

# 2. Basic Statistics

Essential tools for data analysis

# Outline

## Theory:

- Probabilities:
  - Probability measures, events, random variables, conditional probabilities, dependence, expectations, etc
- Bayes rule
- Parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum a Posteriori (MAP)

## Application:

Naive Bayes Classifier for

- Spam filtering
- “Mind reading” = fMRI data processing

# What is the probability?

## Probabilities



Bayes



Kolmogorov

# Probability

- Sample space, Events,  $\sigma$ -Algebras
- Axioms of probability, probability measures
  - What defines a reasonable theory of uncertainty?
- Random variables:
  - discrete, continuous random variables
- Joint probability distribution
- Conditional probabilities
- Expectations
- Independence, Conditional independence

# Sample space

**Def:** A *sample space*  $\Omega$  is the set of all possible outcomes of a (conceptual or physical) random experiment. ( $\Omega$  can be finite or infinite.)

## Examples:

- $\Omega$  may be the set of all possible outcomes of a dice roll (1,2,3,4,5,6)
- Pages of a book opened randomly. (1-157)
- Real numbers for temperature, location, time, etc



# Events

We will ask the question:

**What is the probability of a particular event?**

**Def: *Event A is a subset of the sample space  $\Omega$***

***Examples:***

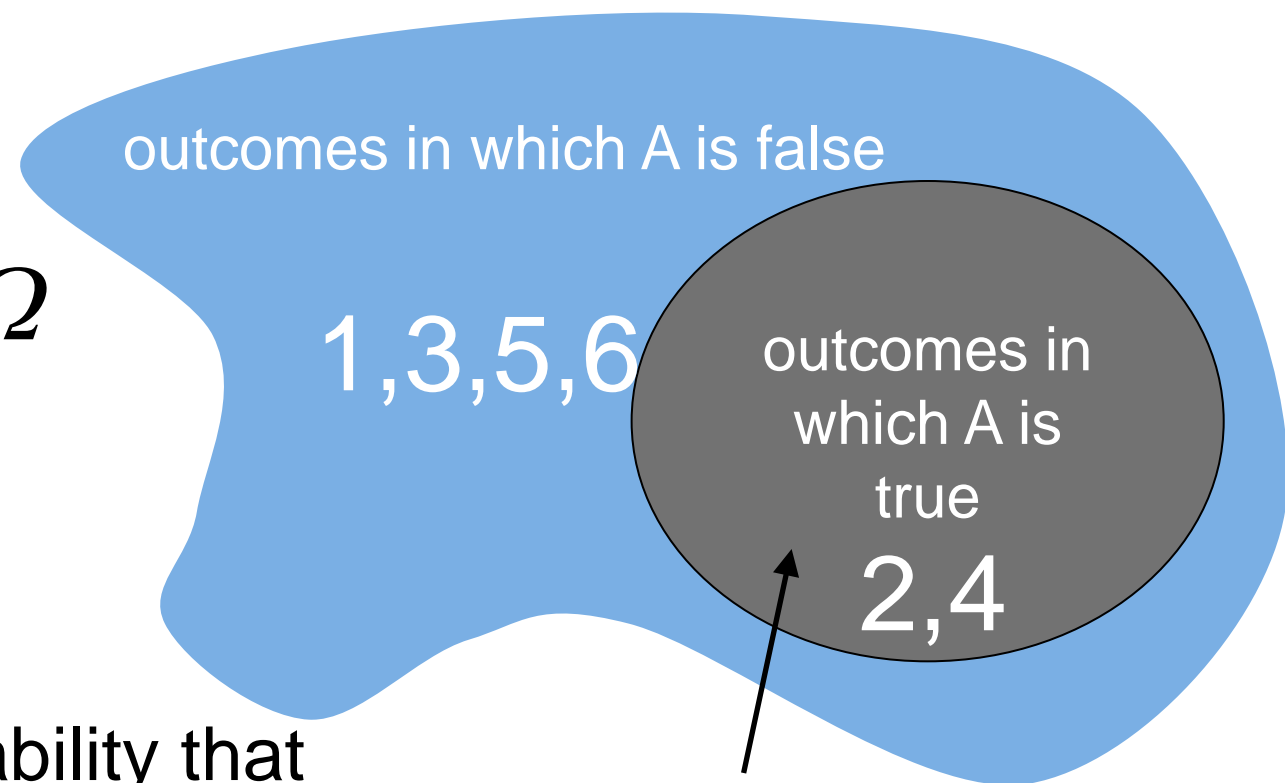
What is the probability of

- *the book is open at an odd number*
- *rolling a dice the number  $< 4$*
- *a random person's height  $X : a < X < b$*

# Probability

**Def:** *Probability  $P(A)$ , the probability that event (subset)  $A$  happens*, is a function that maps the event  $A$  onto the interval  $[0, 1]$ .  $P(A)$  is also called the **probability measure** of  $A$ .

*sample space  $\Omega$*



**Example:** What is the probability that the number on the dice is 2 or 4?

$P(A)$  is the volume of the area.

What defines a reasonable  
theory of uncertainty?



# Kolmogorov Axioms

(i) Nonnegativity:  $P(A) \geq 0$  for each  $A$  event.

(ii)  $P(\Omega) = 1$ .

(iii)  $\sigma$ -additivity: For disjoint sets (events)  $A_i$ , we have

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

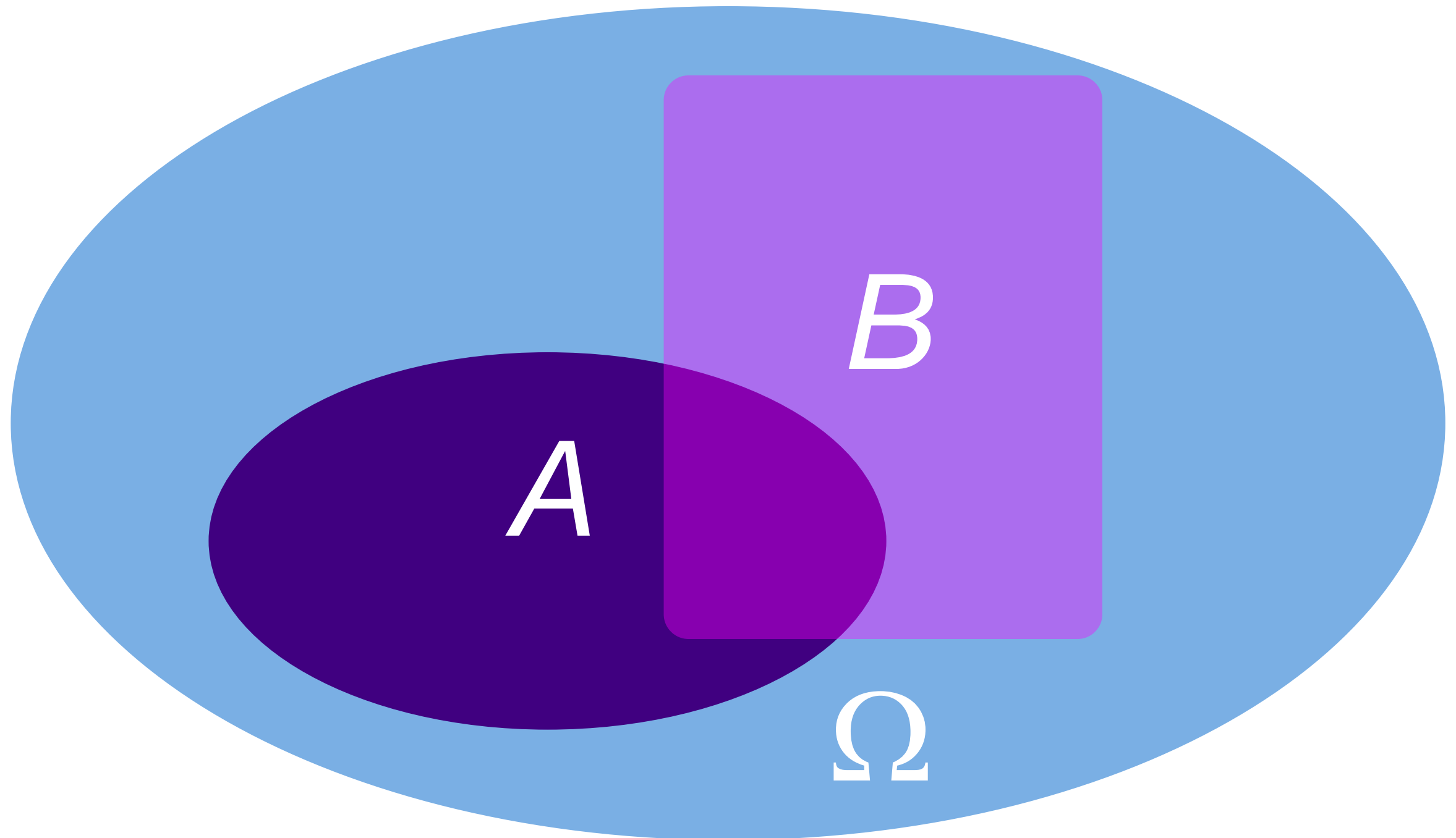
## Consequences:

$$P(\emptyset) = 0.$$

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

$$P(A^c) = 1 - P(A).$$

# Venn Diagram



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

# Random Variables

**Def:** Real valued **random variable** is a function of the outcome of a randomized experiment

$$X : \Omega \rightarrow \mathbb{R}$$

$$P(a < X < b) \doteq P(\omega : a < X(\omega) < b)$$

$$P(X = a) \doteq P(\omega : X(\omega) = a)$$

**Examples:**

- **Discrete random variable examples ( $\Omega$  is discrete):**
- $X(\omega) = \text{True}$  if a randomly drawn person ( $\omega$ ) from our class ( $\Omega$ ) is female
- $X(\omega) = \text{The hometown } X(\omega) \text{ of a randomly drawn person } (\omega) \text{ from our class } (\Omega)$

# Random Variables

Sometimes  $\Omega$  can be quite abstract

$$\Omega = [0, \infty) \times \{1, \dots, 145\}$$

$$\omega = (\omega_1, \omega_2) \in \Omega$$

**Continuous random variable:**

Let  $X(\omega_1, \omega_2) = \omega_1$  be the heart rate of a randomly drawn person  $(\omega = \omega_1, \omega_2)$  in our class  $\Omega$

$$P(a < X < b) \doteq P((\omega_1, \omega_2) : a < X(\omega_1, \omega_2) < b)$$

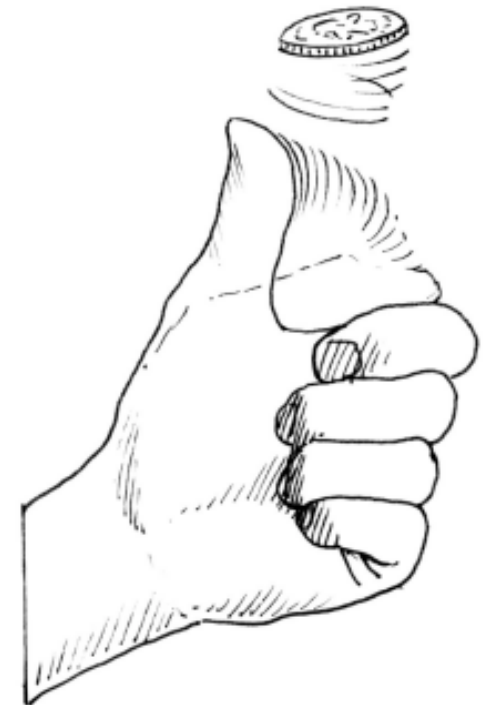
What discrete distributions do we know?

# Discrete Distributions

- Bernoulli distribution:  $\text{Ber}(p)$

$\Omega = \{\text{head}, \text{tail}\}$   $X(\text{head}) = 1$ ,  $X(\text{tail}) = 0$ .

$$P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases}$$



- Binomial distribution:  $\text{Bin}(n, p)$

Suppose a coin with head prob.  $p$  is tossed  $n$  times. What is the probability of getting  $k$  heads and  $n-k$  tails?

