Machine Learning

Prof. Barbara Caputo
Dip. Ingegneria Informatica, Automatica e Gestionale, Roma

SAPIENZA
UNIVERSITÀ DI ROMA

Useful info (1)

Teacher: Barbara Caputo,
www.dis.uniroma1.it/~barbaracaputo

Assistant: Fabio M. Carlucci,

Where/how to find us:
Email: caputo@dis.uniroma1.it,
fabiom.carlucci@dis.uniroma1.it

Office: B109 (BC), A003 (IK)

Q/A time: Fri 9:30-10:30/after the lectures (BC)
You can come at other times at your own risk!

Useful info (2)

Exam modality:

Homeworks (lab experiences) +oral

For AIML students, the modality will remain as last year (homeworks +written exam)

Course Web site:
https://sites.google.com/site/machinelearning20162017/

To get slides (after lecture) send and email to F. M. Carlucci for access to course Dropbox

Outline

Introduction and basics (a crash course on probability)
Bayes decision theory
Principal Component Analysis
Regression
Non-parametric methods: K-NN
Discriminative Methods: Perceptron, Neural Networks,
Deep learning,
SVM, Kernels
Learning Theory: regularization, risk minimization,
VC dimension
Unsupervised Learning: clustering,
semi-supervised learning
Probabilistic Representation and modeling:
graphical models, Bayes nets, HMM, Reinforcement learning,
topic models.

Outline for AIML students

Introduction and basics (a crash course on probability)
Bayes decision theory
Principal Component Analysis
Non-parametric methods: K-NN
Discriminative Methods: Perceptron, SVM, Kernels
Learning Theory: regularization, risk minimization,
Unsupervised Learning: clustering,

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, Yahoo!...**

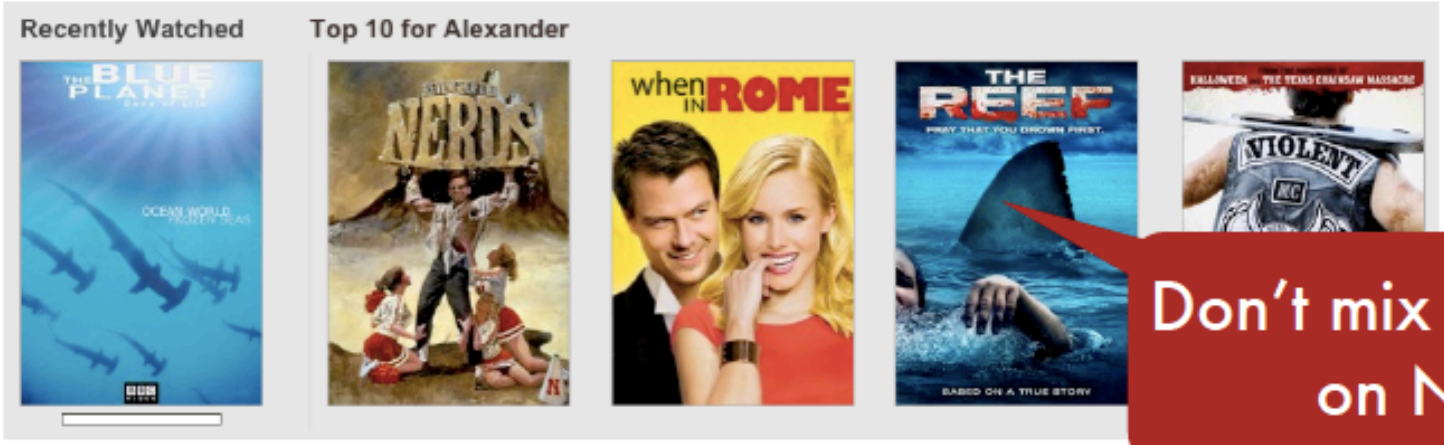
Machine Learning

- 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

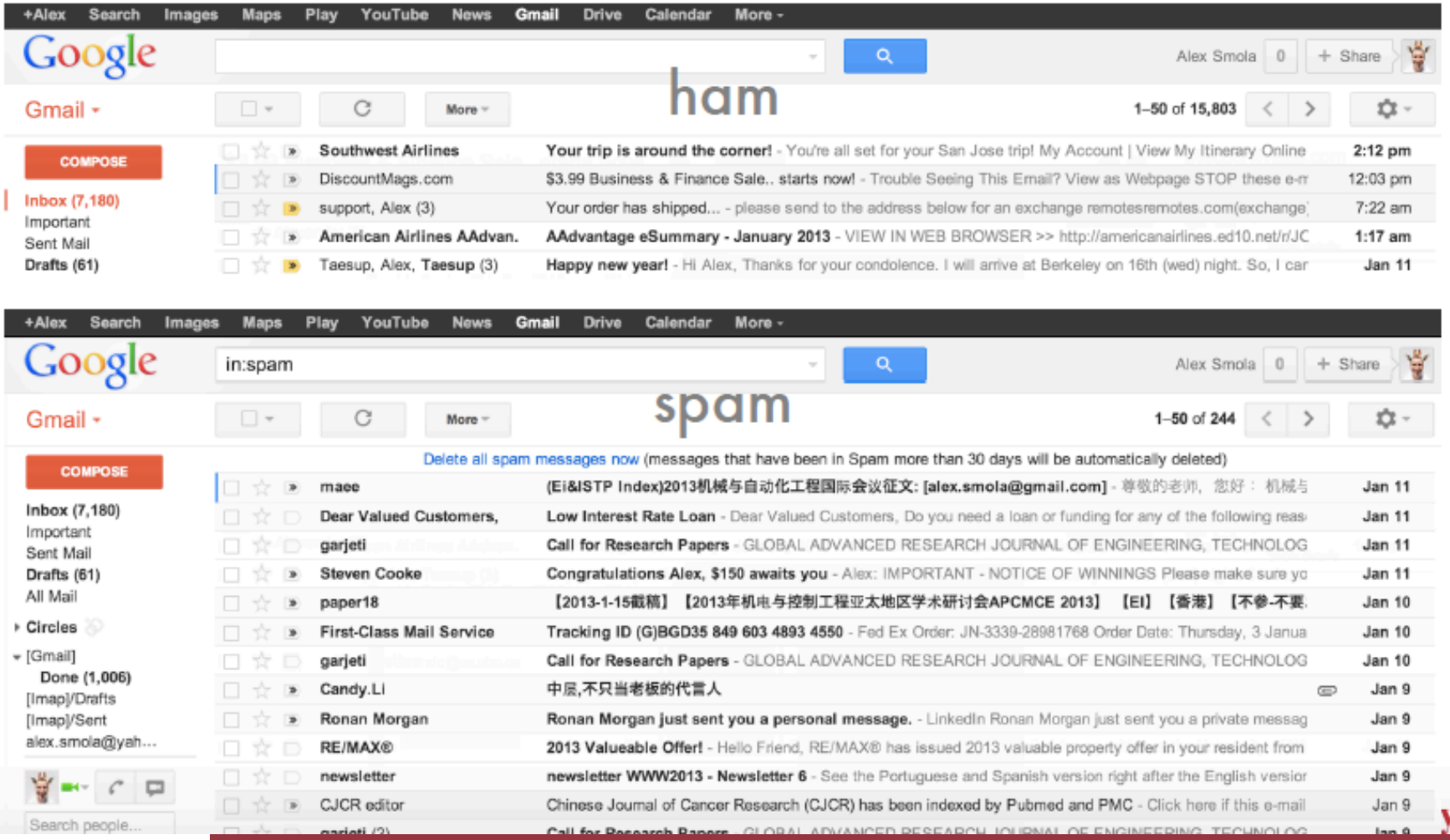
Probabilistic Graphical Models: Principles and T... by Daphne Koller
(5)
$71.52

