Artificial Intelligence and Machine Learning

Barbara Caputo

Useful info (1)

Teacher: Barbara Caputo,
https://scholar.google.it/citations
?hl=en&user=mHbdIAwAAAAJ&view_op=list_works

Assistant: Paolo Russo

Where/how to find us:
Email: barbara.caputo@iit.it
paolo.russo@iit.it

Office: 4E-45 (BC), Lab 7 –R3BC31 (PR)

Q/A time: after the lectures (BC)/by appointment
You can come any time at your own risk!

Useful info (1)

Textbooks/Useful books:

K. P. Murphy. Machine Learning: a probabilistic perspective. MIT Press.

B. Schoelkopf, A. Smola. Learning with Kernels. MIT Press.

W. Feller. An introduction to probability theory and its applications. Vol 1. Wiley Ed.

I. Goodfellows, Y. Bengio, A. Courville. Deep Learning. MIT Press.

Useful info (2)

Exam modality:

Homeworks (lab experiences) +written exam +oral exam

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svm, cnn???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svn, cnn???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

Useful info (2)

Exam modality:

-Homeworks during the lectures: 3, on three key topics (PCA, svn, cnn???). Reports to be submitted by end of the lectures (19/01/2019, 23:59 CET). Every experience graded (2 points), if overall points <4 no admission to written exam.

- Written exam: exercises as done during lectures. If grade <18, no admission to oral exam.

- Oral exam: three questions, pen and paper (demonstrations).

Questions?

Why to do this exam?

A basic set of tools that you will need for:

- **your M. Sc./PhD Thesis**
- **your Start Up**
- **Wall Street**
- **Google, Facebook, Amazon, Bosch, Huawei, Samsung...**

Artificial Intelligence and Machine Learning

AI: theory and algorithms that enable computers to mimic human intelligence

ML: a subset of AI that includes statistical techniques enabling machines to improve at tasks with experience.

Artificial Intelligence and Machine Learning

AI: theory and algorithms that enable computers to mimic human intelligence

ML: a subset of AI that includes statistical techniques enabling machines to improve at tasks with experience.

This course focuses on ML only, and specifically on a specific family of ML algorithms (discriminative methods)

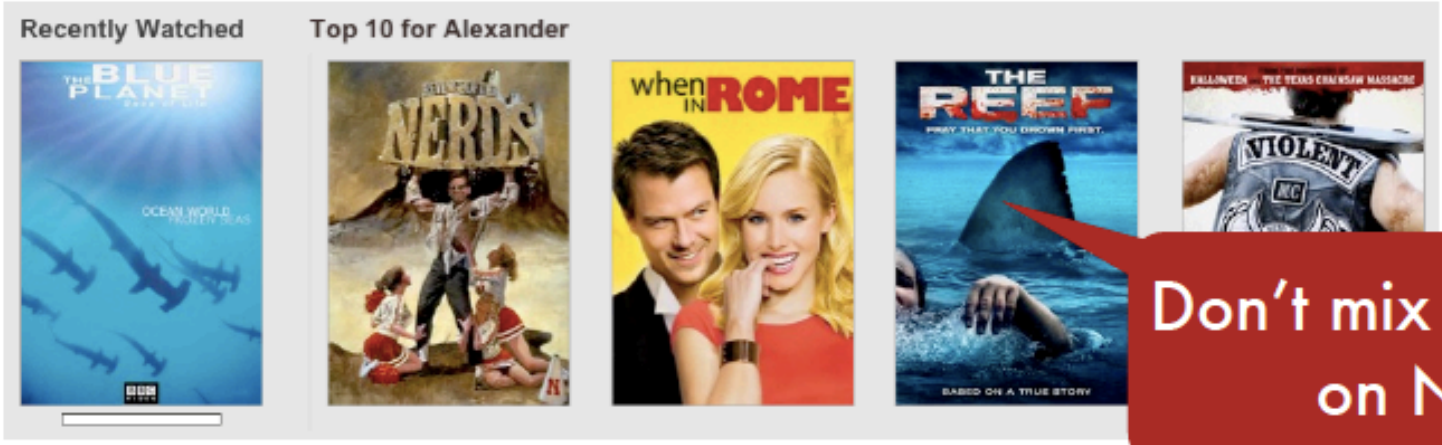
Machine Learning

- Programming computers to use example data or past experience

Machine Learning

- Programming computers to use example data or past experience
- Well-Posed Learning Problems
  - A computer program is said to learn from *experience E*
  - with respect to *class of tasks T* and *performance measure P*,
  - if its performance at tasks *T*, as measured by *P*, improves with experience *E*.

Collaborative Filtering



Don't mix preferences on Netflix!

Customers Who Bought This Item Also Bought

Convex Optimization by Stephen Boyd
(11)
$65.78

Point Processes (Chapman & Hall / CRC Monographs on S... by D.R. Cox
$125.47

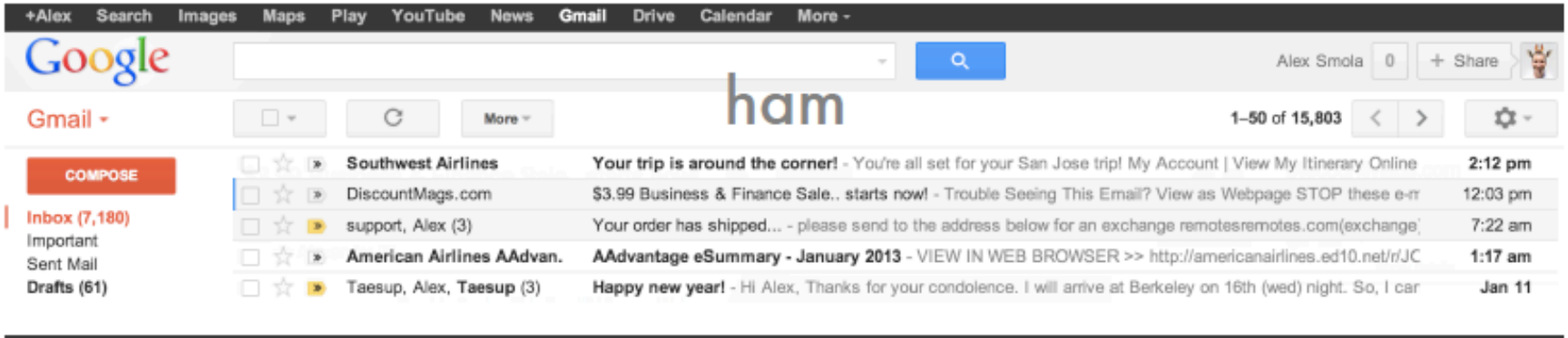
Probabilistic Graphical Models: Principles and T... by Daphne Koller
(5)
$71.52