$\Omega = \{\text{possible } n \text{ long head/tail series}\}$ ,  $|\Omega| = 2^n$

$K(\omega) = \text{number of heads in } \omega = (\omega_1, \dots, \omega_n) \in \{\text{head}, \text{tail}\}^n = \Omega$

$$P(K = k) = P(\omega : K(\omega) = k) = \sum_{\omega: K(\omega)=k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

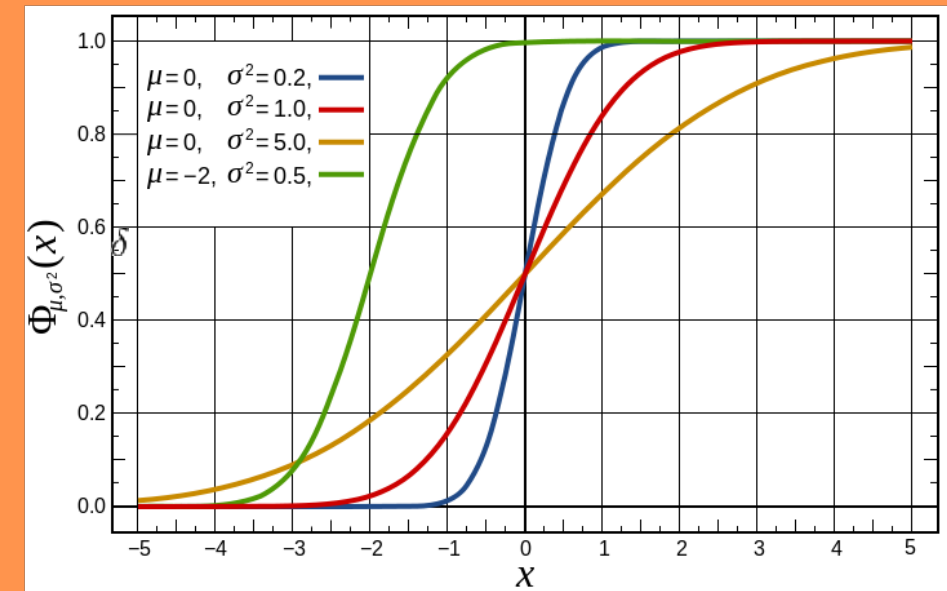
# Continuous Distribution

**Def:** continuous probability distribution: its cumulative distribution function is absolutely continuous.

**Def:** cumulative distribution function

USA:  $F_X(z) = P(X \leq z)$

Hungary:  $F_X(z) = P(X < z)$



**Def :** Let  $F(-\infty) = 0$ .  $F : (-\infty, \infty) \rightarrow \mathbb{R}$  is **absolutely continuous**

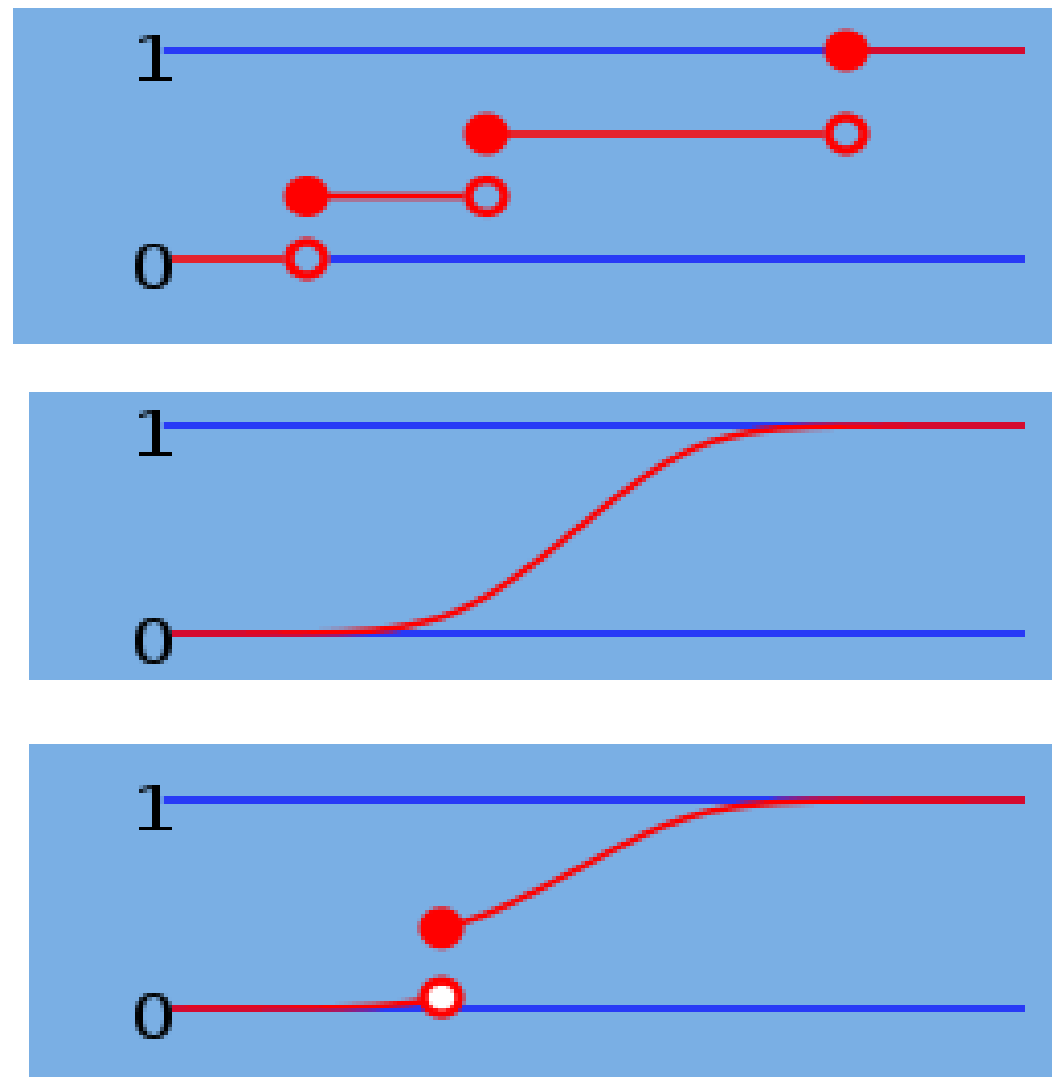
$$F(x) = \int_{-\infty}^x f(t)dt \text{ for some function } f.$$

**Def :**  $f$  is called the density of the distribution.

**Properties :**  $\frac{d}{dx}F(x) = f(x)$

$$F(x) = \int_{-\infty}^x f(t)dt$$

# Cumulative Distribution Function (cdf)



## From top to bottom:

- the cumulative distribution function of a **discrete** probability distribution
- **continuous** probability distribution,
- a distribution which has both a **continuous part** and a **discrete part**.



# Cumulative Distribution Function (cdf)

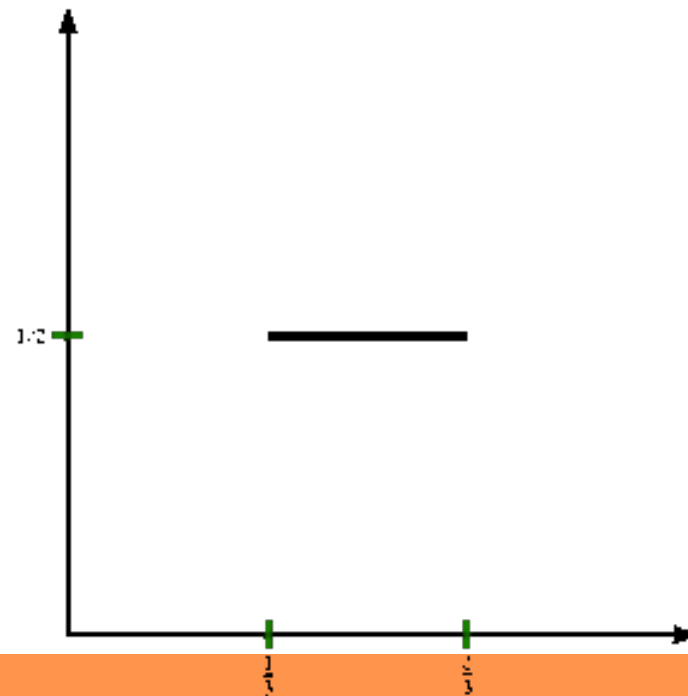
If the CDF is **absolute continuous**, then the distribution has **density** function.

$$\frac{d}{dx}F(x) = f(x) \qquad F(x) = \int_{-\infty}^x f(t)dt$$

Why do we need **absolute** continuity?

**Continuity** of the CDF is not enough to have density function??

$$F(x) \neq \int_{-\infty}^x f(t)dt = 0$$



**Cantor function:** F continuous everywhere, has zero derivative ( $f=0$ ) almost everywhere, F goes from 0 to 1 as x goes from 0 to 1, and takes on every value in between.  $\Rightarrow$  there is **no density** for the Cantor function CDF.

# Probability Density Function (pdf)

## Pdf properties:

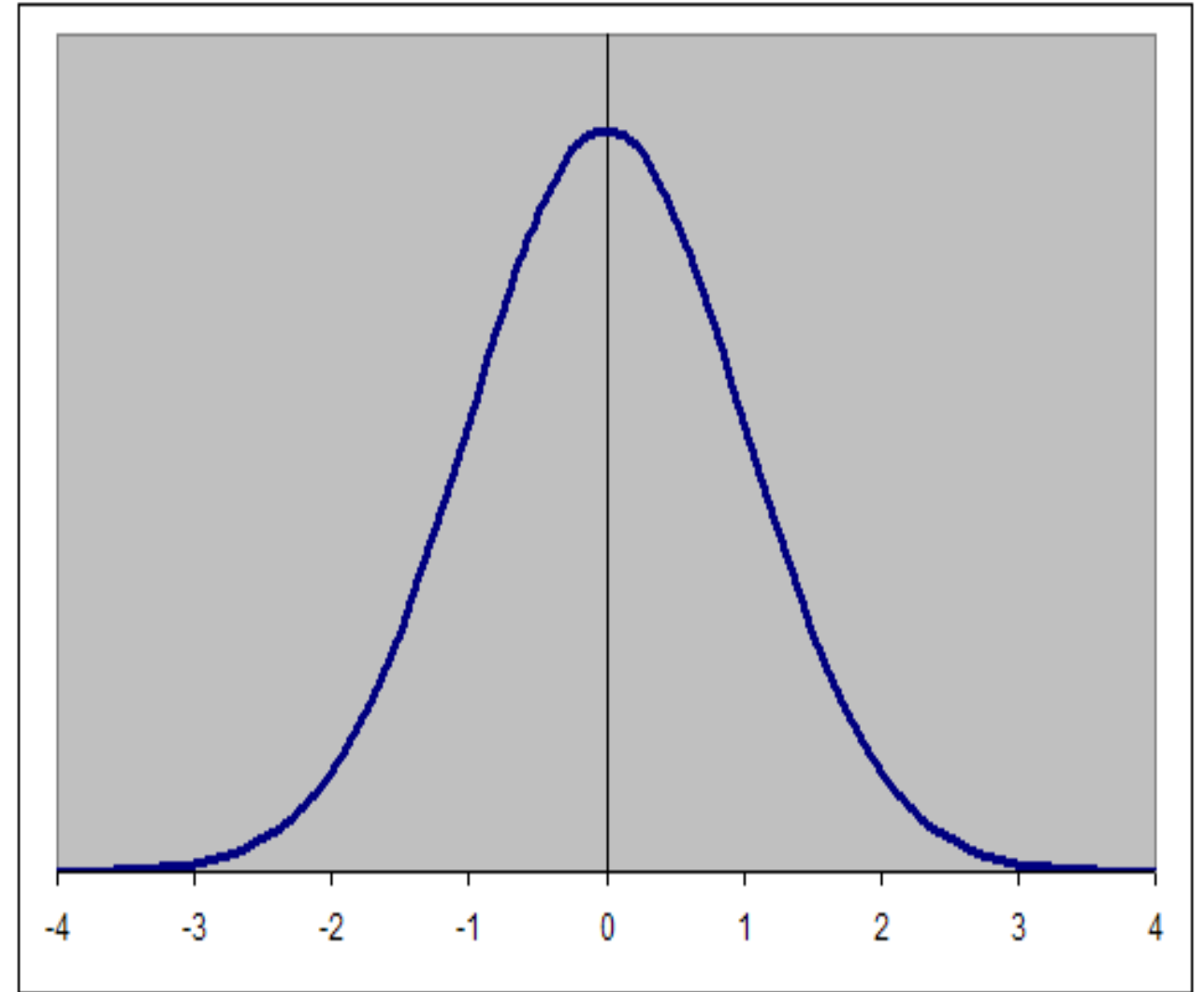
$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$f(x) = \frac{d}{dx} F(x)$$

$$F(x) = \int_{-\infty}^x f(t) dt$$

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$



Intuitively, one can think of  $f(x)dx$  as being the probability of  $X$  falling within the infinitesimal interval  $[x, x + dx]$ .  $P(x < X < x + dx) = f(x)dx$

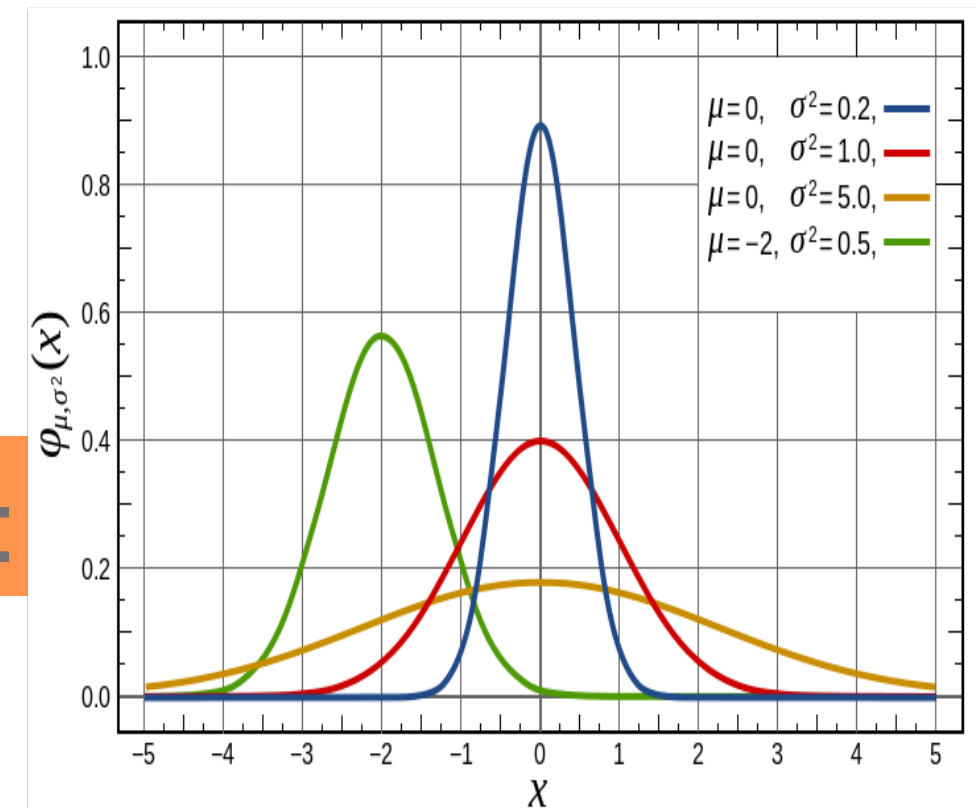
# Moments

**Expectation:** average value, mean, 1<sup>st</sup> moment:

$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$

**Variance:** the spread, 2<sup>nd</sup> moment:

$$E(X) = \begin{cases} \sum_{i \in \Omega} [x_i - E(X)]^2 p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - E(x))^2 p(x) dx & \text{continuous} \end{cases}$$