Amazon books

Imitation Learning in Games

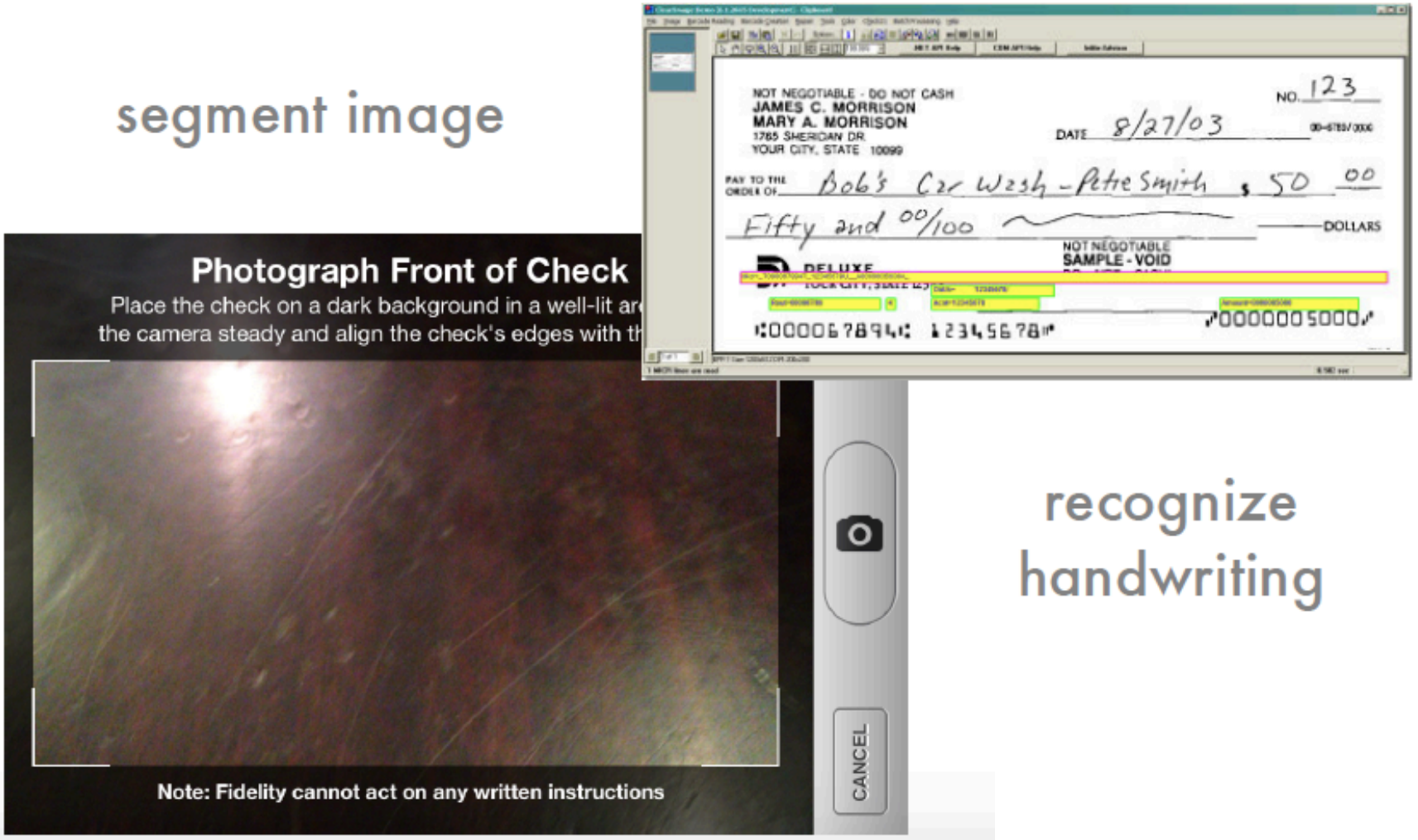
Black & White
Lionsgate Studios

Spam Filtering



Cheque reading

segment image



recognize handwriting

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

