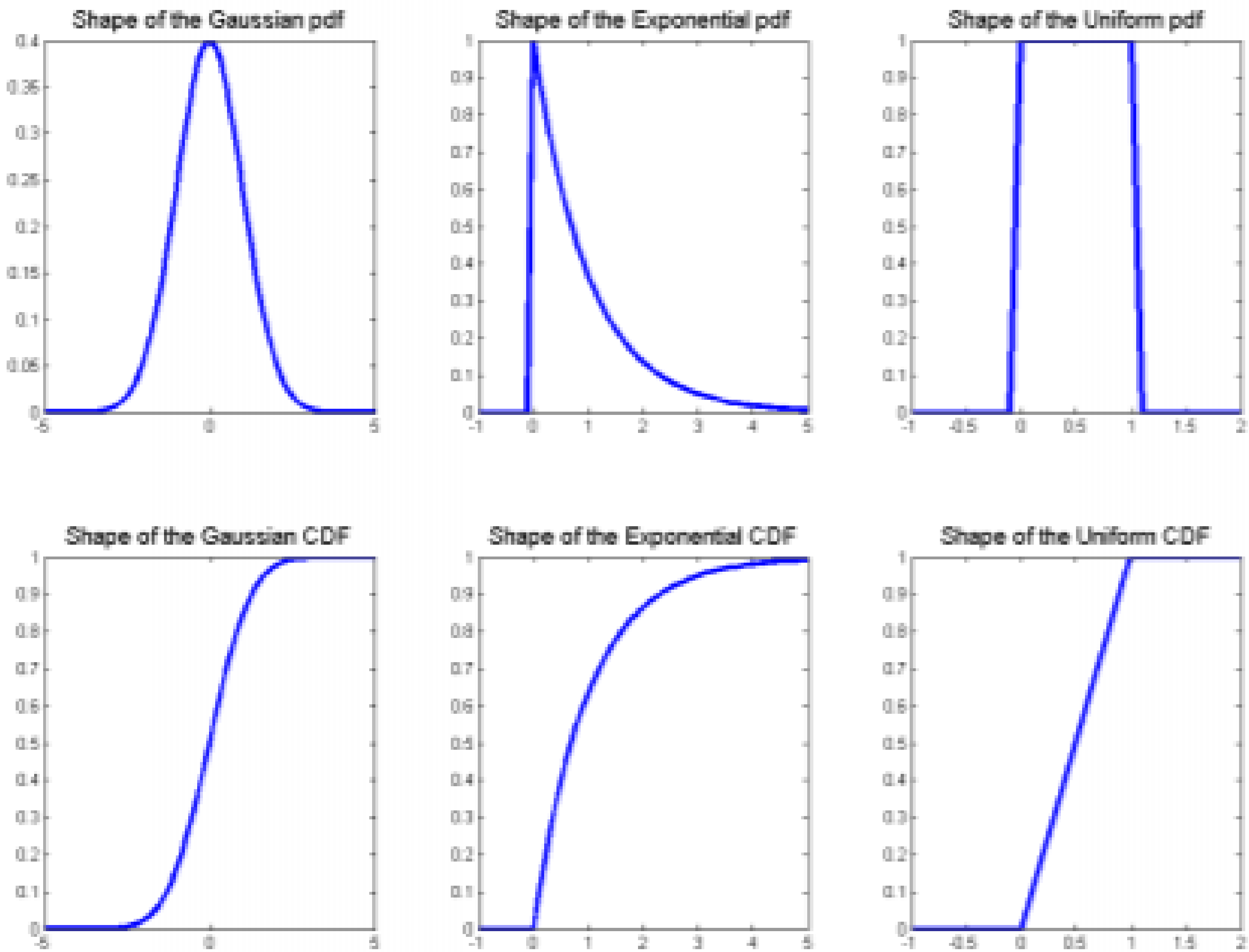
- $X \sim Exponential(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Normal(\mu, \sigma^2)$: also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Example CDFs and PDFs