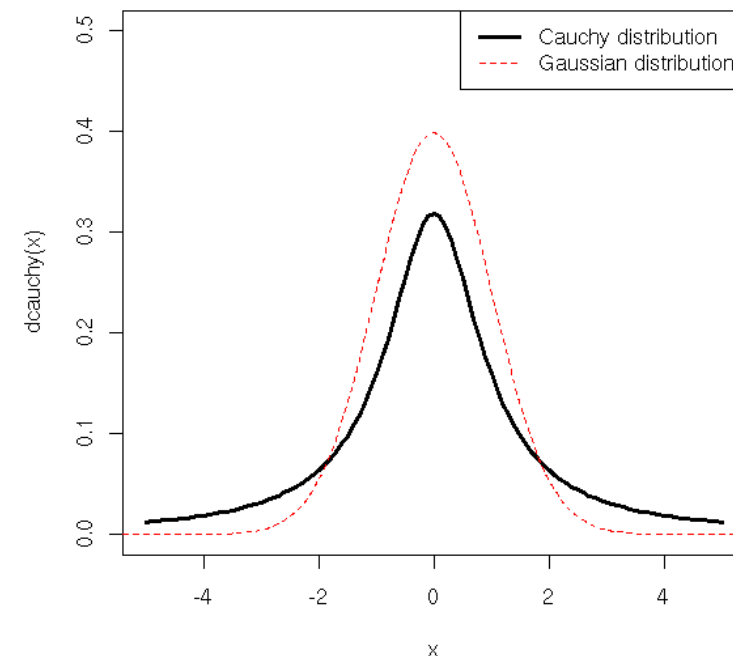


# Warning!

Moments may not always exist!

## Cauchy distribution

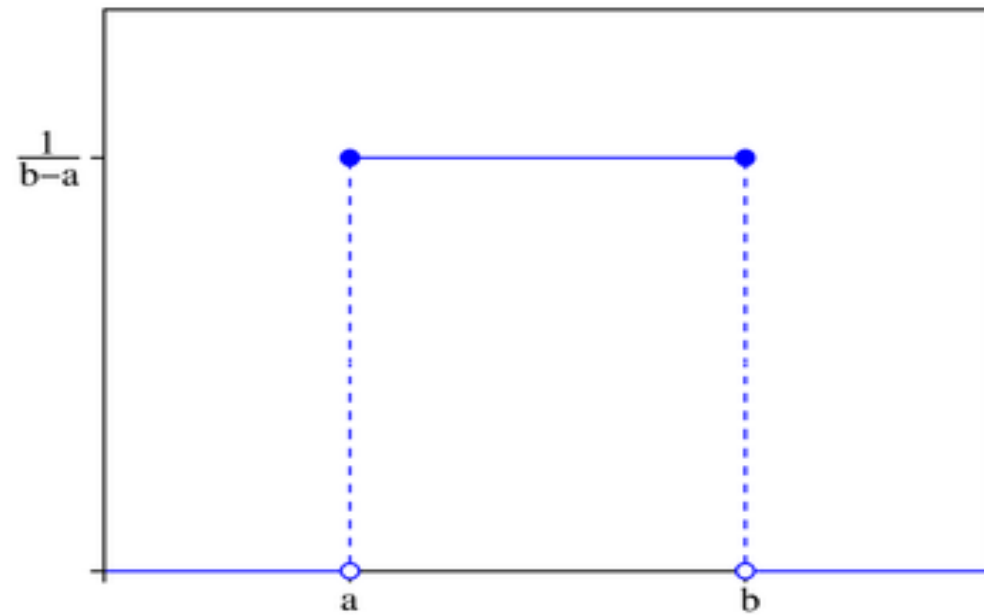
$$p(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$$



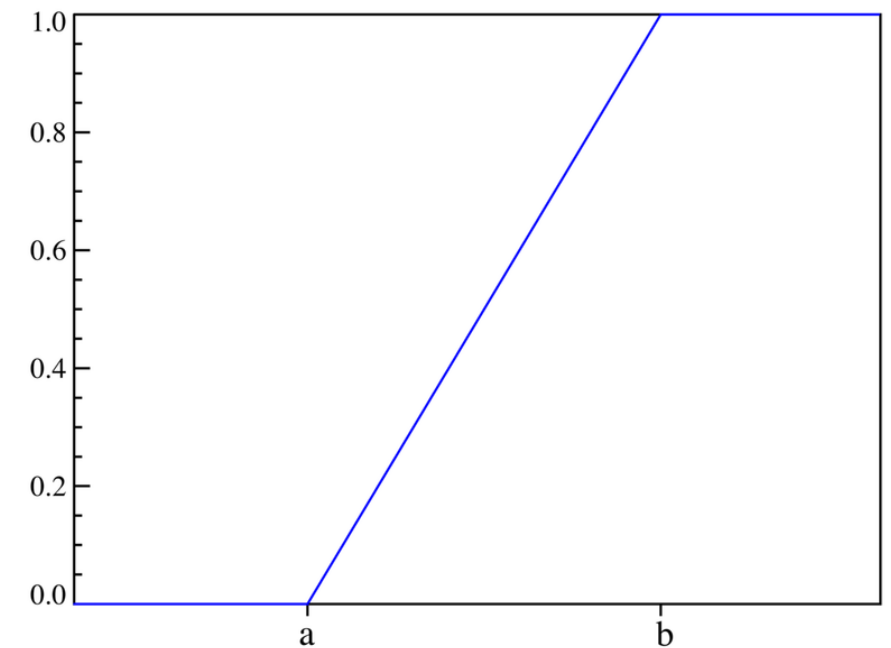
For the mean to exist the following integral would have to converge

$$\begin{aligned} \int_{-\infty}^{\infty} |x| p(x) dx &= \int_{-\infty}^{\infty} |x| \frac{1}{\pi} \frac{1}{1 + x^2} dx = 2 \int_0^{\infty} x \frac{1}{\pi} \frac{1}{1 + x^2} dx \\ &\geq \frac{1}{\pi} \int_1^{\infty} \frac{2x}{1 + x^2} dx \geq \frac{1}{\pi} \int_1^{\infty} \frac{1}{x} dx = \infty \end{aligned}$$

# Uniform Distribution



PDF

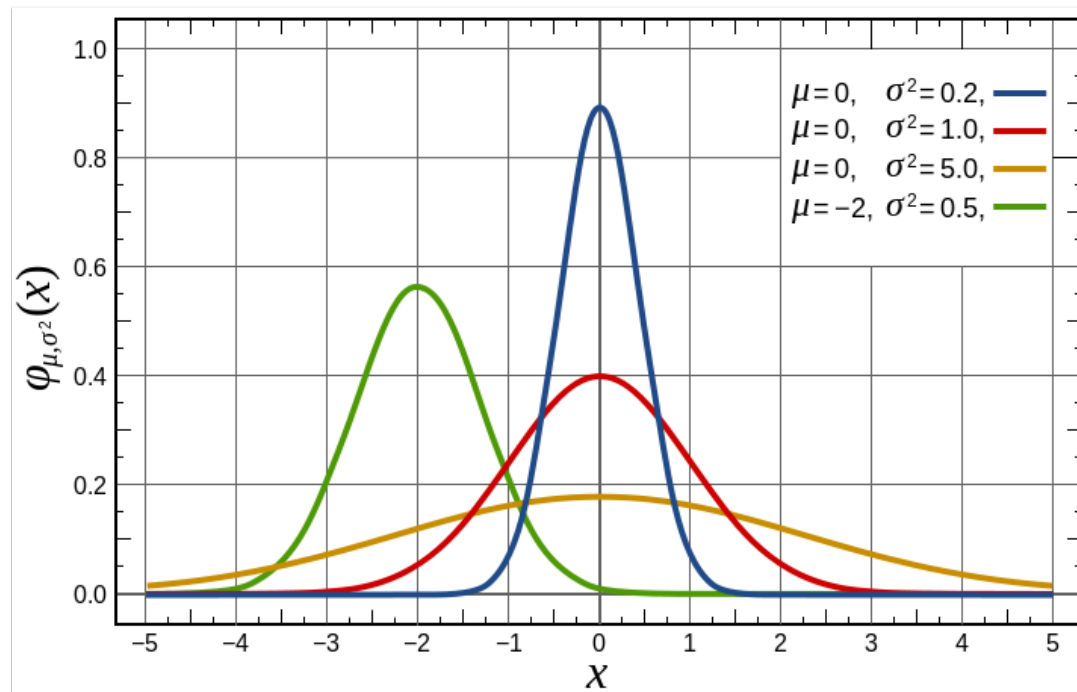


CDF

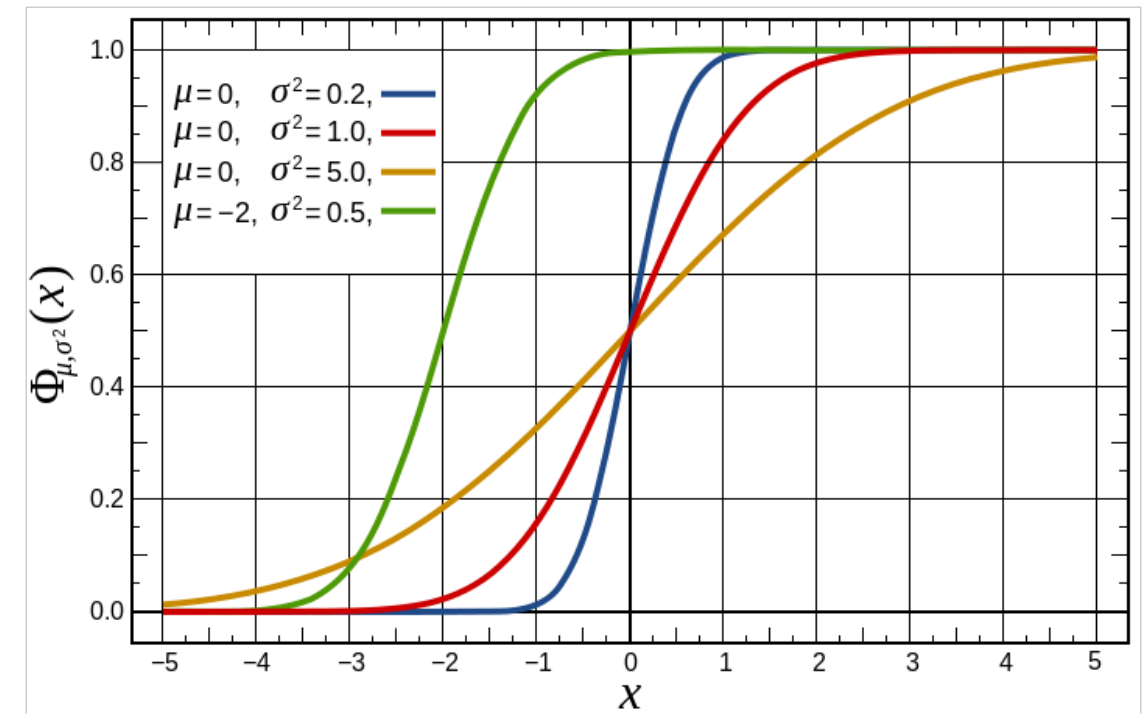
$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{Otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x \leq b \\ 1 & b < x \end{cases}$$

# Normal (Gaussian) Distribution



PDF



CDF

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

$$F(x) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2\sigma^2}} \right) \right]$$

# Multivariate (Joint) Distribution

We can generalize the above ideas from 1-dimension to any finite dimensions.