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

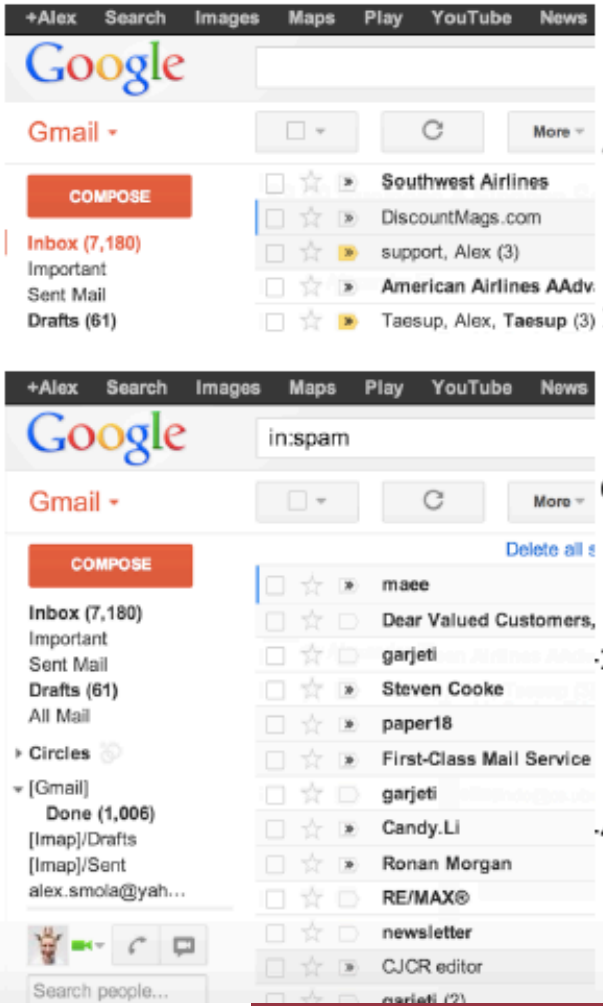
Supervised Learning $y = f(x)$

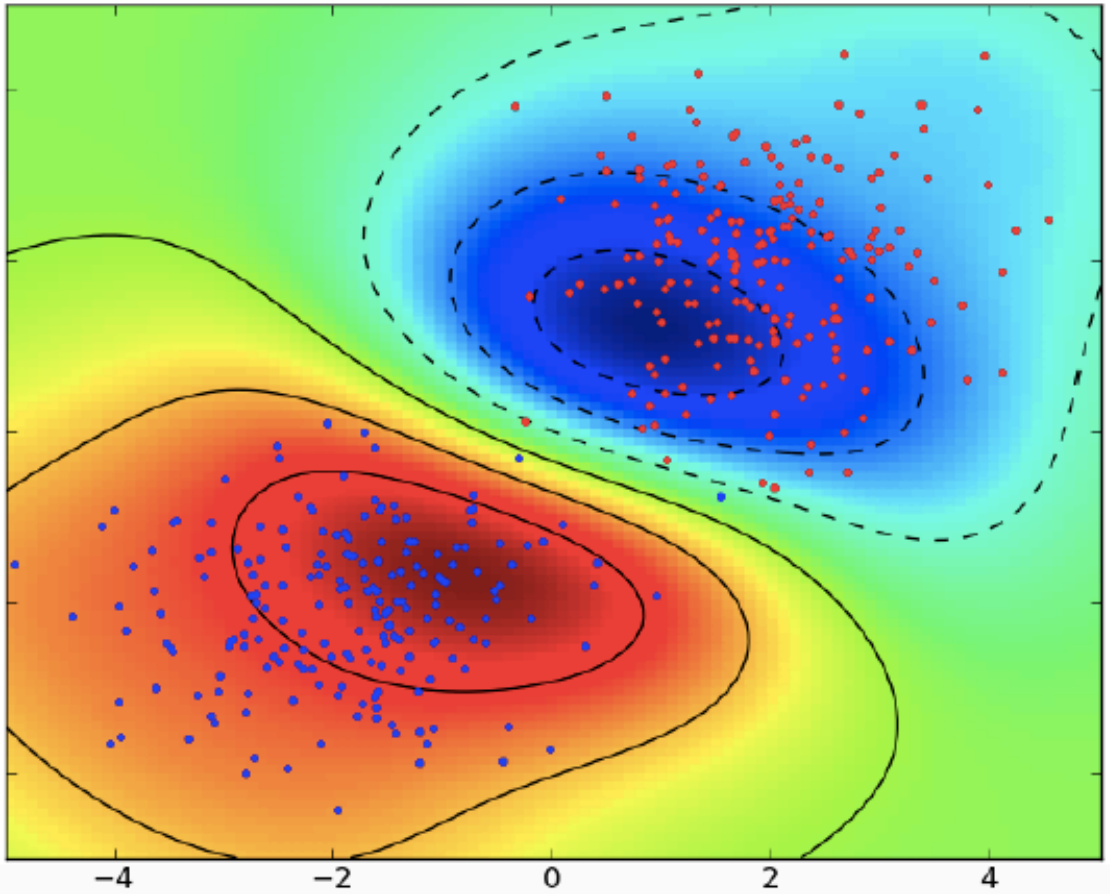
- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, … k}
- Regression
  Given x find y in R (or R^d)