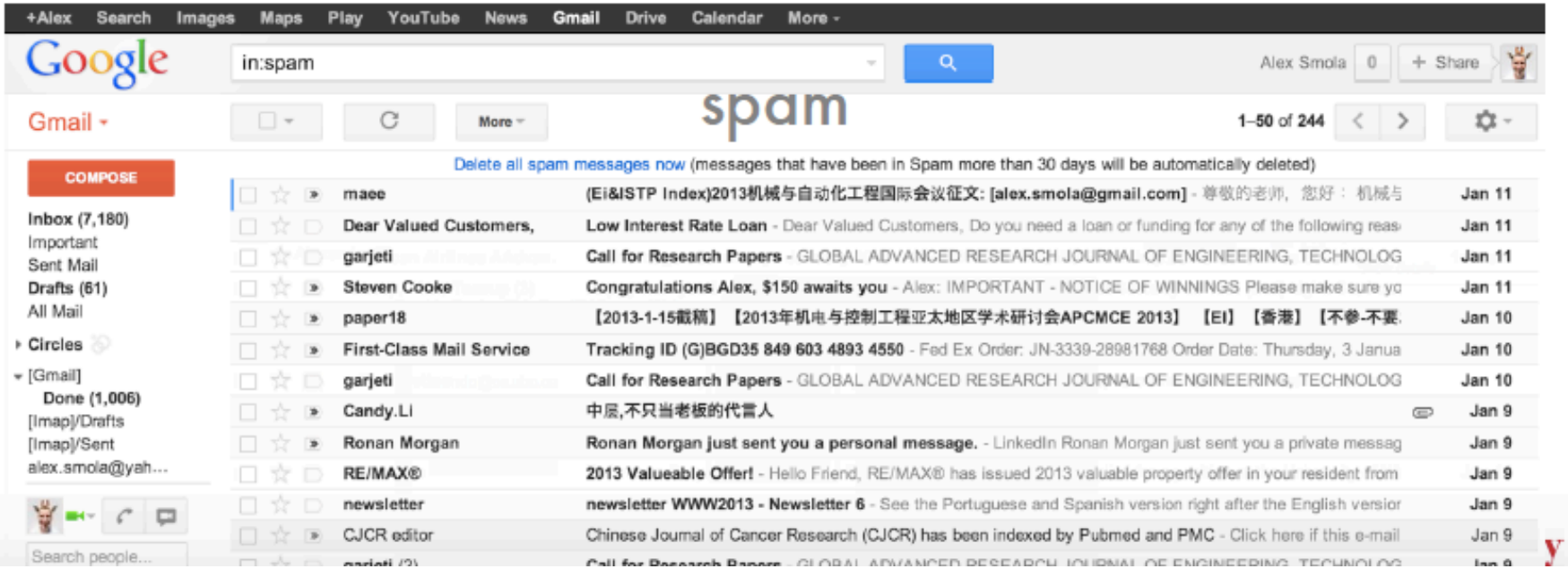
Amazon books

Imitation Learning in Games

Black & White
Lionsgate Studios

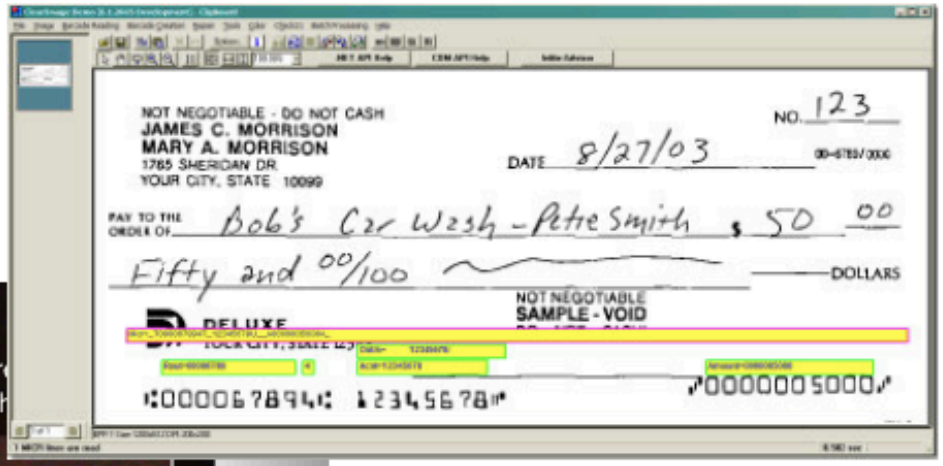
Spam Filtering

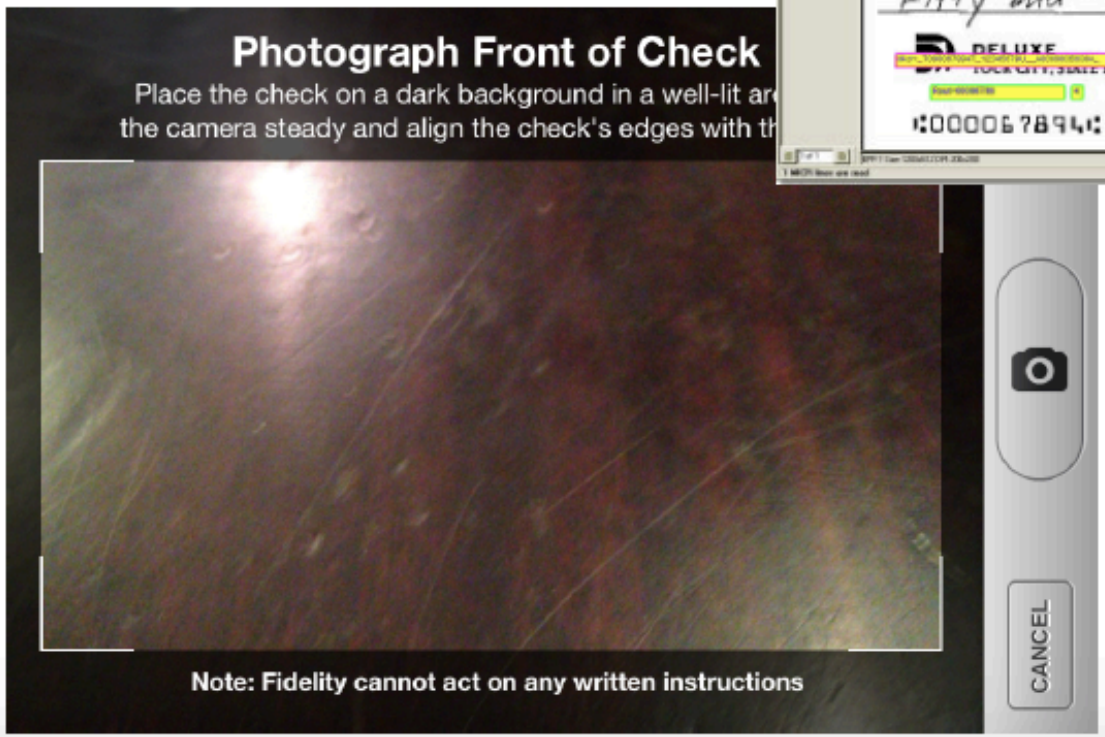




Cheque reading

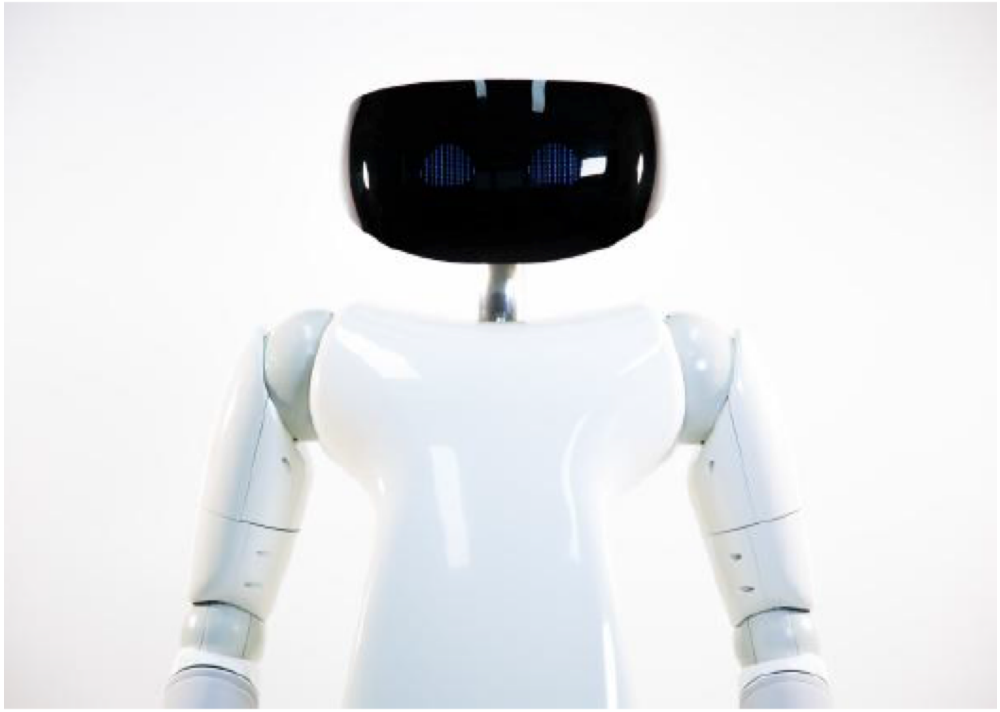
segment image





recognize handwriting

Autonomous systems –here we need more than ML!



Programming with Data

- Want adaptive robust and fault tolerant systems
- Rule-based implementation is (often)
  - difficult (for the programmer)
  - brittle (can miss many edge-cases)
  - becomes a nightmare to maintain explicitly
  - often doesn't work too well (e.g. OCR)

Programming with Data

- Want adaptive robust and fault tolerant systems
- Rule-based implementation is (often)
  - difficult (for the programmer)
  - brittle (can miss many edge-cases)