$$P(a \leq X \leq b, c \leq Y \leq d) = ?$$

$$P(a_1 \leq X_1 \leq b_1, \dots, a_d \leq X_d \leq b_d) = ?$$

Discrete distribution:

$$P(\text{headache} \wedge \text{no flu}) = 7/80$$

$$P(\text{headache}) = 7/80 + 1/80$$

$$P(X = \text{headache}, Y = \text{flu}) = 1/80$$

	Flu	No Flu
Headache	1/80	7/80
No Headache	1/80	71/80

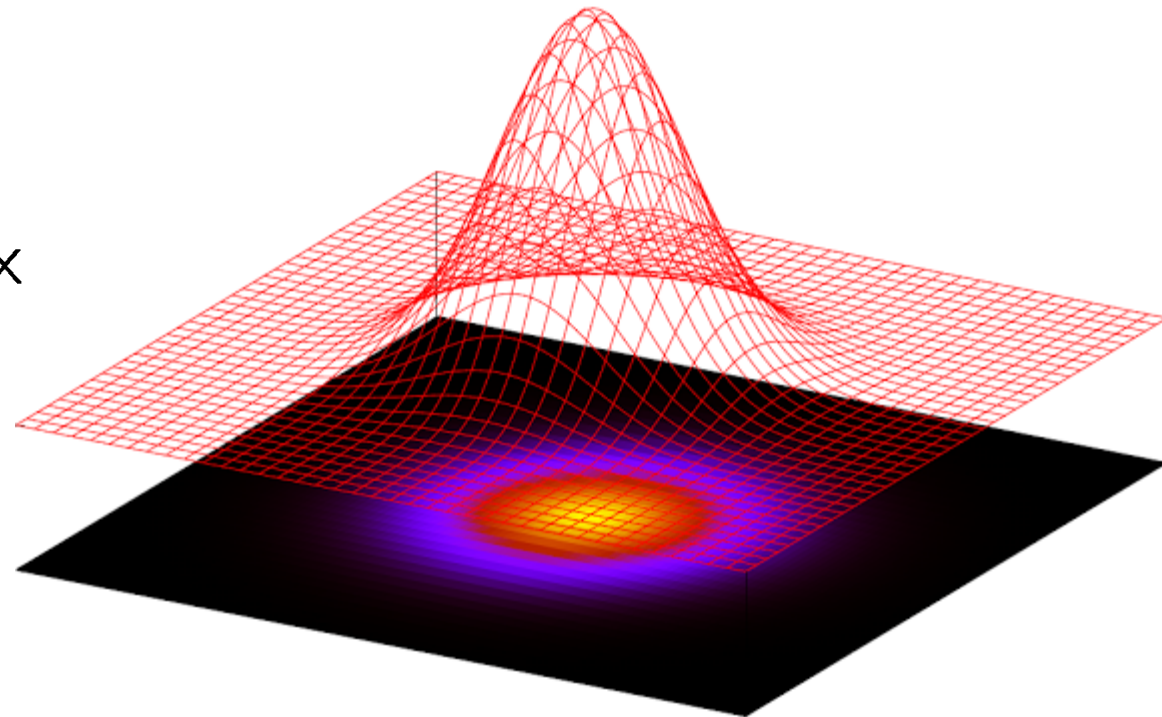
# Multivariate Gaussian distribution

For  $A \subset \mathbb{R}^d$ ,  $P([X_1, \dots, X_d] \in A) = \int_A f(x_1, \dots, x_d) dx_1 \cdots dx_d$

$$F_X(z_1, \dots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \dots, x_d) dx_1 \cdots dx_d \quad \text{Multivariate CDF}$$

$\mu \in \mathbb{R}^d$  : mean vector

$\Sigma \in \mathbb{R}^{d \times d}$  : covariance matrix



<http://www.moserware.com/2010/03/computing-your-skill.htm>

$$f_X(x_1, \dots, x_d) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp \left( -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)$$



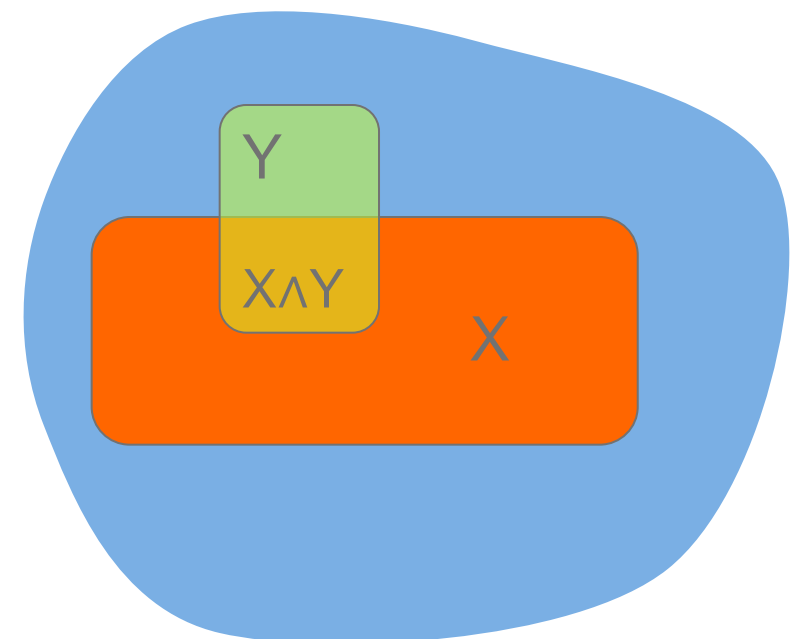
# Conditional Probability

$P(X|Y)$  = Fraction of worlds in which  $X$  event is true given  $Y$  event is true.

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

$$P(\text{flu}|\text{headache}) = \frac{P(\text{flu, headache})}{P(\text{headache})} = \frac{1/80}{1/80 + 7/80}$$

	Flu	No Flu
Headache	1/80	7/80
No Headache	1/80	71/80



# Independence

**Independent random variables:**

$$P(X, Y) = P(X)P(Y)$$

$$P(X|Y) = P(X)$$

Y and X don't contain information about each other.

Observing Y doesn't help predicting X.

Observing X doesn't help predicting Y.

## **Examples:**

Independent: Winning on roulette this week and next week.

Dependent: Russian roulette

# Conditionally Independent

**Conditionally independent:**

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Knowing Z makes X and Y independent

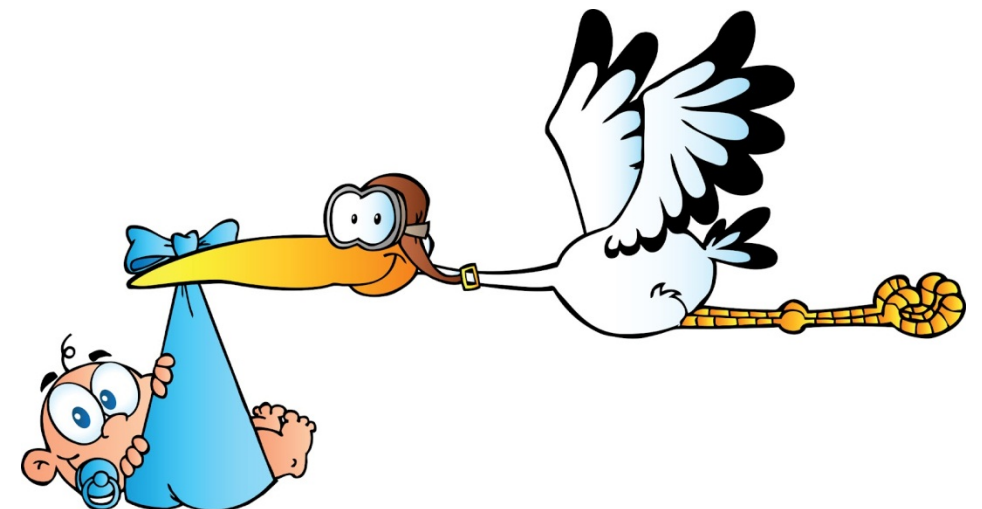
**Examples:**

Dependent: show size and reading skills

Conditionally independent: show size and reading skills given **age**

**Storks deliver babies:**

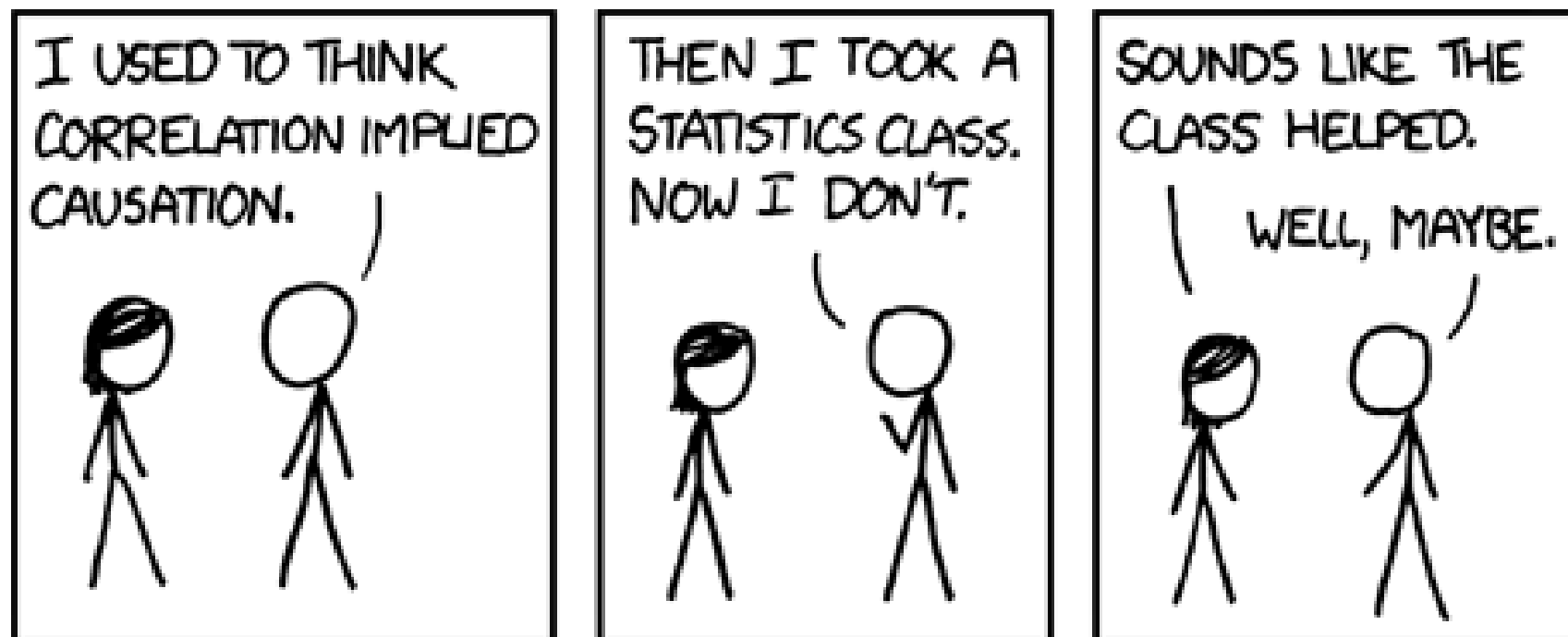
Highly statistically significant correlation exists between stork populations and human birth rates across Europe



# Conditionally Independent

**London taxi drivers:** A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains...



# Conditional Independence

Formally:  $X$  is **conditionally independent** of  $Y$  given  $Z$ :

$$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

$$P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain})$$

Equivalent to:

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

# Bayes Rule

# Chain Rule & Bayes Rule

Chain rule:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Bayes rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes rule is important for reverse conditioning.

# AIDS test (Bayes rule)

## Data

- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

Probability of having AIDS if test is positive:

$$\begin{aligned} P(a = 1|t = 1) &= \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1)} \\ &= \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1|a = 1)P(a = 1) + P(t = 1|a = 0)P(a = 0)} \\ &= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091 \end{aligned}$$

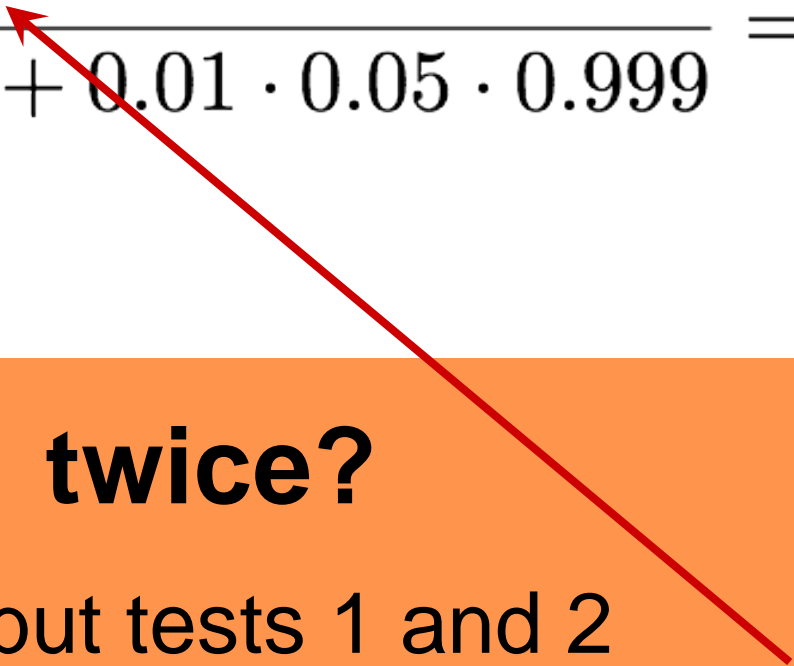
Only 9%!...