often with loss

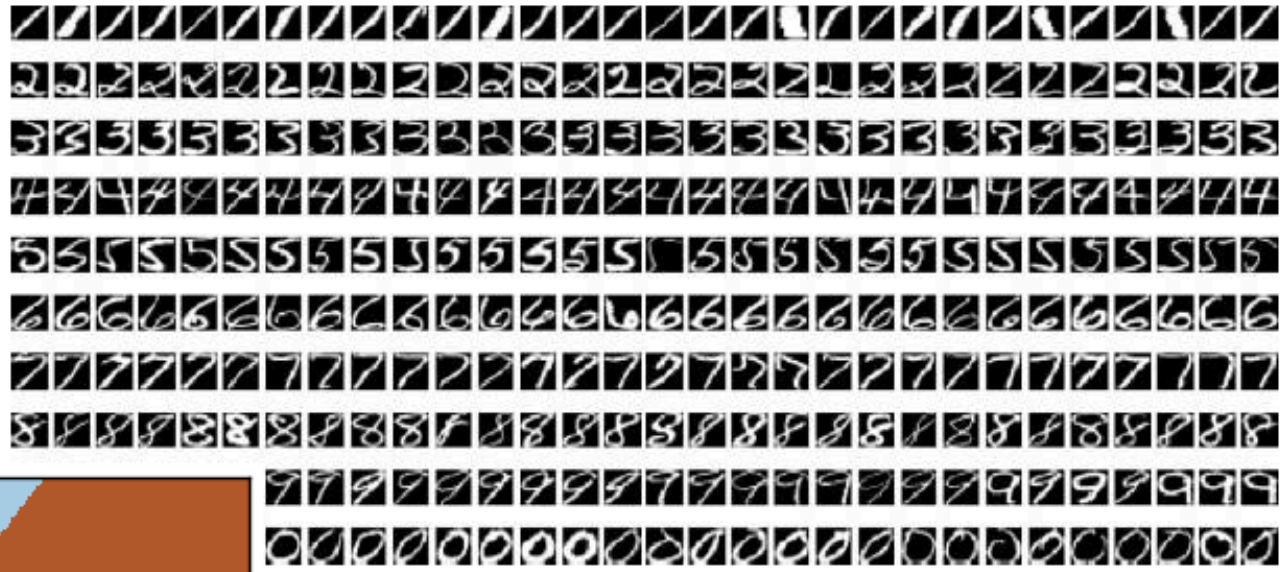
$$l(y, f(x))$$

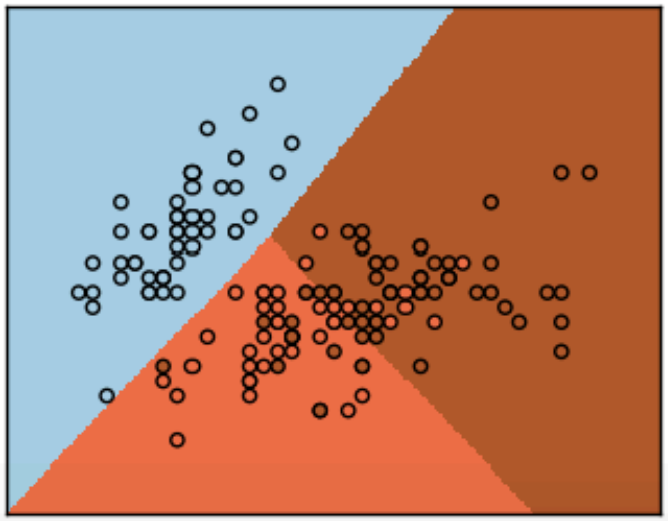
Binary Classification





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

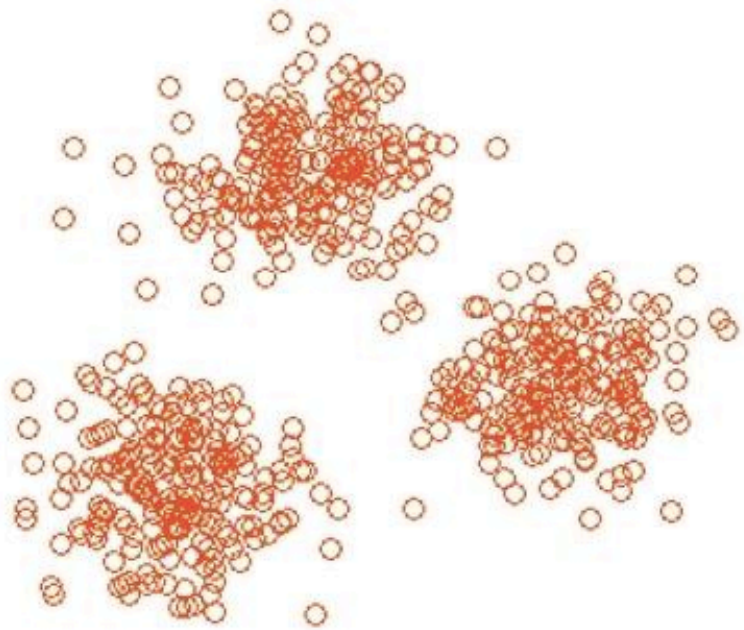
Unsupervised Learning

- 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

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

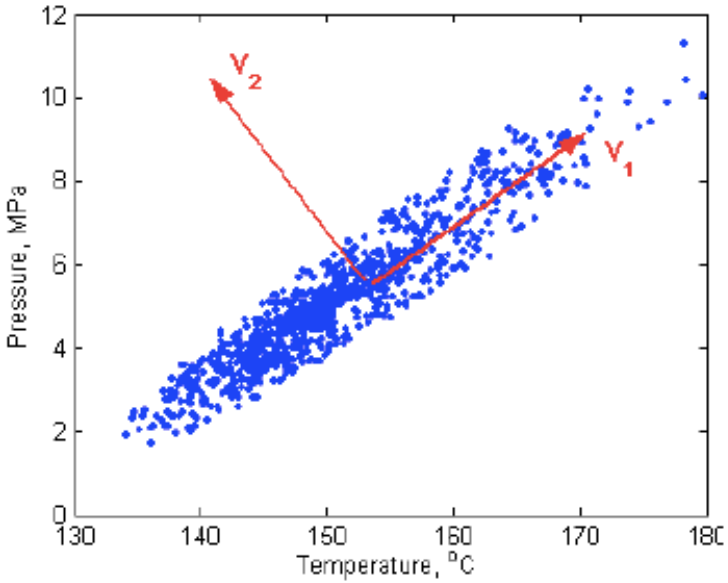
Clustering

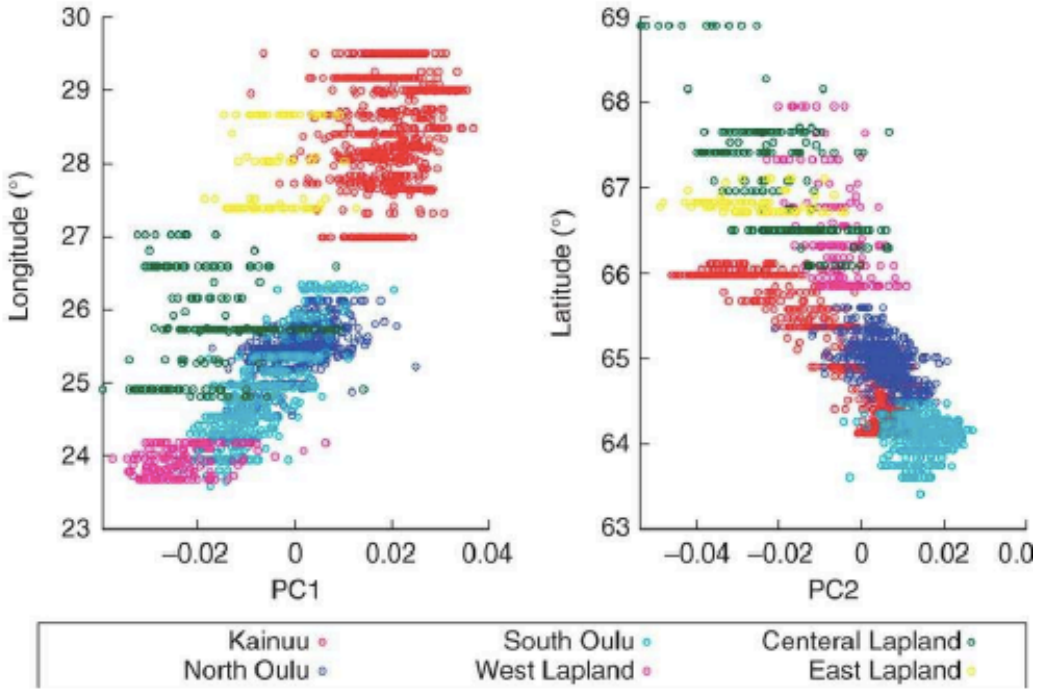


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

...