  - becomes a nightmare to maintain explicitly
  - often doesn't work too well (e.g. OCR)
- Usually easy to obtain examples of what we want
  IF x THEN DO y
- Collect many pairs (x_i, y_i)
- Estimate function f such that $f(x_i) = y_i$ (supervised learning)
- Detect patterns in data (unsupervised learning)

Problem Prototypes

Supervised Learning $y = f(x)$

Supervised Learning $y = f(x)$

- Binary classification
  Given x find y in {-1, 1}

Supervised Learning $y = f(x)$

- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, ... k}

often with loss

$$l(y, f(x))$$

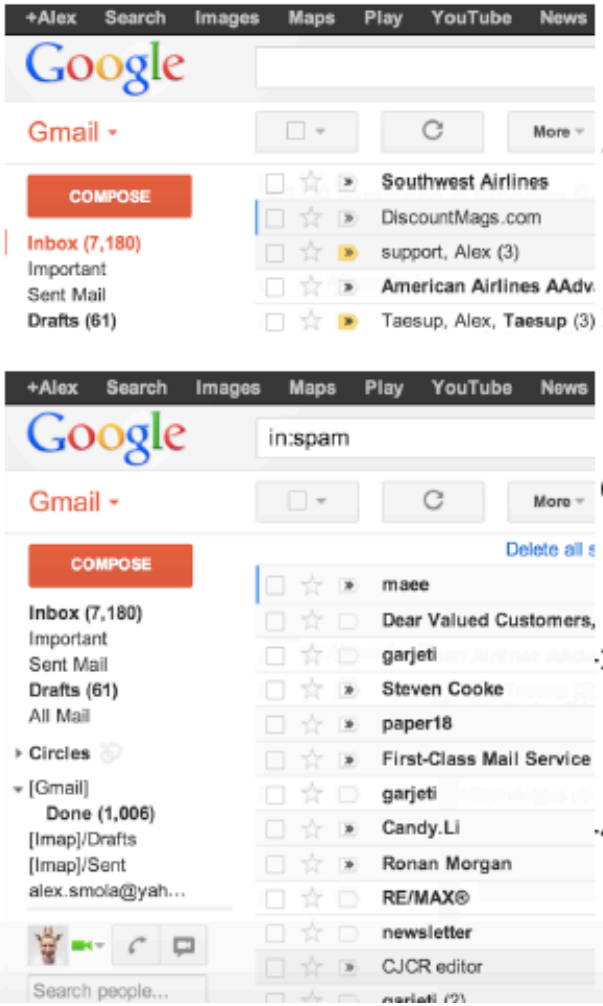
Supervised Learning $y = f(x)$

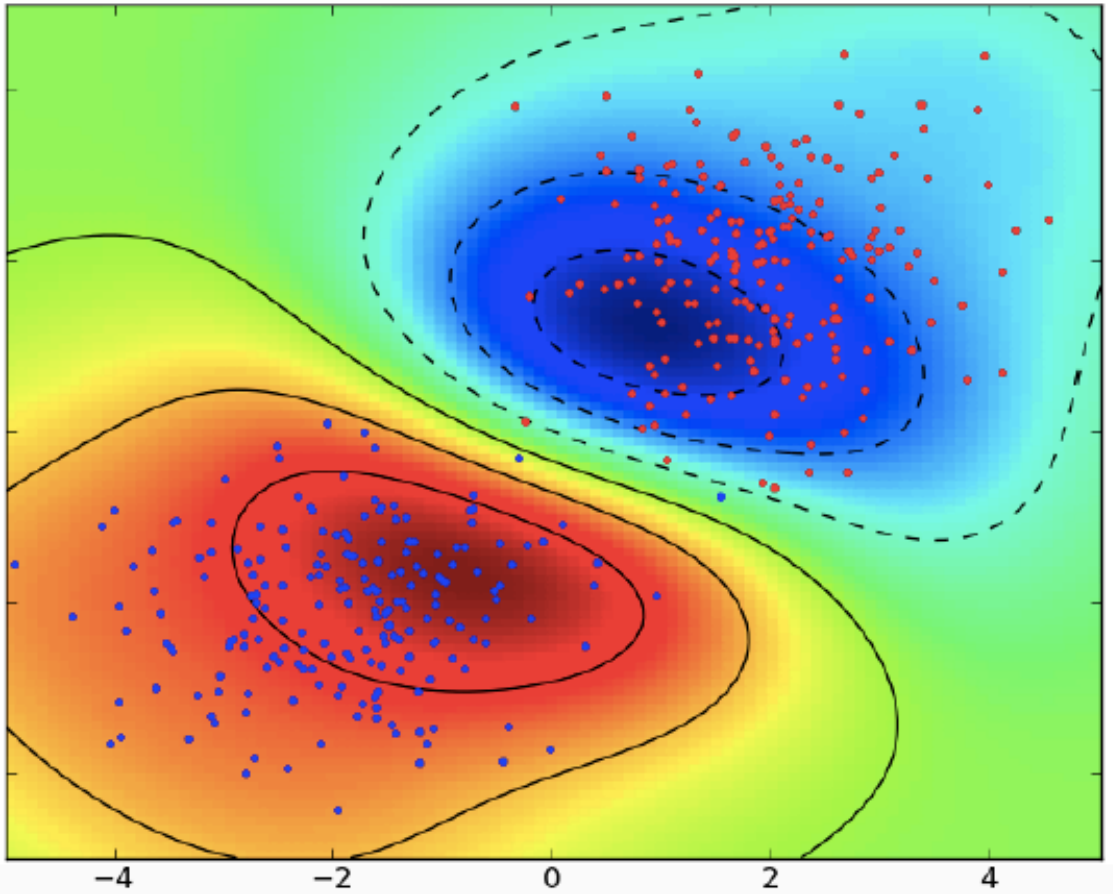
- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, ... k}
- Regression
  Given x find y in R (or R^d)