# Improving the diagnosis

## Use a follow-up test!

- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

$$\begin{aligned} P(a = 0 | t_1 = 1, t_2 = 1) &= \frac{P(t_1 = 1, t_2 = 1 | a = 0)P(a = 0)}{P(t_1 = 1, t_2 = 1 | a = 1)P(a = 1) + P(t_1 = 1, t_2 = 1 | a = 0)P(a = 0)} \\ &= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357 \\ P(a = 1 | t_1 = 1, t_2 = 1) &= 0.643 \end{aligned}$$


## Why can't we use Test 1 twice?

Outcomes are **not** independent but tests 1 and 2

are **conditionally independent**  $p(t_1, t_2 | a) = p(t_1 | a) \cdot p(t_2 | a)$

# Application: Document Classification, Spam filtering



# Data for spam filtering

- date
- time
- recipient path
- IP number
- sender
- encoding
- many more features

Delivered-To: [alex.smola@gmail.com](mailto:alex.smola@gmail.com)  
Received: by 10.216.47.73 with SMTP id s51cs361171web;  
Tue, 3 Jan 2012 14:17:53 -0800 (PST)  
Received: by 10.213.17.145 with SMTP id s17mr2519891eba.147.1325629071725;  
Tue, 03 Jan 2012 14:17:51 -0800 (PST)  
Return-Path: <alex+caf\_alex.smola@gmail.com@smola.org>  
Received: from mail-ey0-f175.google.com (mail-ey0-f175.google.com [209.85.215.175])  
by mx.google.com with ESMTPS id n4si29264232eef.57.2012.01.03.14.17.51  
(version=TLSv1/SSLv3 cipher=OTHER);  
Tue, 03 Jan 2012 14:17:51 -0800 (PST)  
Received-SPF: neutral (google.com: 209.85.215.175 is neither permitted nor denied by best guess record for domain of alex+caf\_alex.smola@gmail.com@smola.org) client-ip=209.85.215.175;  
Authentication-Results: mx.google.com; spf=neutral (google.com: 209.85.215.175 is neither permitted nor denied by best  
guess record for domain of alex+caf\_alex.smola@gmail.com@smola.org)  
smtp.mail=alex+caf\_alex.smola@gmail.com@smola.org; dkim=pass (test mode) header.i=@googlemail.com  
Received: by eaal1 with SMTP id l1so15092746eaa.6  
for <alex.smola@gmail.com>; Tue, 03 Jan 2012 14:17:51 -0800 (PST)  
Received: by 10.205.135.18 with SMTP id ie18mr5325064bkc.72.1325629071362;  
Tue, 03 Jan 2012 14:17:51 -0800 (PST)  
X-Forwarded-To: [alex.smola@gmail.com](mailto:alex.smola@gmail.com)  
X-Forwarded-For: [alex@smola.org](mailto:alex@smola.org) [alex.smola@gmail.com](mailto:alex.smola@gmail.com)  
Delivered-To: [alex@smola.org](mailto:alex@smola.org)  
Received: by 10.204.65.198 with SMTP id k6cs206093bki;  
Tue, 3 Jan 2012 14:17:50 -0800 (PST)  
Received: by 10.52.88.179 with SMTP id bh19mr10729402vdb.38.1325629068795;  
Tue, 03 Jan 2012 14:17:48 -0800 (PST)  
Return-Path: <althoff.tim@googlemail.com>  
Received: from mail-vx0-f179.google.com (mail-vx0-f179.google.com [209.85.220.179])  
by mx.google.com with ESMTPS id dt4si11767074vdb.93.2012.01.03.14.17.48  
(version=TLSv1/SSLv3 cipher=OTHER);  
Tue, 03 Jan 2012 14:17:48 -0800 (PST)  
Received-SPF: pass (google.com: domain of [althoff.tim@googlemail.com](mailto:althoff.tim@googlemail.com) designates 209.85.220.179 as permitted sender)  
client-ip=209.85.220.179;  
Received: by vcbf13 with SMTP id f13so11295098vcb.10  
for <alex@smola.org>; Tue, 03 Jan 2012 14:17:48 -0800 (PST)  
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;  
d=googlemail.com; s=gamma;  
h=mime-version:sender:date:x-google-sender-auth:message-id:subject  
:from:to:content-type;  
bh=WcBdZ5sXac25dpH02XcRyDOdts993hKwsAVXpGrFh0w=;  
b=WK2B2+ExWnf/gvTkw6uUvKuP4XeoKnIJq3USYtM0RARK8dSFjyOQsIHeAP9Yssxp6O  
7ngGoTzYqd+ZsyJfvQcLAWp1PCJhG8AMcnqWkx0NMeoFvlp2HQooZwxSOCx5ZRgY+7qX  
ulbbdna4IUDXj6UFe16SpLDCkptd8OZ3gr7+o=  
MIME-Version: 1.0  
Received: by 10.220.108.81 with SMTP id e17mr24104004vcp.67.1325629067787;  
Tue, 03 Jan 2012 14:17:47 -0800 (PST)  
Sender: [althoff.tim@googlemail.com](mailto:althoff.tim@googlemail.com)  
Received: by 10.220.17.129 with HTTP; Tue, 3 Jan 2012 14:17:47 -0800 (PST)  
Date: Tue, 3 Jan 2012 14:17:47 -0800  
X-Google-Sender-Auth: 6bwi6D17HjZIkxOEol38NZzyeHs  
Message-ID: <CAFJJHDGPBW+SdZg0MdAABiAKyDk9tpeMoDijYGjoGO-WC7osg@mail.gmail.com>  
Subject: CS 281B. Advanced Topics in Learning and Decision Making  
From: Tim Althoff <[althoff@eecs.berkeley.edu](mailto:althoff@eecs.berkeley.edu)>  
To: [alex@smola.org](mailto:alex@smola.org)  
Content-Type: multipart/alternative; boundary=f46d043c7af4b07e8d04b5a7113a  
  
--f46d043c7af4b07e8d04b5a7113a  
Content-Type: text/plain; charset=ISO-8859-1

# Naïve Bayes Assumption

**Naïve Bayes assumption:** Features  $X_1$  and  $X_2$  are conditionally independent given the class label  $Y$ :

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

More generally:

$$P(X_1 \dots X_d|Y) = \prod_{i=1}^d P(X_i|Y)$$

How many parameters to estimate?

( $X$  is composed of  $d$  binary features, e.g. presence of “earn”  
 $Y$  has  $K$  possible class labels)

**$(2^d - 1)K$  vs  $(2 - 1)dK$**

# Naïve Bayes Classifier

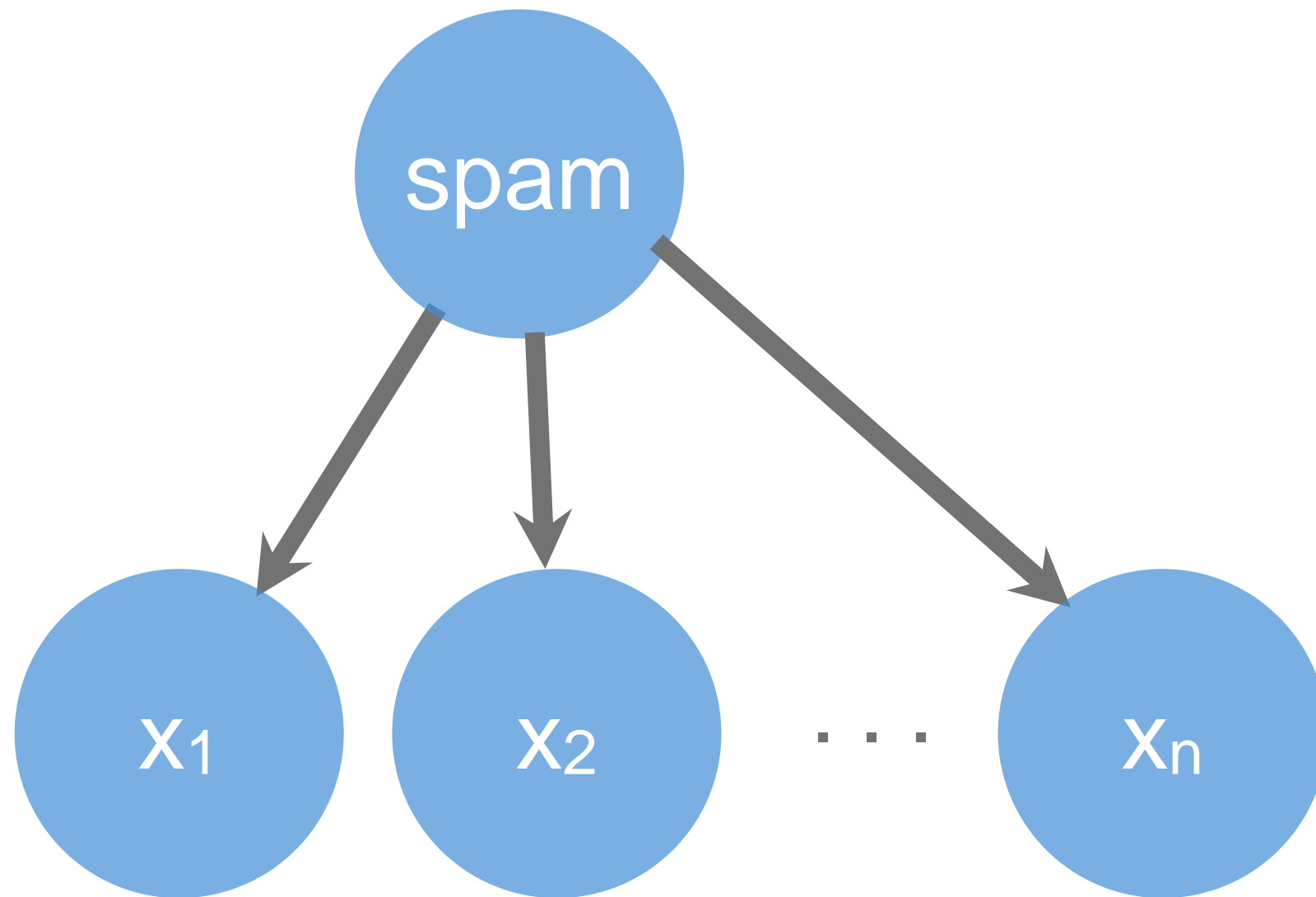
## Given:

- Class prior  $P(Y)$
- $d$  conditionally independent features  $X_1, \dots, X_d$  given the class label  $Y$
- For each  $X_i$ , we have the conditional likelihood  $P(X_i/Y)$

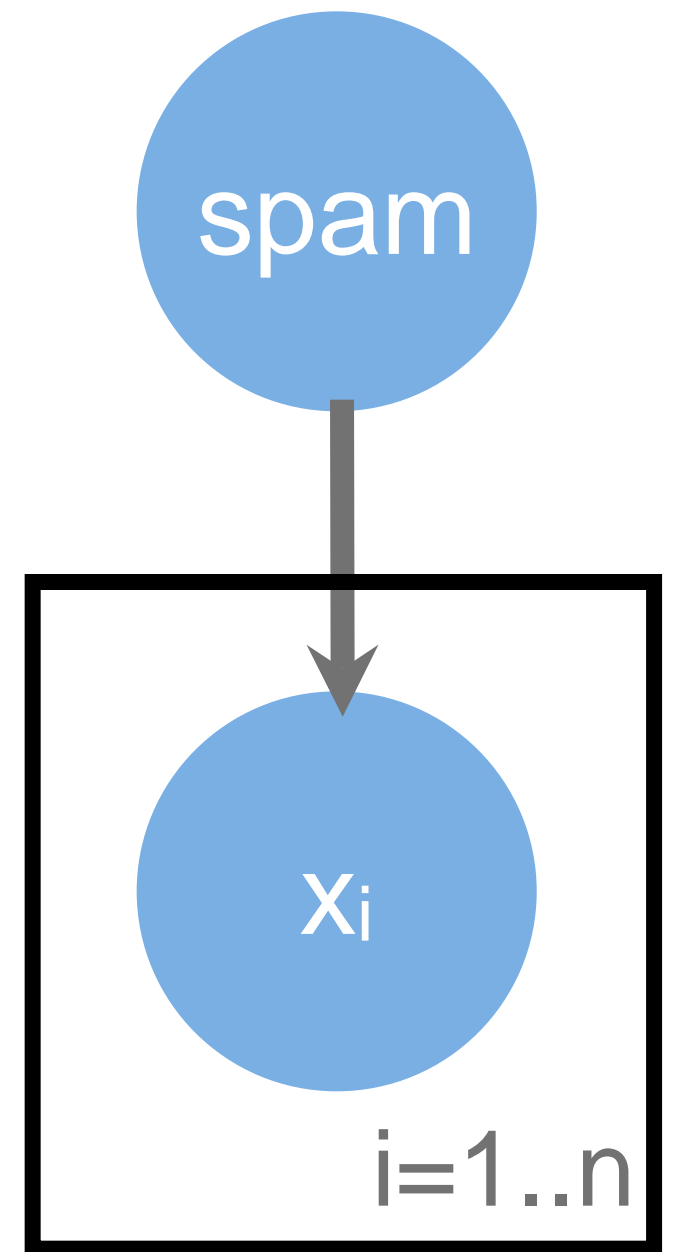
## Decision rule:

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg \max_y P(x_1, \dots, x_d | y) P(y) \\ &= \arg \max_y \prod_{i=1}^d P(x_i|y) P(y) \end{aligned}$$

# A Graphical Model



$$P(X_1, \dots, X_d \mid \text{spam}) = \prod_{i=1}^d P(X_i \mid \text{spam})$$



# Naïve Bayes Algorithm for discrete features

Training Data:  $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$      $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$   
 $n$   $d$  dimensional features + class labels

$$f_{NB}(\mathbf{x}) = \arg \max_y \prod_{i=1}^d P(x_i|y) P(y) \quad \text{We need to estimate these probabilities!}$$

Estimate them with Relative Frequencies!