Expectation and variance

Expectation for discrete random variable X

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x).$$

Properties

- $E[ag(X)] = a\,E[g(X)]$
- Linearity: $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$

Variance

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

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

Continuous RV

Expectation for continuous random variable X

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

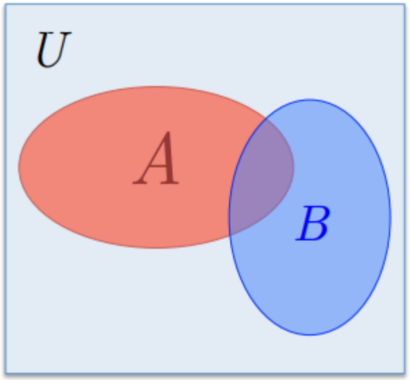
Variance is similar!

Example: Let X be uniform RV on [a,b]

- What is the CDF and PDF?
- Compute the expectation and variance of X

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

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

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

Def: Events A and B are **independent** if and only if

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

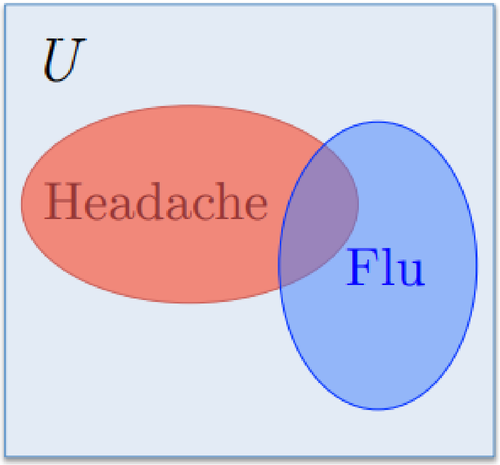
If A and B are independent

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

Inference from Conditional Probability

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

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



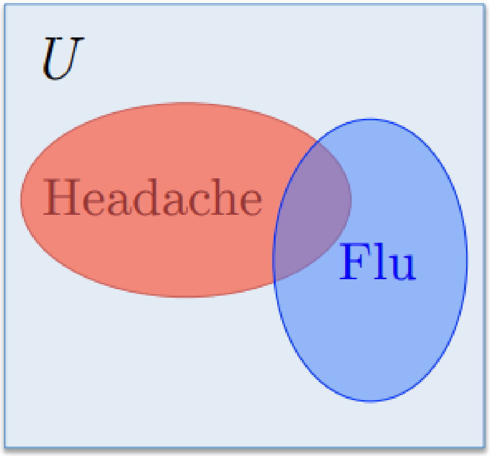
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

"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:

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

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

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

Bayes' 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:

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

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**

Linear algebra

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

Norms

Vector norms: A norm of a vector ||x|| is informally a measure of the "length" of the vector.

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

– Common norms: L_1, L_2 (Euclidean)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

– L_{infinity}

$$\|x\|_\infty = \max_i |x_i|$$

Vector products

We will use lower case letters for vectors The elements are referred by x_i.

- Vector dot (inner) product:

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ x_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

If u•v=0, $||u||_2$!= 0, $||v||_2$!= 0 → *u and v are orthogonal*

If u•v=0, $||u||_2$ = 1, $||v||_2$ = 1 → *u and v are orthonormal*

- Vector outer product:

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

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.

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

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

Properties

- Associativity

$$(AB)C = A(BC)$$ ✓

- Distributivity

$$A(B + C) = AB + AC$$ ✓

- Commutativity

$$AB = BA$$ ✗

Special matrices

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$ Diagonal

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$$ Upper-triangular

$$\begin{pmatrix} a & b & 0 & 0 \\ c & d & e & 0 \\ 0 & f & g & h \\ 0 & 0 & i & j \end{pmatrix}$$ Tri-diagonal

$$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}$$ Lower-triangular

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ I (Identity matrix)

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 $A = A^T$

References

Probability

- Review notes from Stanford's machine learning class
- Sam Roweis's probability review

Linear algebra

- Review notes from Stanford's machine learning class
- Sam Roweis's linear algebra review