often with loss

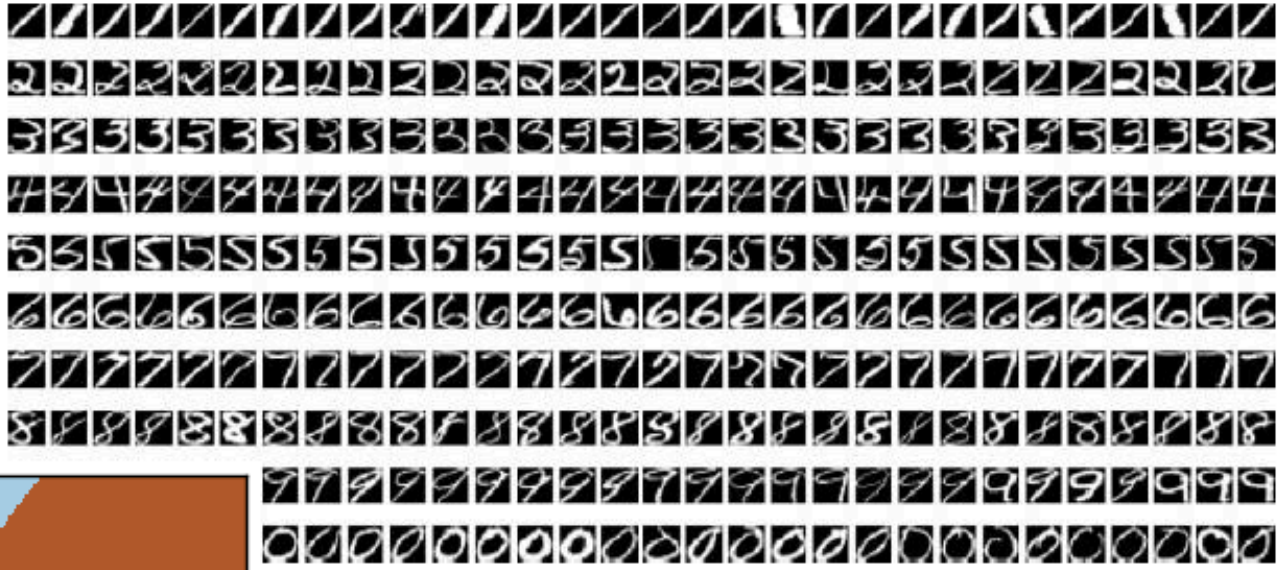
$$l(y, f(x))$$

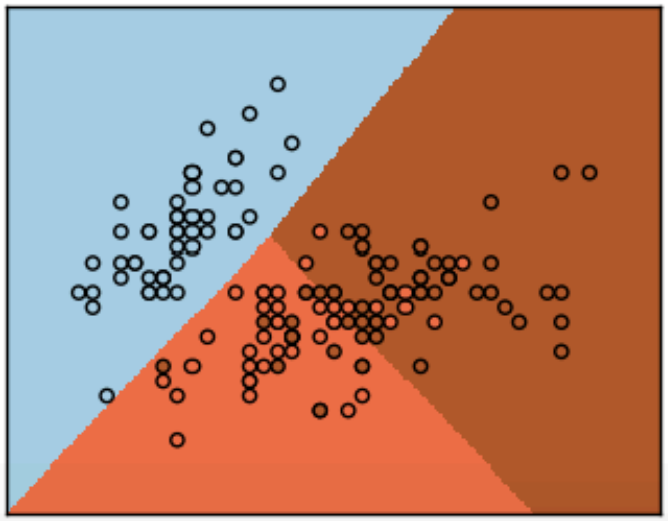
Binary Classification





Multiclass Classification





map image x to digit y

Unsupervised Learning

Unsupervised Learning

- Given data x, ask a good question ... about x or about model for x

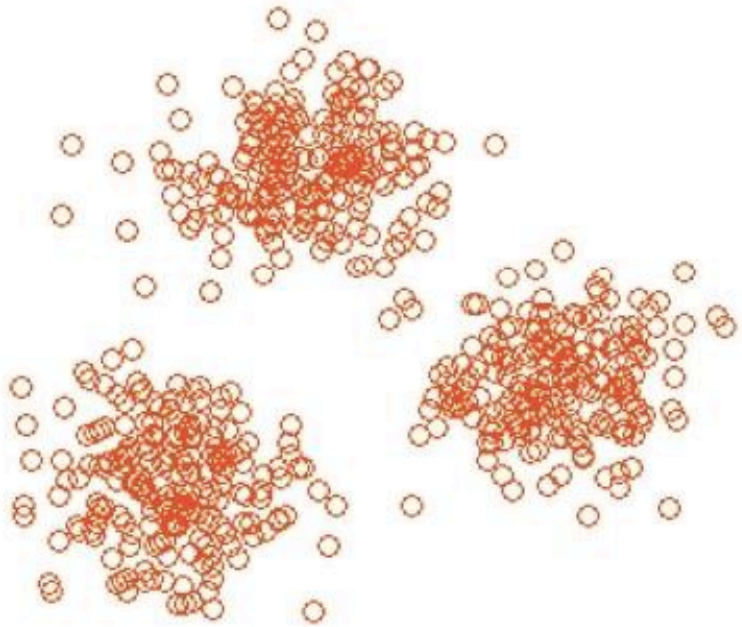
Unsupervised Learning

- Given data x, ask a good question … about x or about model for x
- Clustering
  Find a set of prototypes representing the data

Unsupervised Learning

- Given data x, ask a good question … about x or about model for x
- Clustering
  Find a set of prototypes representing the data
- Principal Components
  Find a subspace representing the data