For Class Prior

$$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

For Likelihood

$$\frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

NB Prediction for test data:

$$X = (x_1, \dots, x_d)$$

$$Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

# Subtlety: Insufficient training data

What if you never see a training instance where  $X_1 = a$  when  $Y = b$ ?

For example,

there is no  $X_1 = \text{'Earn'}$  when  $Y = \text{'SpamEmail'}$  in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a | Y = b) = 0$$

$$\Rightarrow P(X_1 = a, X_2 \dots X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^d P(X_i | Y) = 0$$

Thus, no matter what the values  $X_2, \dots, X_d$  take:

$$P(Y = b | X_1 = a, X_2, \dots, X_d) = 0$$

What now???



# Parameter estimation: MLE, MAP

Estimating Probabilities



# Flipping a Coin

I have a coin, if I flip it, what's the probability it will fall with the head up?

Let us flip it a few times to estimate the probability:



The estimated probability is:  $\frac{3}{5}$  "Frequency of heads"

**Why???... and How good is this estimation???**

# MLE for Bernoulli distribution

Data,  $D =$



$$D = \{X_i\}_{i=1}^n, \quad X_i \in \{H, T\}$$

$$P(\text{Heads}) = \theta, \quad P(\text{Tails}) = 1 - \theta$$

Flips are **i.i.d.**:

- **Independent** events
- **Identically distributed** according to Bernoulli distribution

MLE: Choose  $\theta$  that maximizes the probability of observed data

# Maximum Likelihood Estimation

MLE: Choose  $\theta$  that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D \mid \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i \mid \theta) && \text{Independent draws} \\ &= \arg \max_{\theta} \prod_{i: X_i=H} \theta \prod_{i: X_i=T} (1 - \theta) && \text{Identically distributed} \\ &= \arg \max_{\theta} \underbrace{\theta^{\alpha_H} (1 - \theta)^{\alpha_T}}_{J(\theta)}\end{aligned}$$

# Maximum Likelihood Estimation

MLE: Choose  $\theta$  that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D \mid \theta) \\ &= \arg \max_{\theta} \underbrace{\theta^{\alpha_H} (1 - \theta)^{\alpha_T}}_{J(\theta)}\end{aligned}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T - 1} \Big|_{\theta = \hat{\theta}_{MLE}} = 0$$

$$\alpha_H (1 - \theta) - \alpha_T \theta \Big|_{\theta = \hat{\theta}_{MLE}} = 0$$

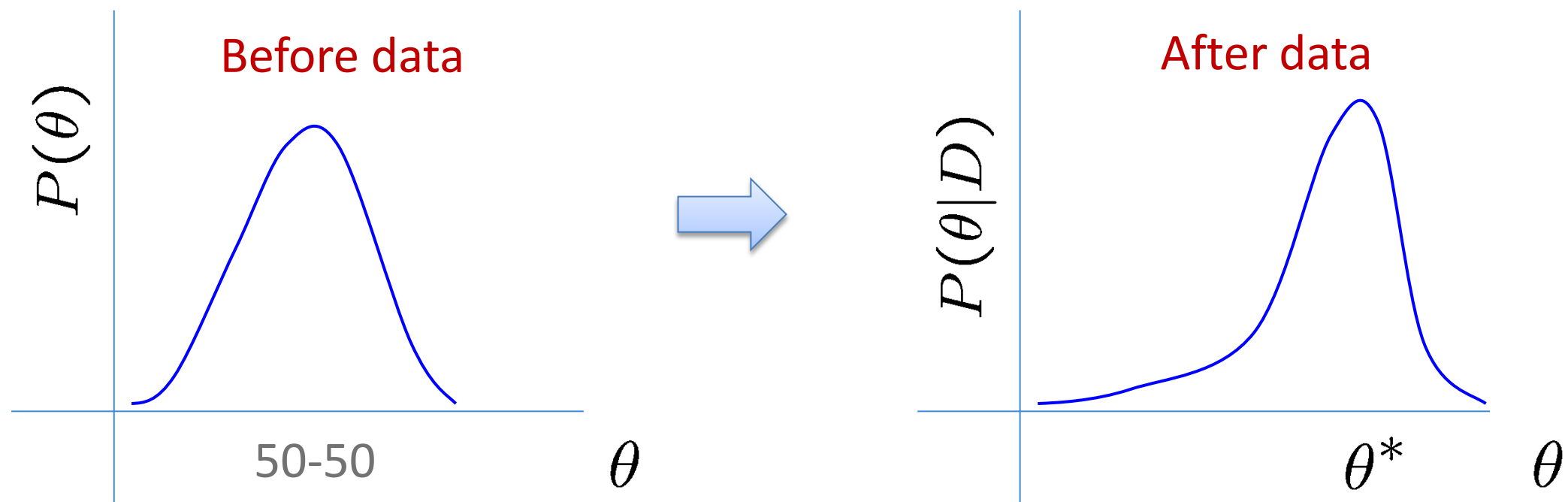
$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

# What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

## The Bayesian way...

Rather than estimating a single  $\theta$ , we obtain a distribution over possible values of  $\theta$



# Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

posterior                      likelihood   prior



**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

# MAP estimation for Binomial distribution

## Coin flip problem

Likelihood is Binomial  $P(\mathcal{D} | \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

$\Rightarrow$  posterior is Beta distribution

$$P(\theta | D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$P(\theta)$  and  $P(\theta | D)$  have the same form! [Conjugate prior]

$$\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} P(D | \theta) P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$



# MLE vs. MAP

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta|D) \\ &= \arg \max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?

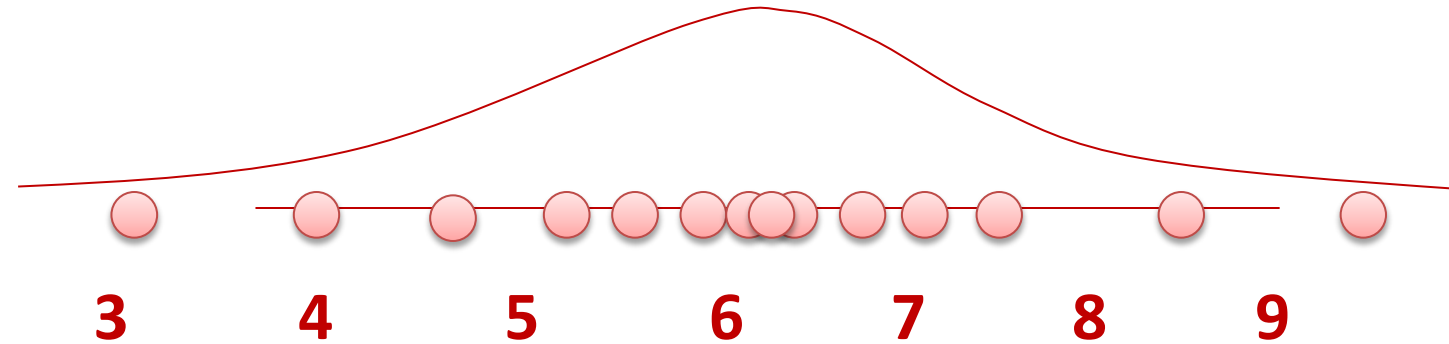
# Bayesians vs. Frequentists

You are no good when sample is small



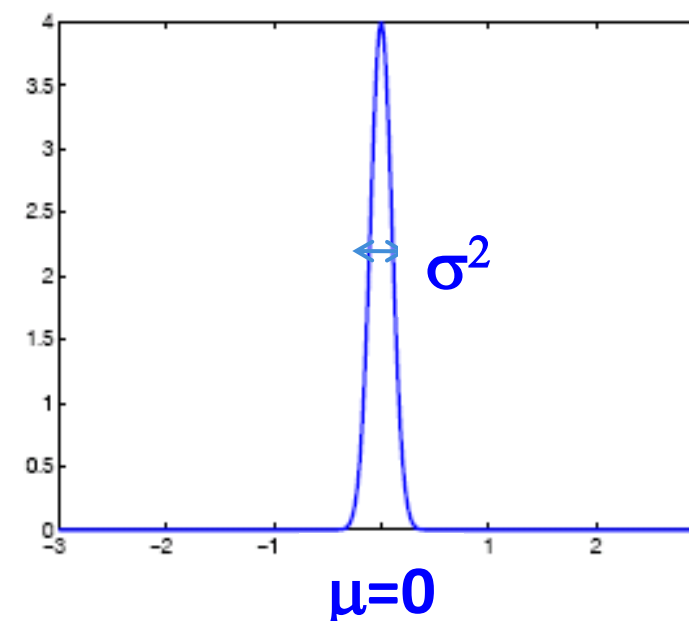
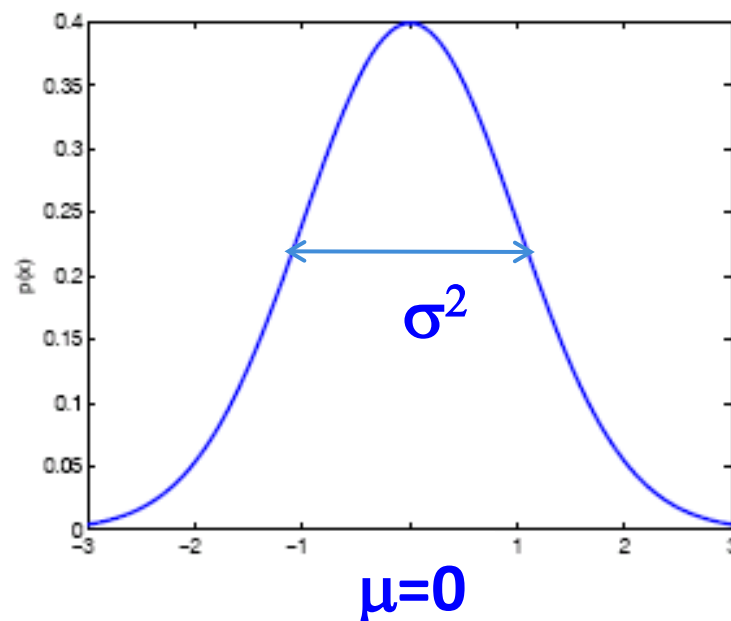
You give a different answer for different priors

# What about continuous features?



Let us try Gaussians...

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma)$$



# MLE for Gaussian mean and variance

Choose  $\theta = (\mu, \sigma^2)$  that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(X_i | \theta) && \text{Independent draws} \\ &= \arg \max_{\theta} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \mu)^2 / (2\sigma^2)} && \text{Identically distributed} \\ &= \arg \max_{\theta = (\mu, \sigma^2)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (X_i - \mu)^2 / (2\sigma^2)}}_{J(\theta)}\end{aligned}$$

# MLE for Gaussian mean and variance

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

**Note:** MLE for the variance of a Gaussian is **biased**

[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator:  $\hat{\sigma}_{unbiased}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$

# Case Study: Text Classification

# Case Study: Text Classification

- Classify e-mails
  - $Y = \{\text{Spam}, \text{NotSpam}\}$
- Classify news articles
  - $Y = \{\text{what is the topic of the article?}\}$

What about the features  $X$ ?

The text!

# $X_i$ represents $i^{\text{th}}$ word in document

## Article from rec.sport.hockey

---

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e  
From: xxx@yyy.zzz.edu (John Doe)  
Subject: Re: This year's biggest and worst (opinion)  
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he's clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he's only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided



# NB for Text Classification

$P(\mathbf{X}|Y)$  is huge!!!

- Article at least 1000 words,  $\mathbf{X}=\{X_1,\dots,X_{1000}\}$
- $X_i$  represents  $i^{\text{th}}$  word in document, i.e., the domain of  $X_i$  is entire vocabulary, e.g., Webster Dictionary (or more).  
 $X_i \in \{1,\dots,50000\} \Rightarrow K1000^{50000}$  parameters....

NB assumption helps a lot!!!

- $P(X_i=x_i|Y=y)$  is the probability of observing word  $x_i$  at the  $i^{\text{th}}$  position in a document on topic  $y \Rightarrow 1000K50000$  parameters

$$h_{NB}(\mathbf{x}) = \arg \max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

# Bag of words model

Typical additional assumption – **Position in document doesn't matter**:  $P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y)$

- “Bag of words” model – order of words on the page ignored
- Sounds really silly, but often works very well!  $\Rightarrow$  K50000 parameters

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^W P(w|y)^{count_w}$$

**When the lecture is over, remember to wake up the person sitting next to you in the lecture room.**

# Bag of words model

Typical additional assumption – **Position in document doesn't matter**:  $P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y)$

- “Bag of words” model – order of words on the page ignored
- Sounds really silly, but often works very well!

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^W P(w|y)^{count_w}$$

in is lecture lecture next over person remember room  
sitting the the the to to up wake when you

# Bag of words approach

the world of

**TOTAL**



**all about the company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

► All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

# Twenty news groups results

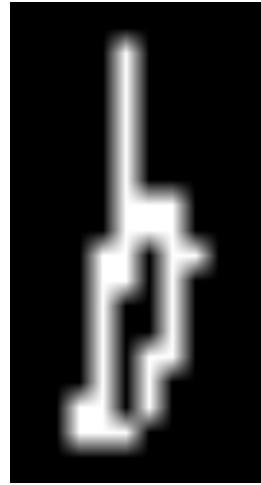
Given 1000 training documents from each group  
Learn to classify new documents according to  
which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

**Naïve Bayes: 89% accuracy**

# What if features are continuous?

Eg., character recognition:  $X_i$  is intensity at  $i^{\text{th}}$  pixel



**Gaussian Naïve Bayes (GNB):**

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Different mean and variance for each class  $k$  and each pixel  $i$ .