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)

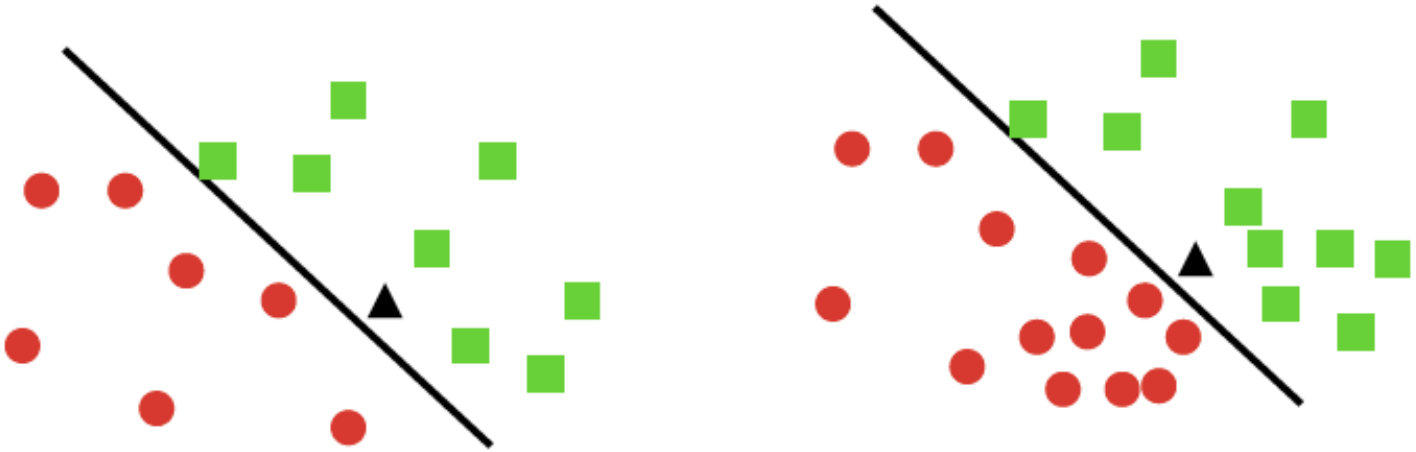
Discriminative vs. Generative (mainly relevant for supervised models)

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

Discriminative vs. Generative (mainly relevant for supervised models)

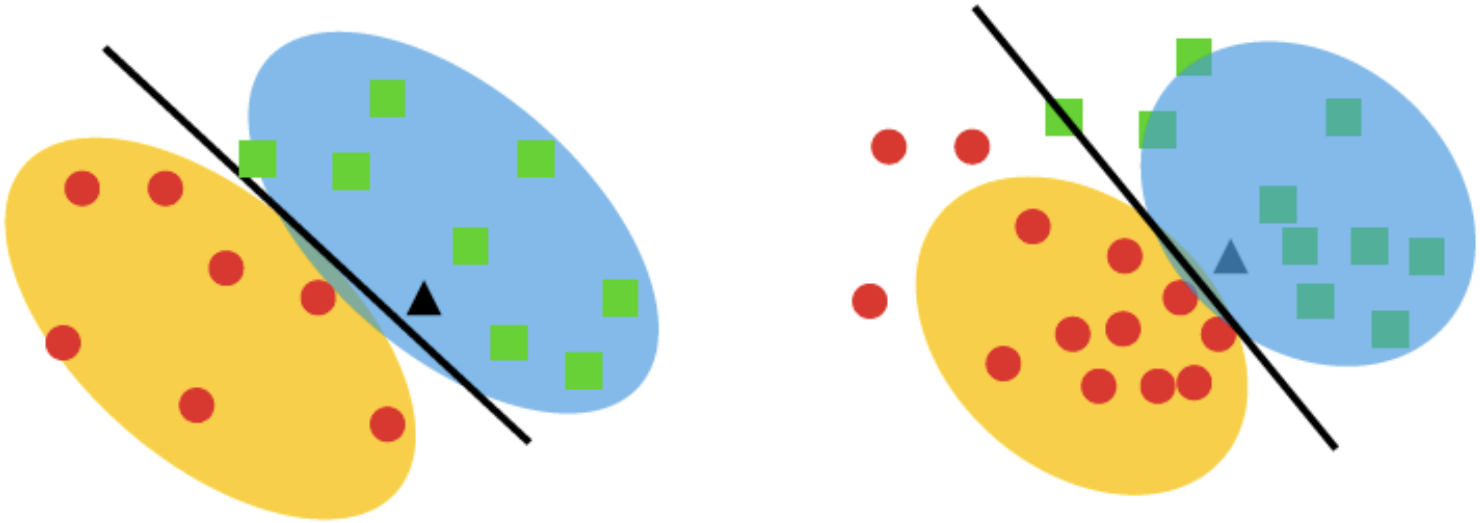
- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions
- Generative models
  - Estimate joint distribution over (x,y)
  - Use conditional probability to infer y|x
  - Often more intuitive
  - Easier to add prior knowledge

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

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

8

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*

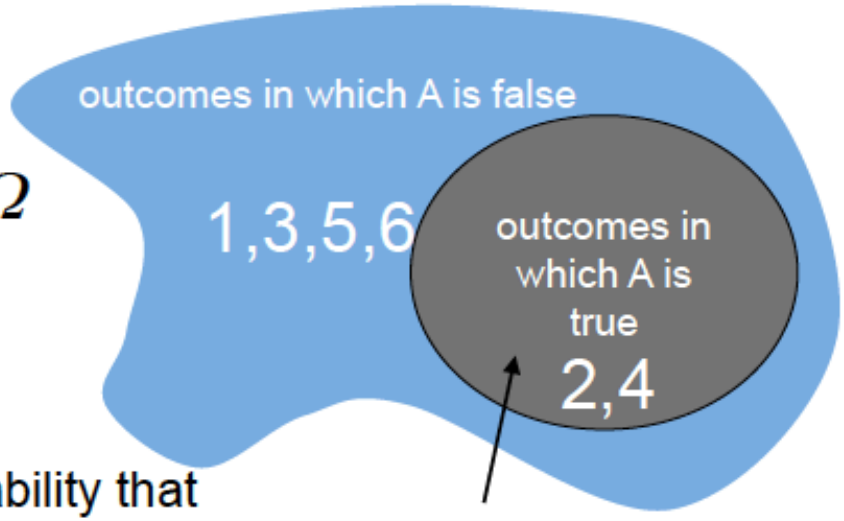
Probability

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.