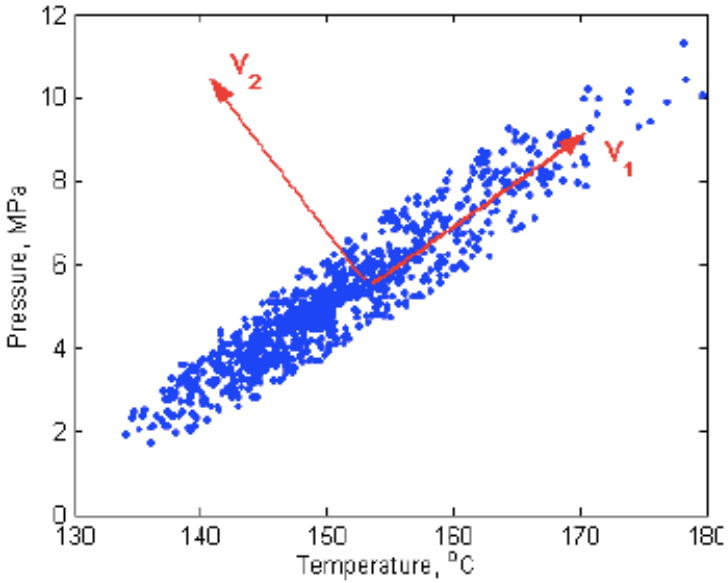
Clustering

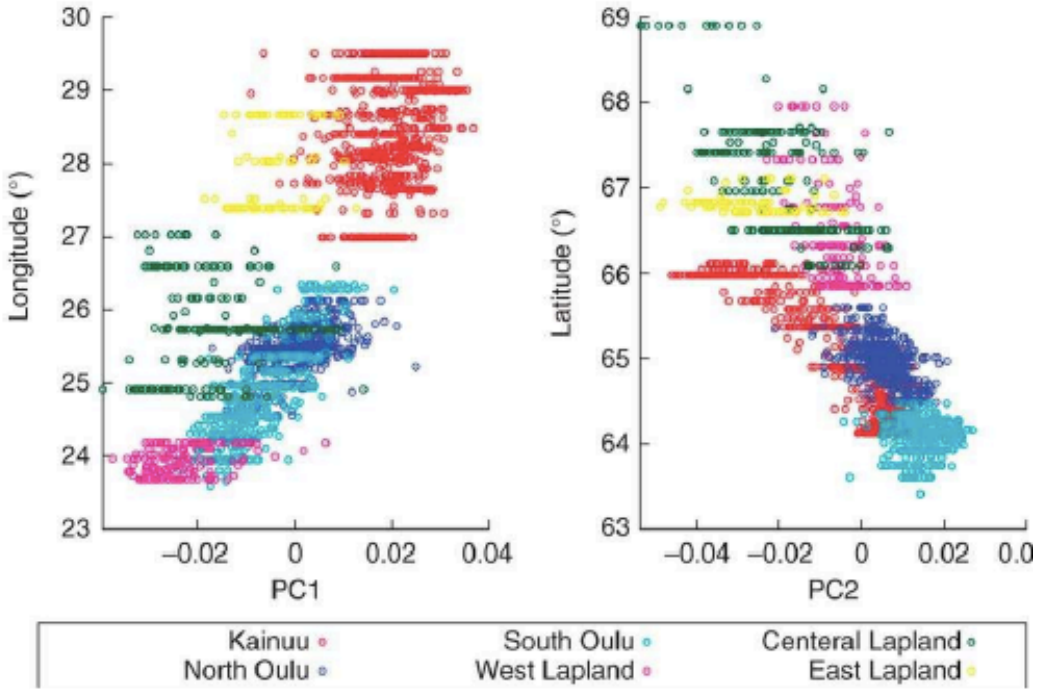


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

...

Principal Components





Variance component model to account for sample structure in genome-wide association studies, **Nature Genetics 2010**

Discriminative vs. Generative (mainly relevant for supervised models)

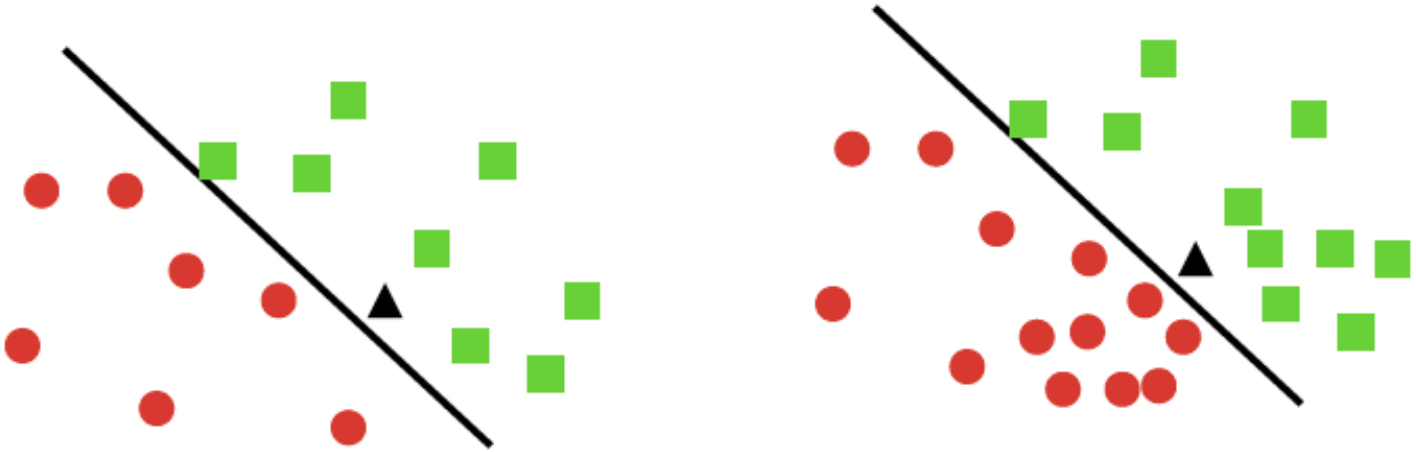
Discriminative vs. Generative (mainly relevant for supervised models)

- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions

Discriminative vs. Generative (mainly relevant for supervised models)

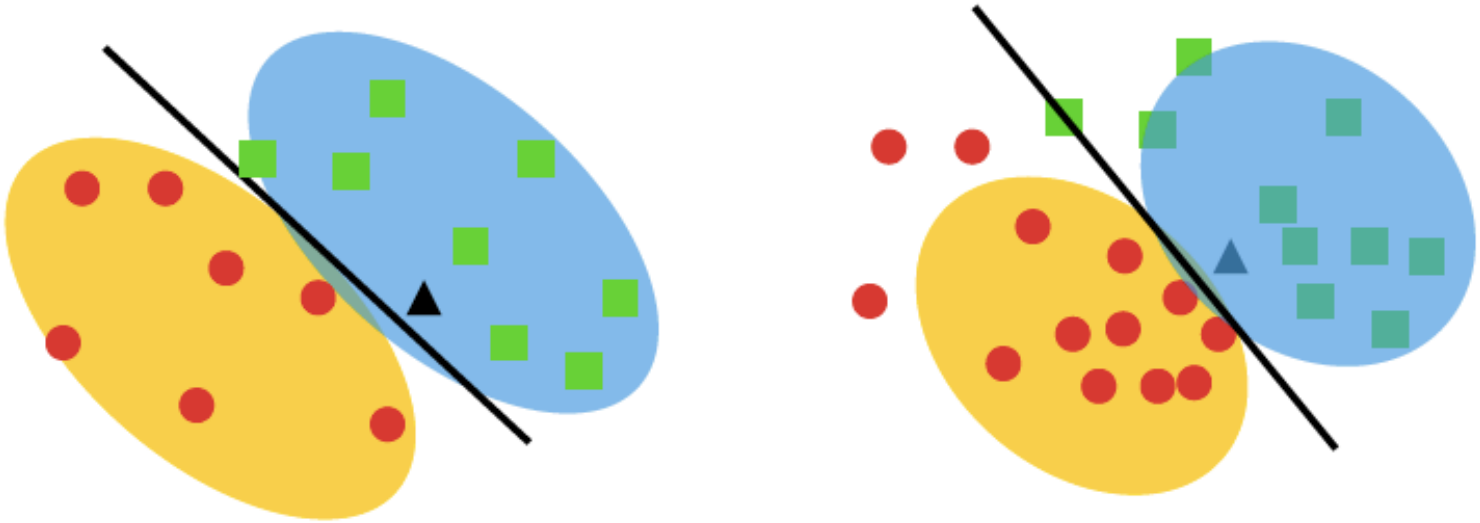
- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions
- Generative models
  - Estimate joint distribution over (x,y)
  - Use conditional probability to infer y|x
  - Often more intuitive
  - Easier to add prior knowledge

Discriminative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative



- Model observations (x,y) first
- Then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

Let’s take a break!