Sometimes assume variance

- is independent of  $Y$  (i.e.,  $\sigma_i$ ),
- or independent of  $X_i$  (i.e.,  $\sigma_k$ )
- or both (i.e.,  $\sigma$ )

# Example: GNB for classifying mental states



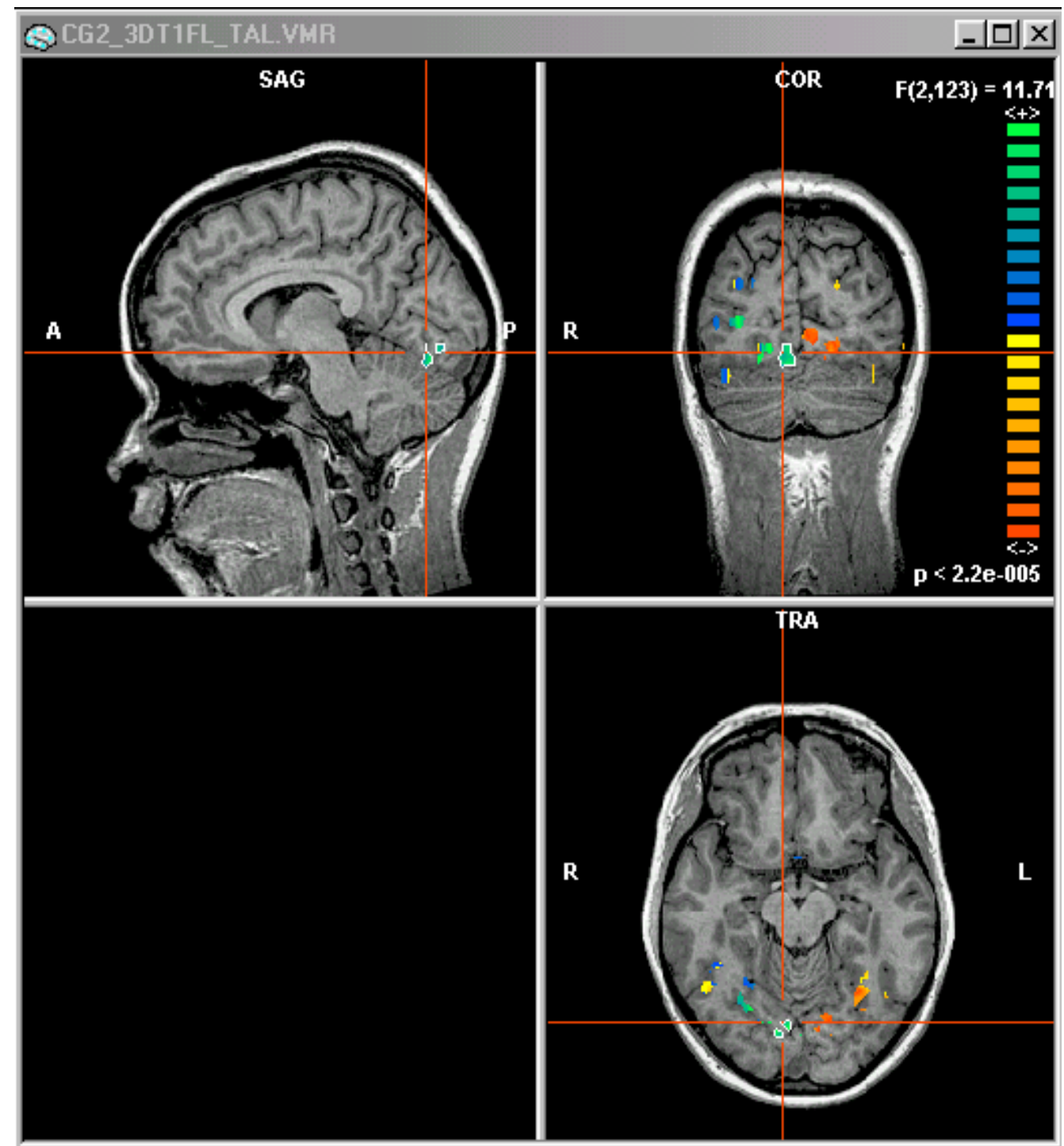
~1 mm resolution

~2 images per sec.

15,000 voxels/image

non-invasive, safe

measures Blood Oxygen  
Level Dependent (BOLD)  
response



[Mitchell et al.]



# Learned Naïve Bayes Models – Means for $P(\text{BrainActivity} \mid \text{WordCategory})$

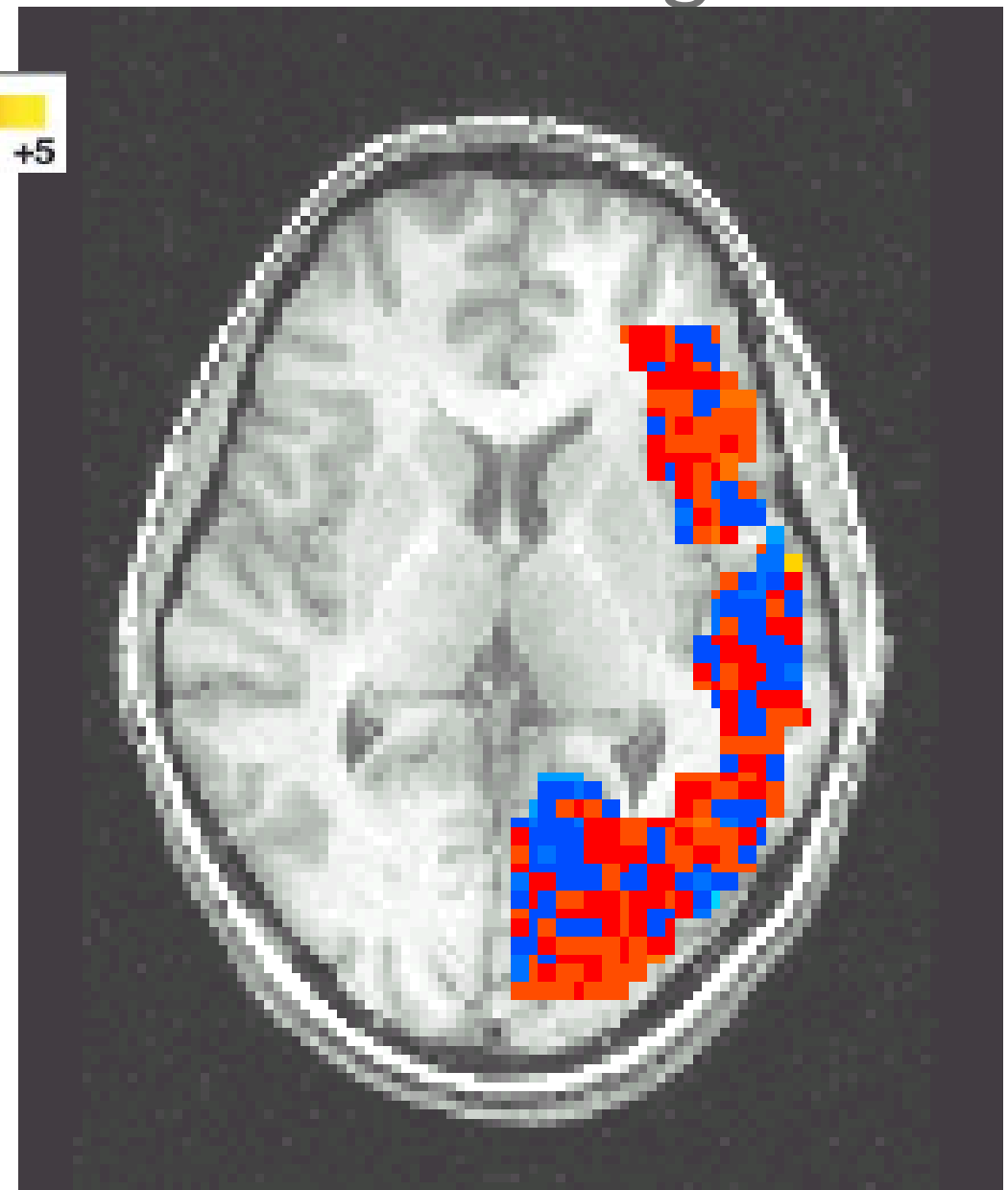
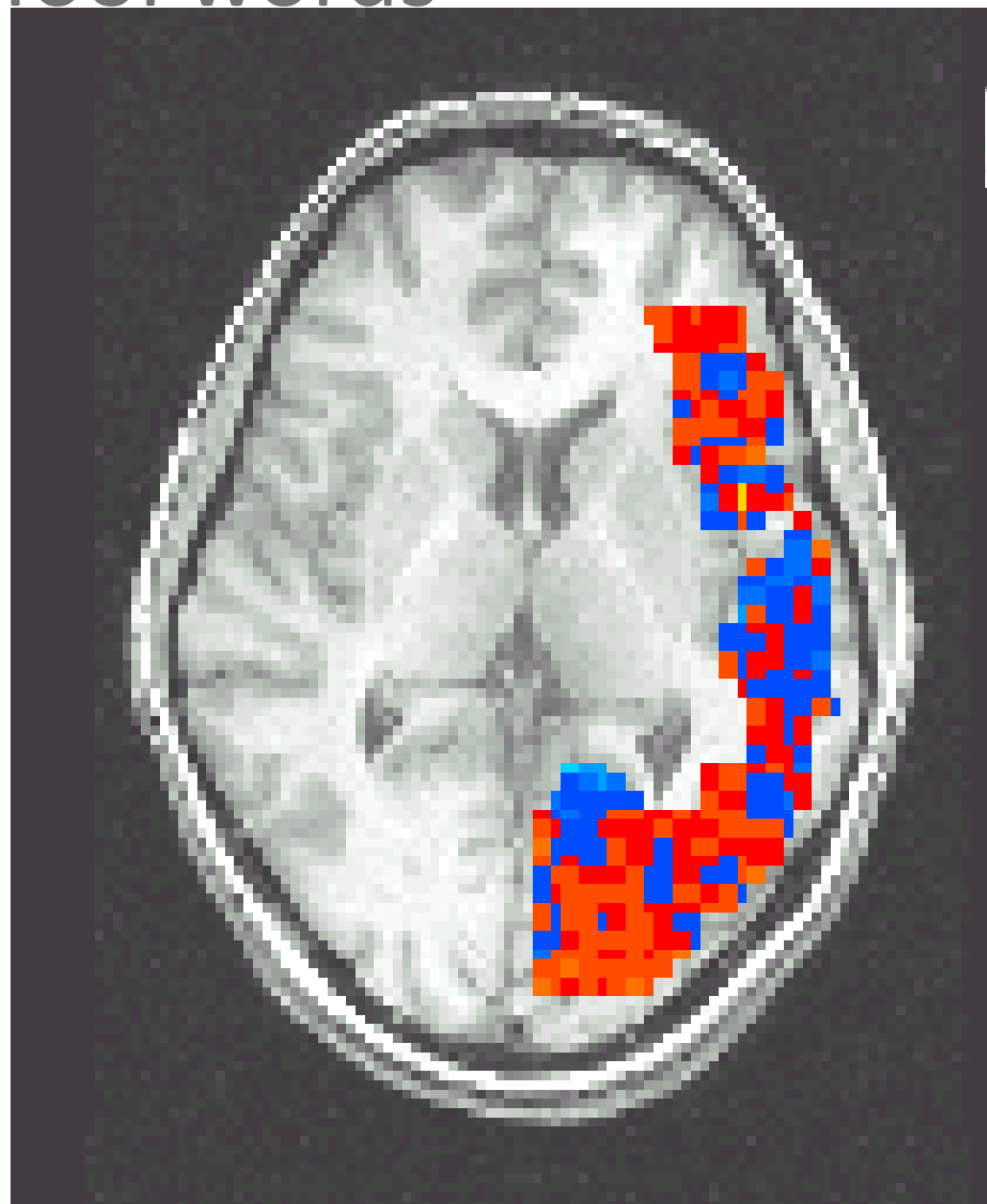
Pairwise classification accuracy:

[Mitchell et al.]