10

5 minutes break

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

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

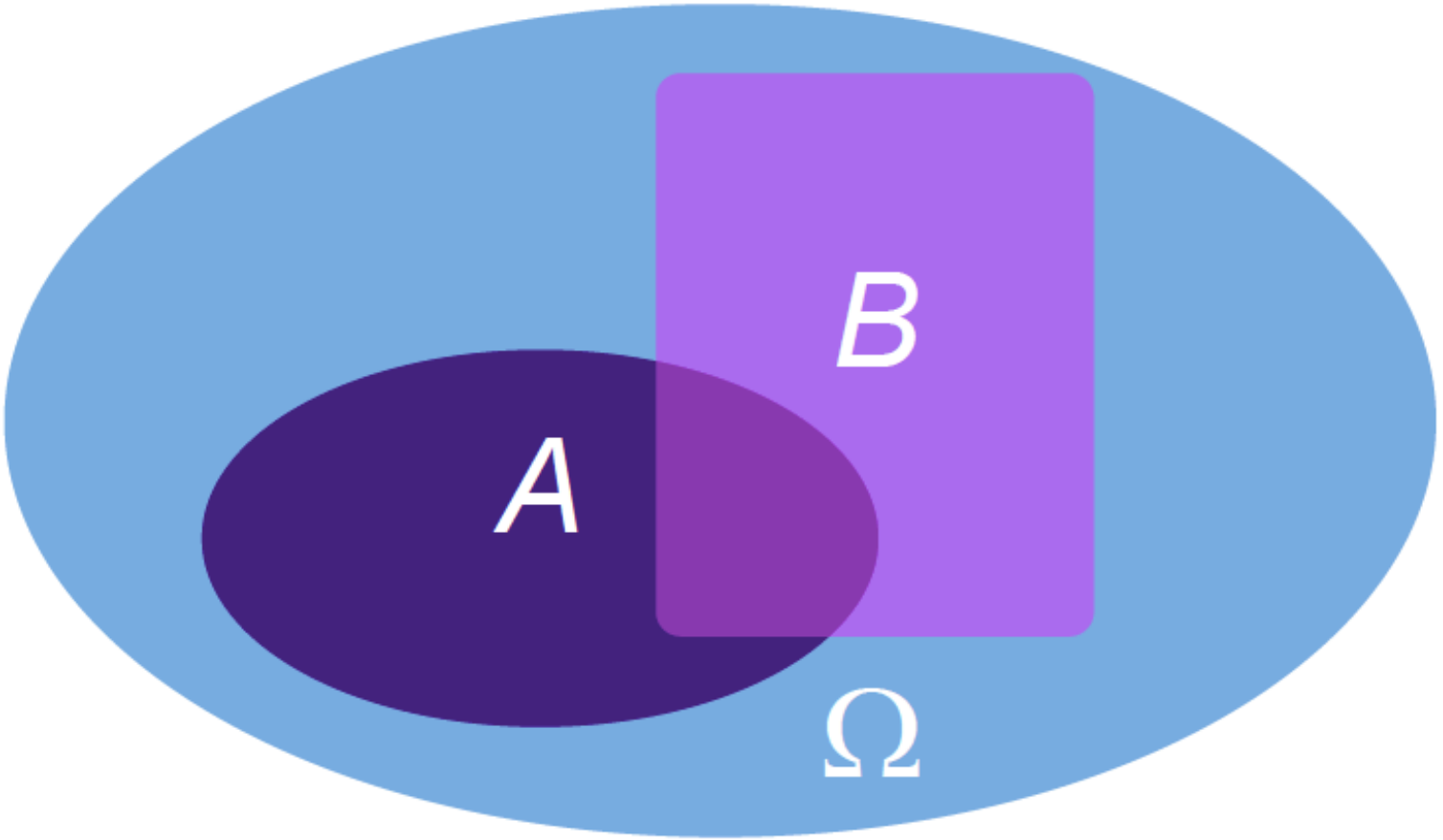
Consequences:

$P(\emptyset) = 0$.

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

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

Venn Diagram



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

13

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

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

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

Examples:

- **Discrete random variable examples (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)

14

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

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

15

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

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

17

Continuous Distribution

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

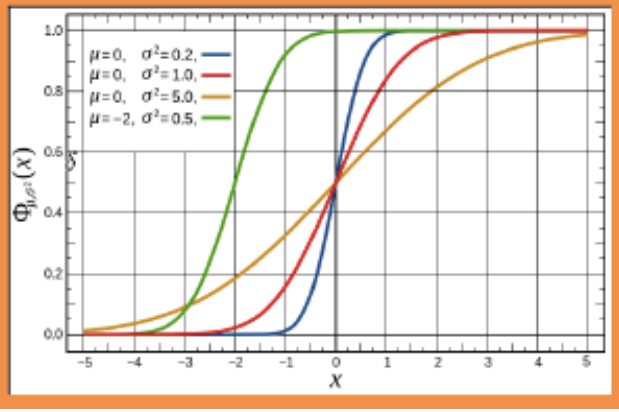
Continuous Distribution

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

Def: cumulative distribution function

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

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



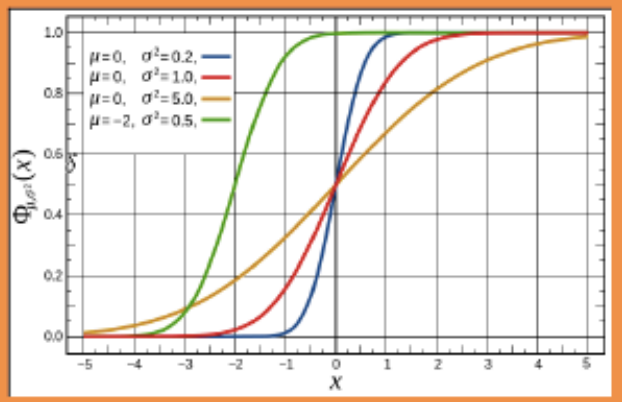
Continuous Distribution

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

Def: cumulative distribution function

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

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



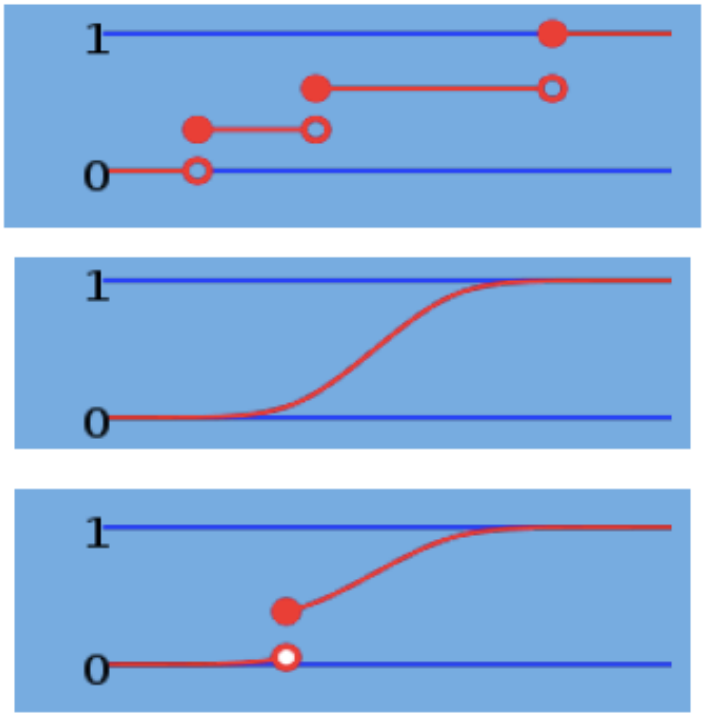
Def : Let $F(-\infty) = 0$. $F : (-\infty, \infty) \to \mathbb{R}$ is **absolutely continuous** $F(x) = \int_{-\infty}^{x} f(t)dt$ for some function f.