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)

What is the probability?

Probabilities

Bayes

Kolmogorov

Sample space

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

Sample space

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

Examples:

– Ω 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?

Events

We will ask the question:
What is the probability of a particular event?

Def: *Event* *A is a* *subset* *of the sample space* Ω

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*

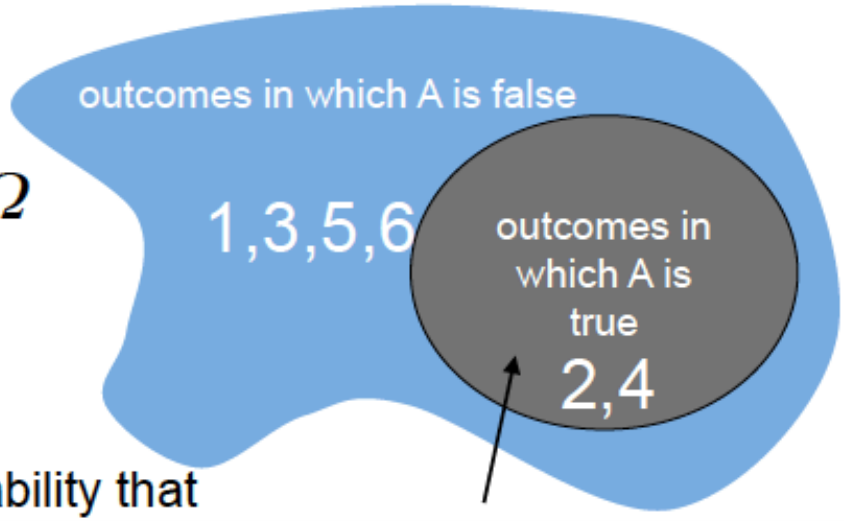
Probability

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*.

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 Ω



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

Kolmogorov Axioms

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

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

(iii) σ-additivity: For disjoint sets (events) A_i, we have

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

Kolmogorov Axioms

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

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

(iii) σ-additivity: For disjoint sets (events) A_i, we have

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

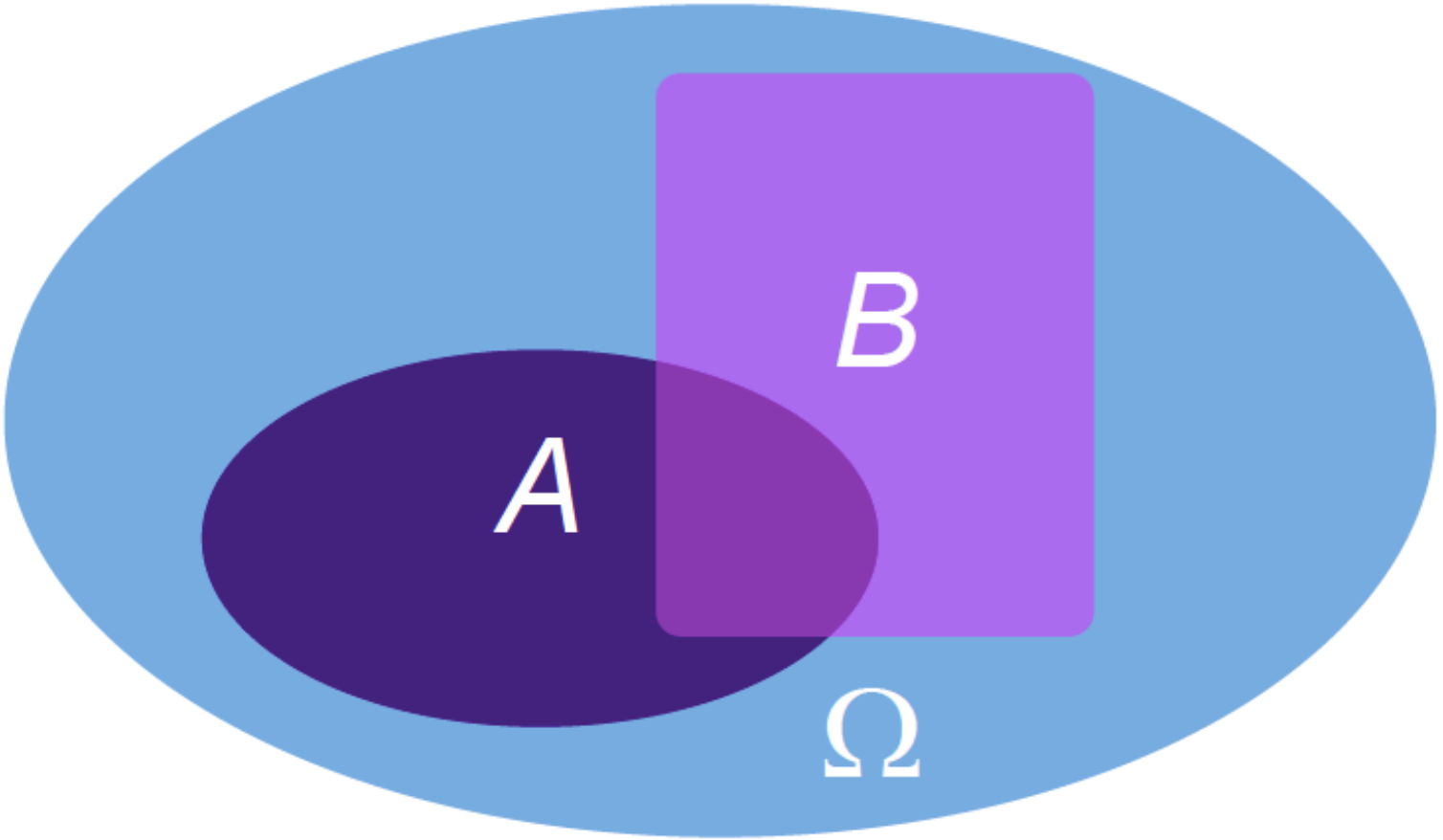
Consequences:

$P(\emptyset) = 0$.

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

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

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

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

Random Variables

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

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

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

Random Variables

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

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

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

Examples:

- **Discrete random variable examples (Ω is discrete):**
- X(ω) = True if a randomly drawn person (ω) from our class (Ω) is female
- X(ω) = The hometown X(ω) of a randomly drawn person (ω) from our class (Ω)

Random Variables

Sometimes Ω can be quite abstract

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

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

Random Variables

Sometimes Ω can be quite abstract

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

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

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 Ω

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

What discrete distributions do we know?

Discrete Distributions

- Bernoulli distribution: Ber(p)

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

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

Discrete Distributions

- Bernoulli distribution: Ber(p)

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

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

- Binomial distribution: 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?