78-99%, 12 participants

Tool words

Building words





# What you should know...

## Naïve Bayes classifier

- What's the assumption
- Why we use it
- How do we learn it
- Why is Bayesian (MAP) estimation important

## Text classification

- Bag of words model

## Gaussian NB

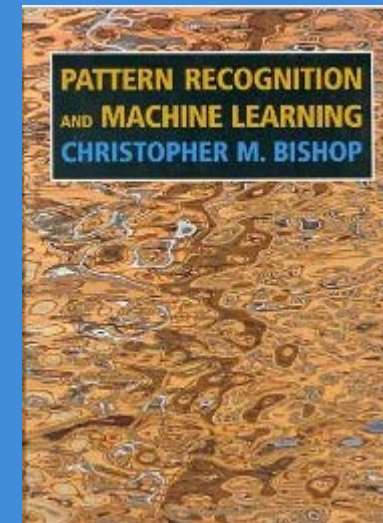
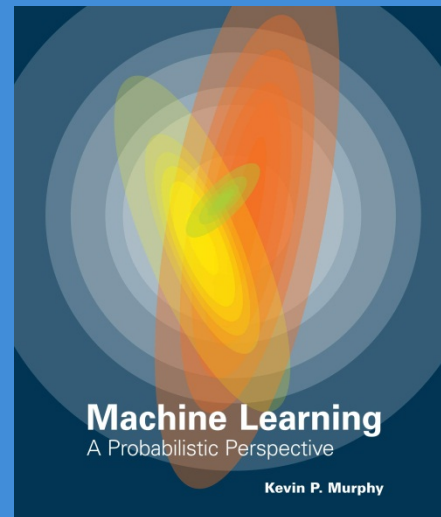
- Features are still conditionally independent
- Each feature has a Gaussian distribution given class

# Further reading

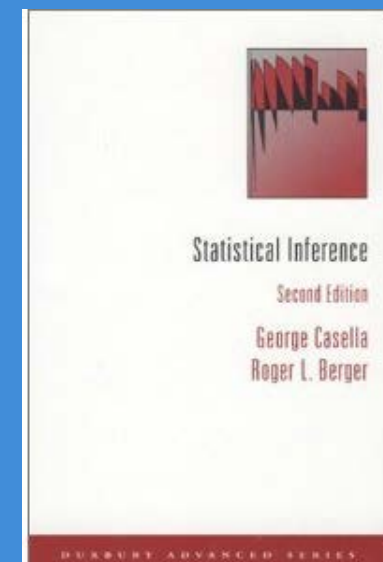
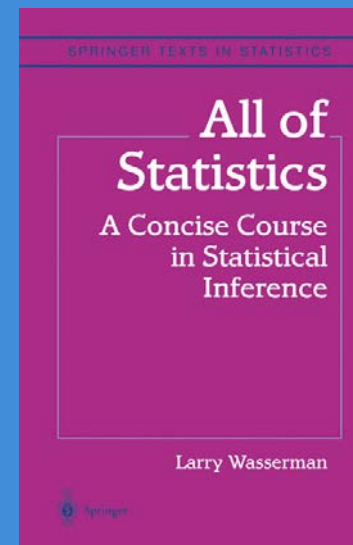
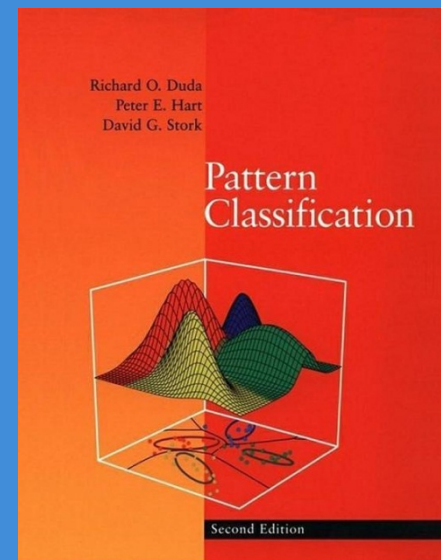
Manuscript (book chapters 1 and 2)

[http://alex.smola.org/teaching/berkeley2012/slides/chapter1\\_2.pdf](http://alex.smola.org/teaching/berkeley2012/slides/chapter1_2.pdf)

## ML Books



## Statistics 101



A tiny bit of extra theory...

# Feasible events = $\sigma$ -algebra

**Def:** A collection of subsets of  $\Omega$  is called a  $\sigma$ -algebra, denoted by  $\mathcal{M}$ , if it satisfies the following 3 properties:

- a.  $\emptyset \in \mathcal{M}$  (the empty set is an element of  $\mathcal{M}$ ).
- b. If  $A \in \mathcal{M}$ , then  $A^c \in \mathcal{M}$   
( $\mathcal{M}$  is closed under complementation).
- c. If  $A_1, A_2, \dots \in \mathcal{M}$ , then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{M}$   
( $\mathcal{M}$  is closed under countable unions).

## **Examples:**

- a. All subsets of  $\Omega = \{1, 2, 3\}$ :  $\{ \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\} \}$
- b.  $\Omega = (-\infty, \infty)$ .  $\mathcal{M} = \sigma((a, b) | a, b \in \mathbb{R})$  (Borel sets)

# Measure

Let  $\Omega$  be a set and  $\mathcal{M}$  a  $\sigma$ -algebra over  $\mathcal{M}$ .

A function  $\mu$  from  $\mathcal{M}$  to  $\mathbb{R} \cup \{\infty\}$  is called a measure if it satisfies the following properties.

(i) Nonnegativity.  $\mu(A) \geq 0$  for each  $A \in \mathcal{M}$ .

(ii)  $\exists E \in \mathcal{M}$  s.t.  $\mu(E) = 0$ , e.g.  $\mu(\emptyset) = 0$ .

(iii)  $\sigma$ -additivity: For disjoint sets  $A_i \in \mathcal{M}$ , we have

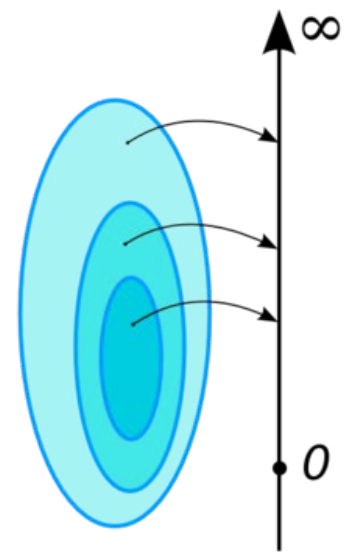
$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

## Consequences:

Monotonicity:  $A_1 \subset A_2$ ,  $A_1, A_2 \in \mathcal{M}$ , then  $\mu(A_1) \leq \mu(A_2)$ .

$$\mu(\text{large rectangle}) = \mu(\text{small rectangle}) + \mu(\text{medium rectangle}) + \mu(\text{tiny rectangle}) + \dots$$

$\sigma$ -additivity



monotonicity

# Important measures

**Counting measure:**  $\mu(A) = |A|$ , number of elements in the subset  $A$ .

**Borel measure:**  $(\mathbb{R}, \mathcal{B} = \sigma((a, b)), \mu)$

$\mu((a, b)) = |b - a|$ , length of the interval.

This is not a complete measure: There are Borel sets with zero measure, whose subsets are not Borel measurable...

**Lebesgue measure:**  $(\mathbb{R}, \mathcal{L} \supset \mathcal{B}, \lambda)$

complete extension of the Borel measure, i.e. extension & every subset of every null set is Lebesgue measurable (having measure zero).

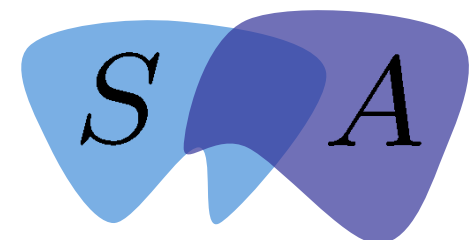
**Lebesgue measure construction:**

Given a subset  $A$ , its Lebesgue outer complete measure  $\lambda^*$  is defined as

$$\lambda^*(A) = \inf \{ \mu(A) \mid A \subset B \in \mathcal{B} \}$$

**Def:**  $A \subset \mathbb{R}$  is Lebesgue measurable if for every  $S \subset \mathbb{R}$

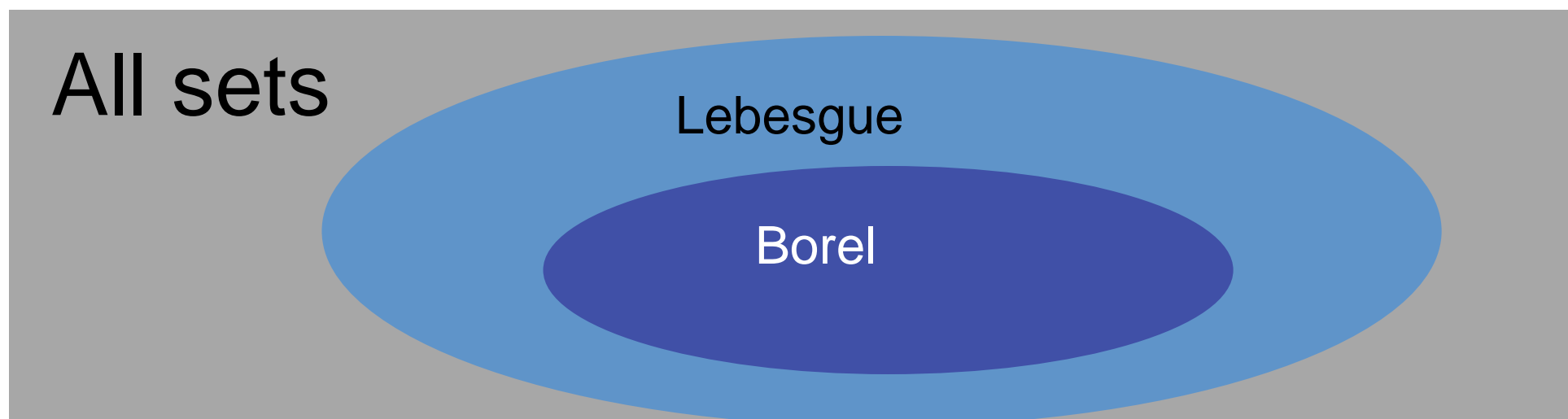
$$\lambda^*(S) = \lambda^*(S \cap A) + \lambda^*(S \setminus A)$$



# Brain Teasers ☺

*These might be surprising:*

- Construct an uncountable Lebesgue set with measure zero.
- Construct a Lebesgue but not Borel set.
- Prove that there are not Lebesgue measurable sets. We can't ask what is the probability of that event!
- Construct a Borel nullset who has a not measurable subset





# The Banach-Tarski paradox (1924)

Given a solid ball in 3-dimensional space, there exists a **decomposition** of the ball into a **finite** number of **non-overlapping** pieces (i.e., subsets), which can then be put back together in a different way to yield **two identical copies** of the original ball.

The reassembly process involves only **moving the pieces around and rotating them, without changing their shape**. However, the pieces themselves are not "solids" in the usual sense, but infinite scatterings of points.

A stronger form of the theorem implies that given any two "reasonable" solid objects (such as a small ball and a huge ball), either one can be reassembled into the other.

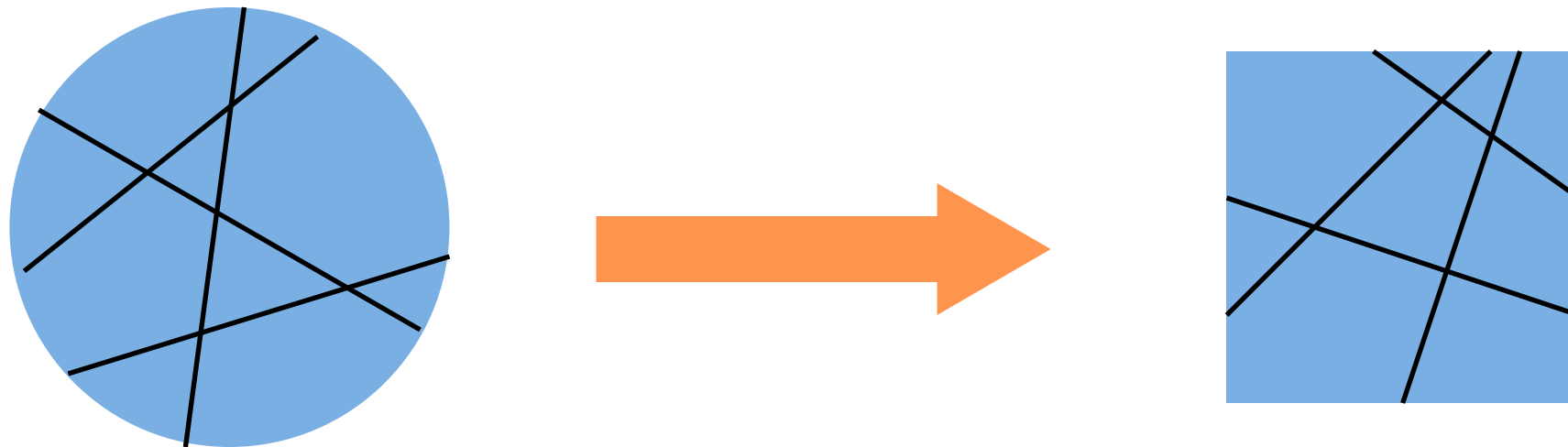
This is often stated colloquially as "a pea can be chopped up and reassembled into the Sun."





# Tarski's circle-squaring problem (1925)

Is it possible to take a disc in the plane, cut it into finitely many pieces, and reassemble the pieces so as to get a square of equal area?



**Miklós Laczkovich** (1990): It is possible using translations only; rotations are not required. It is not possible with scissors. The decomposition is non-constructive and uses about  $10^{50}$  different pieces.

Thanks for your attention 😊

# References

Many slides are recycled from

- Tom Mitchel

[http://www.cs.cmu.edu/~tom/10701\\_sp11/slides](http://www.cs.cmu.edu/~tom/10701_sp11/slides)

- Alex Smola

- Aarti Singh

- Eric Xing

- Xi Chen

- <http://www.math.ntu.edu.tw/~hchen/teaching/StatInference/notes/lecture2.pdf>

- Wikipedia