Def : f is called the density of the distribution.

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

18

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.

19

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???

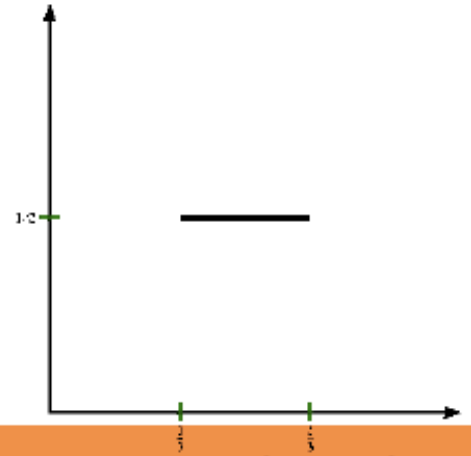
Cumulative Distribution Function (cdf)

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.

20

Probability Density Function (pdf)

Probability Density Function (pdf)

Pdf properties:

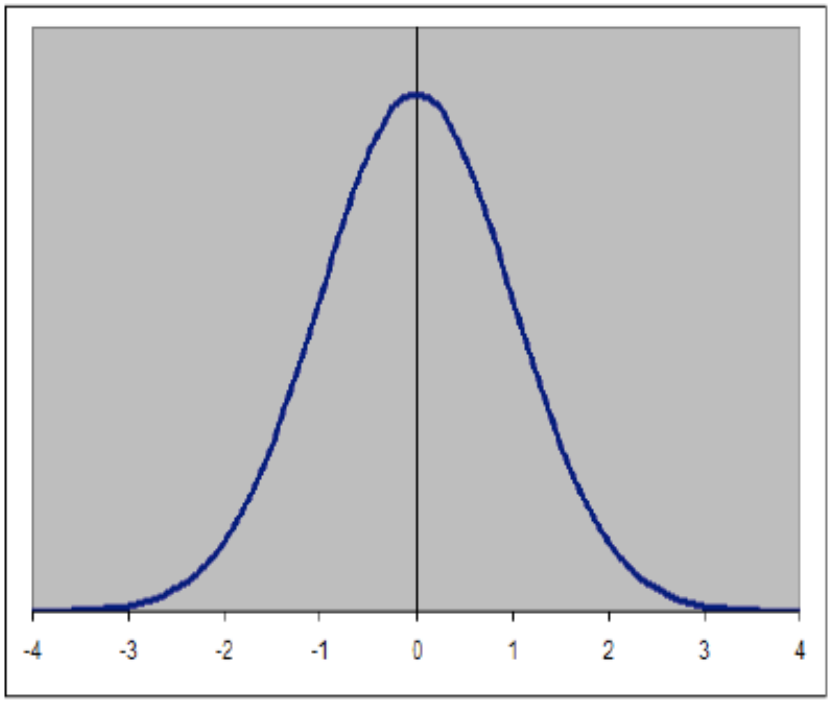
$$f(x) \geq 0$$

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

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

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

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



Probability Density Function (pdf)

Pdf properties:

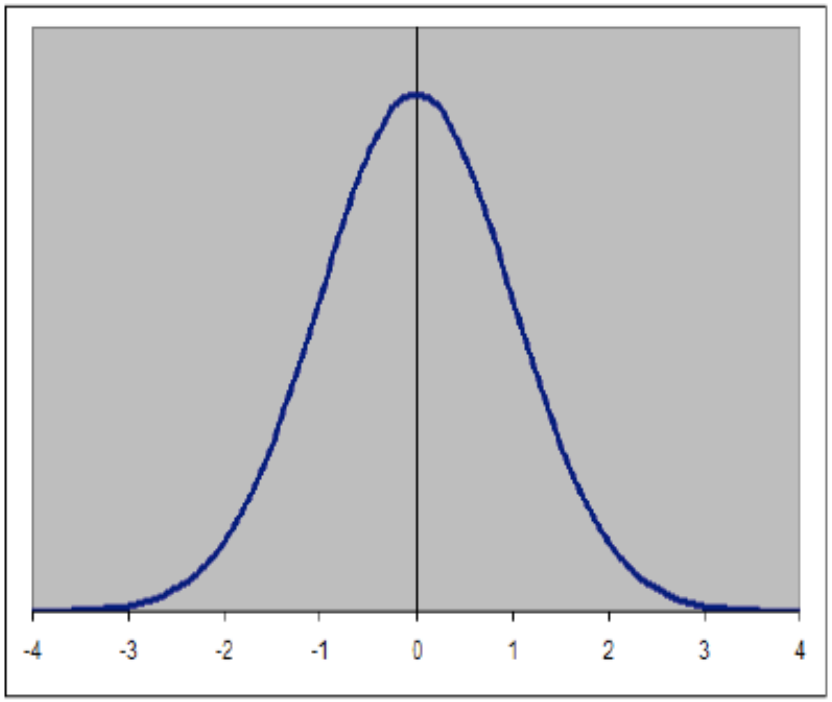
$$f(x) \geq 0$$

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

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

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

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

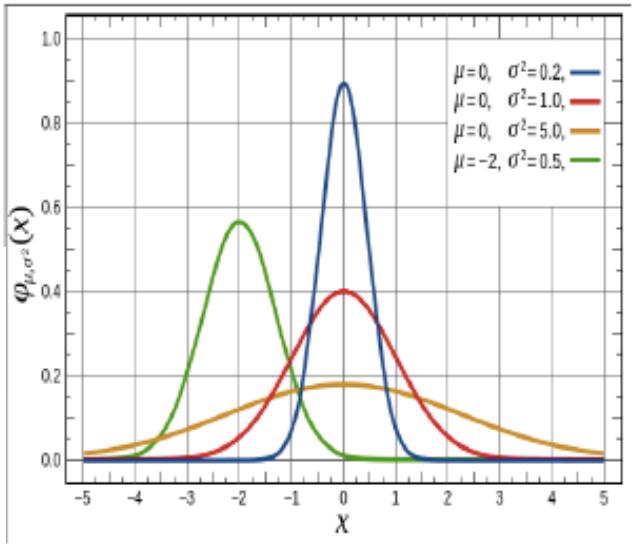


Intuitively, one can think of f(x)dx as being the probability of X falling within the infinitesimal interval [x, x + dx]. $P(x < X < x + dx) = f(x)dx$