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

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

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

Continuous Distribution

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

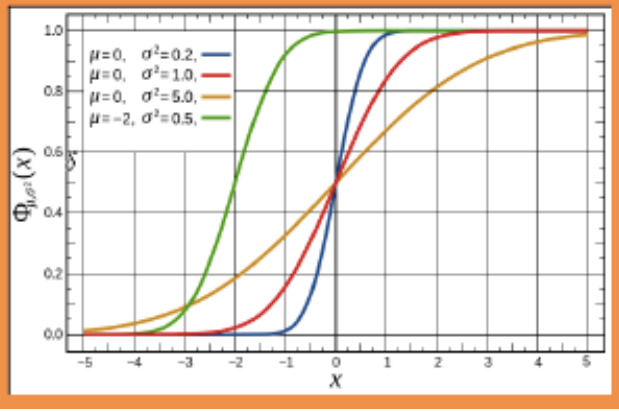
Continuous Distribution

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

Def: cumulative distribution function

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

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



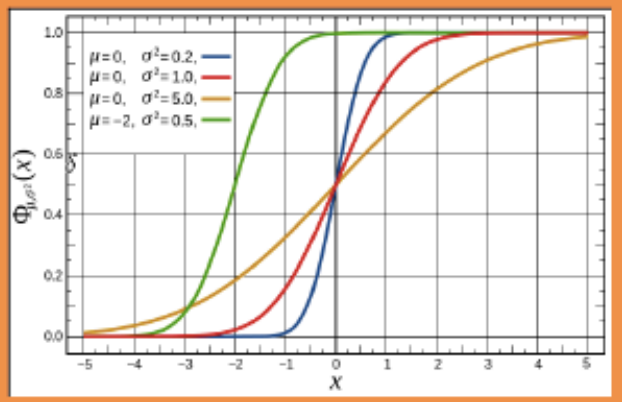
Continuous Distribution

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

Def: cumulative distribution function

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

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

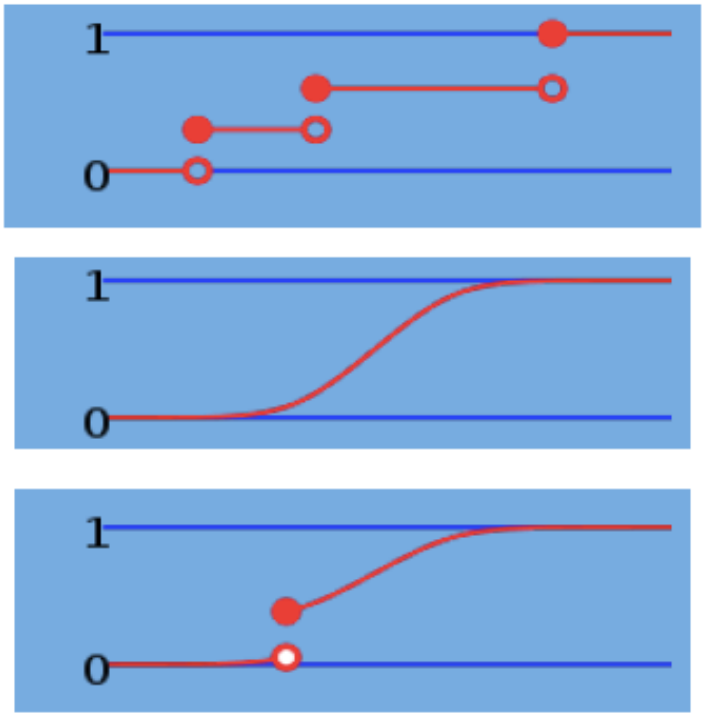


Def : Let $F(-\infty) = 0$. $F : (-\infty, \infty) \to \mathbb{R}$ is **absolutely continuous** $F(x) = \int_{-\infty}^{x} f(t)dt$ for some function f.

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.

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

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???

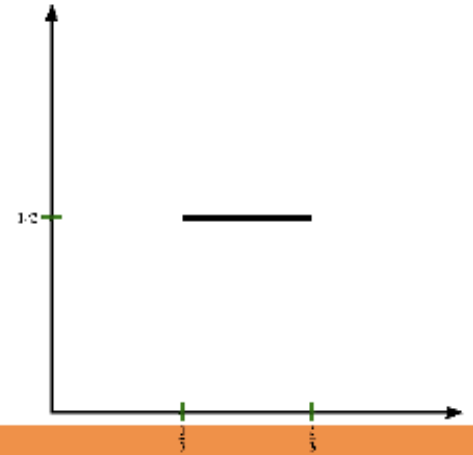
Cumulative Distribution Function (cdf)

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. ⇒ there is **no density** for the Cantor function CDF.

Probability Density Function (pdf)

Probability Density Function (pdf)

Pdf properties:

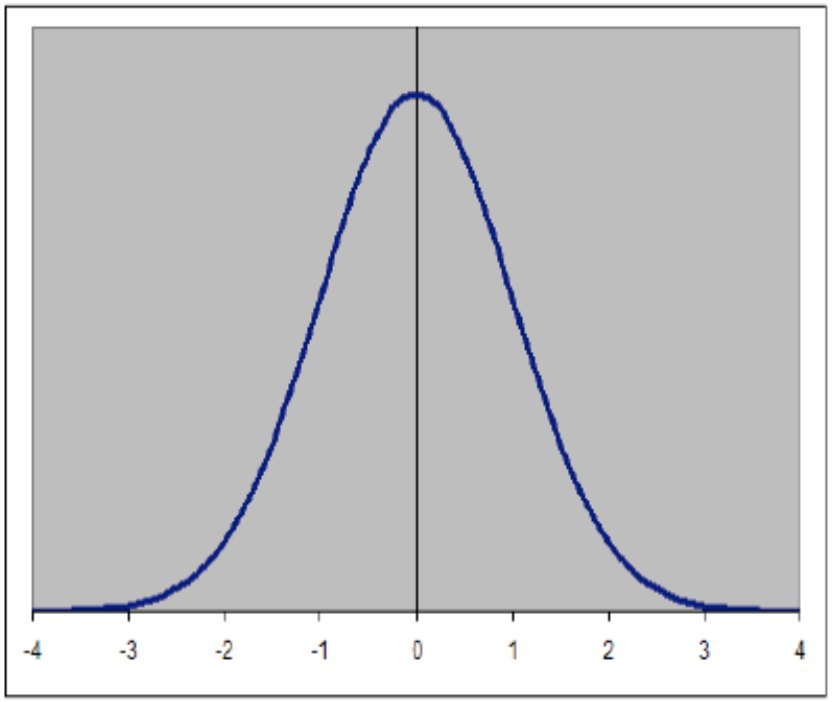
$$f(x) \geq 0$$

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

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

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

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$



Probability Density Function (pdf)

Pdf properties:

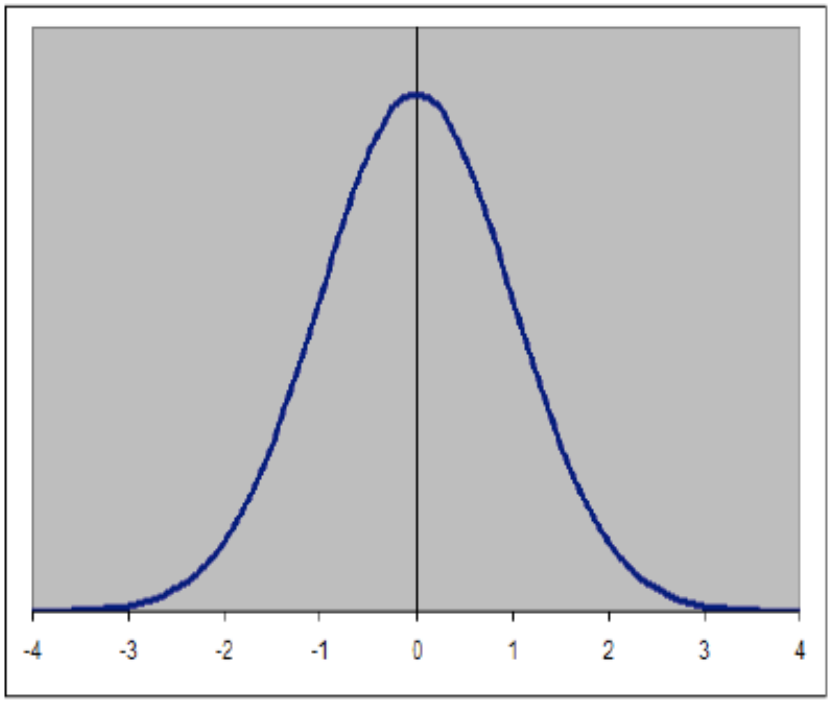
$$f(x) \geq 0$$

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

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

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

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

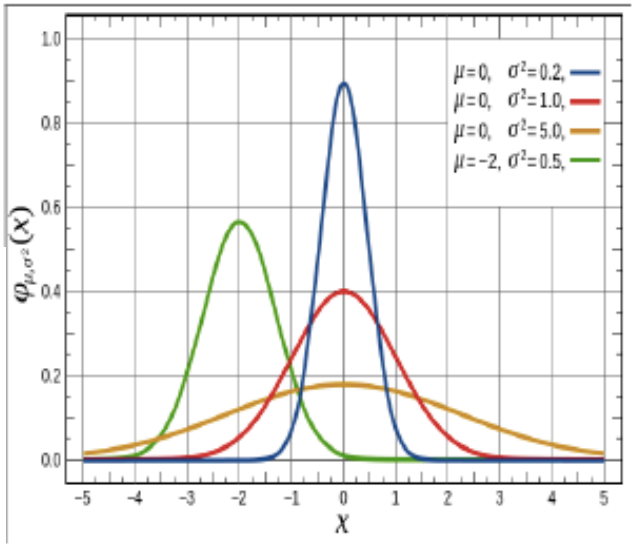


Intuitively, one can think of f(x)dx as being the probability of X falling within the infinitesimal interval [x, x + dx]. $P(x < X < x + dx) = f(x)dx$

Moments

Expectation: average value, mean, 1st moment:

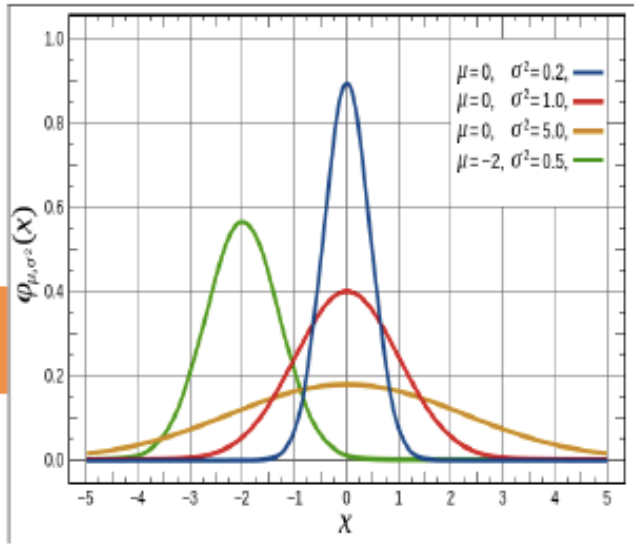
$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$



Moments

Expectation: average value, mean, 1st moment:

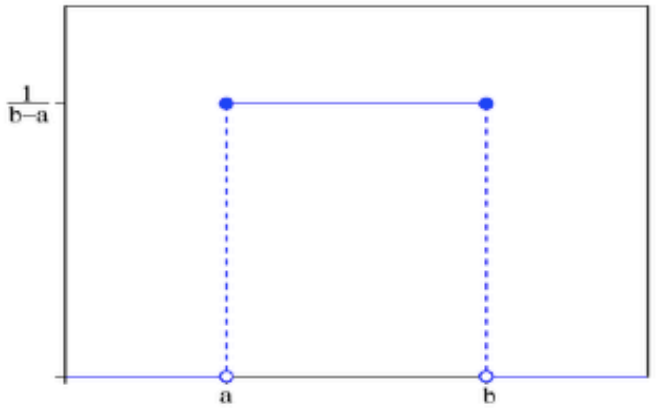
$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$



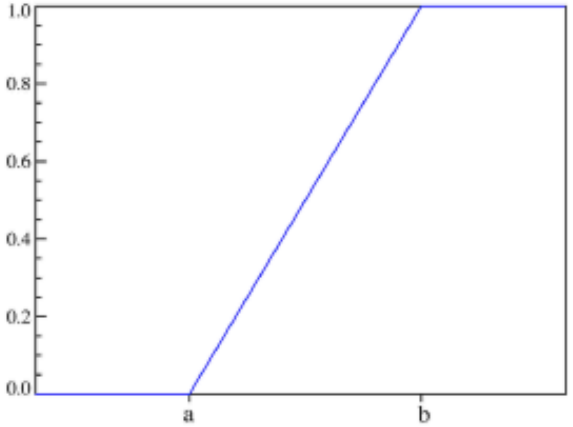
Variance: the spread, 2nd moment:

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

Uniform Distribution



PDF

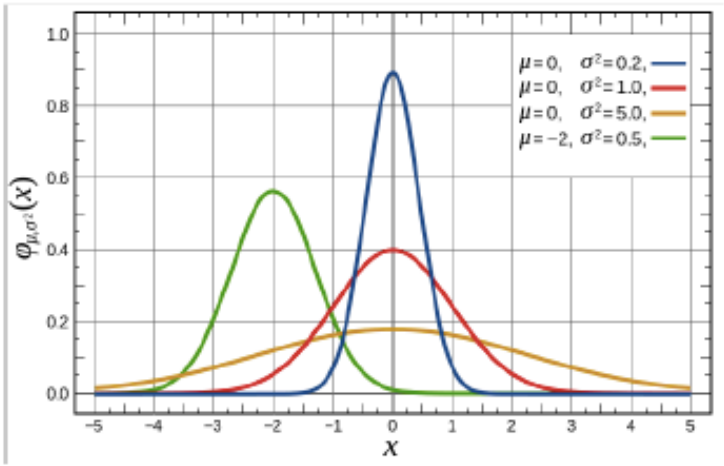


CDF

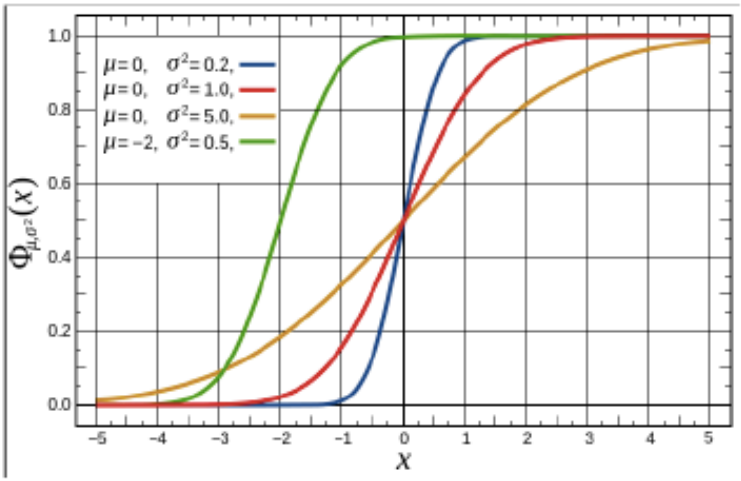
$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{Otherwise} \end{cases}$$

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

Normal (Gaussian) Distribution



PDF



CDF

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

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

That’s all!