Moments

Expectation: average value, mean, 1st moment:

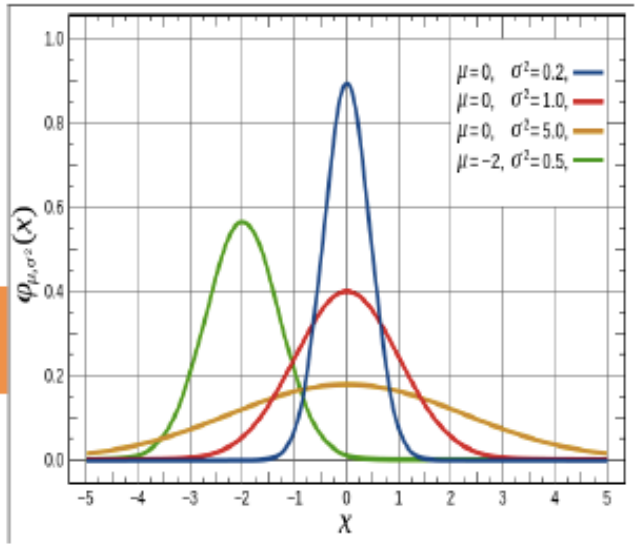
$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$



Moments

Expectation: average value, mean, 1st moment:

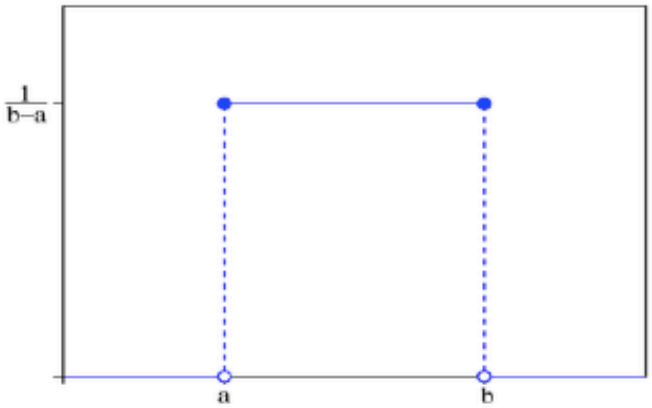
$$E(X) = \begin{cases} \sum_{i \in \Omega} x_i p(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & \text{continuous} \end{cases}$$



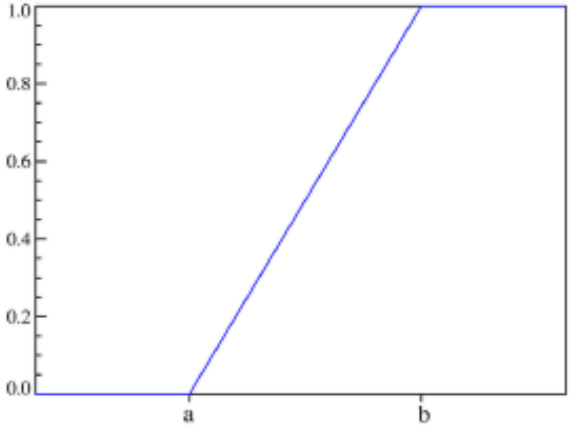
Variance: the spread, 2nd moment:

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

Uniform Distribution



PDF

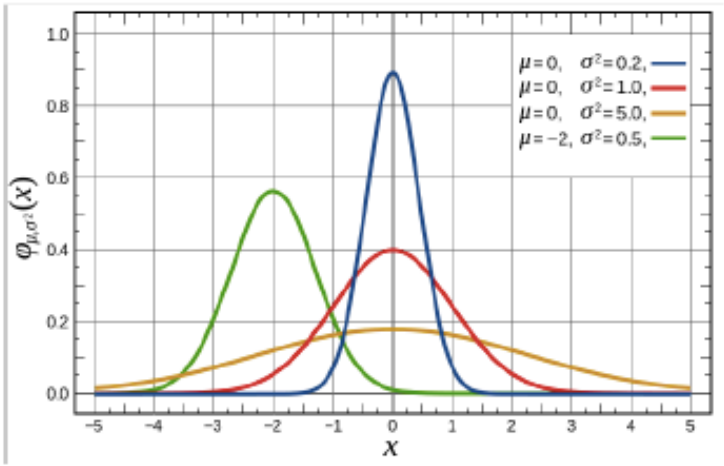


CDF

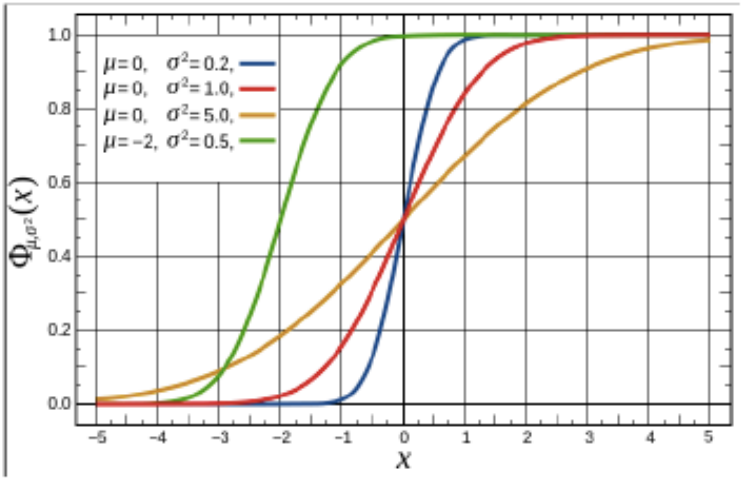
$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{Otherwise} \end{cases}$$

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

Normal (Gaussian) Distribution



PDF



CDF

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

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

Multivariate (Joint) Distribution

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

$P(a \leq X \leq b, c \leq Y \leq d) = ?$ $P(a_1 \leq X_1 \leq b_1, \ldots a_d \leq X_d \leq b_d) = ?$

Multivariate (Joint) Distribution

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

$P(a \leq X \leq b, c \leq Y \leq d) = ?$ $P(a_1 \leq X_1 \leq b_1, \ldots a_d \leq X_d \leq b_d) = ?$

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

26

Multivariate Gaussian distribution

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

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

Multivariate Gaussian distribution

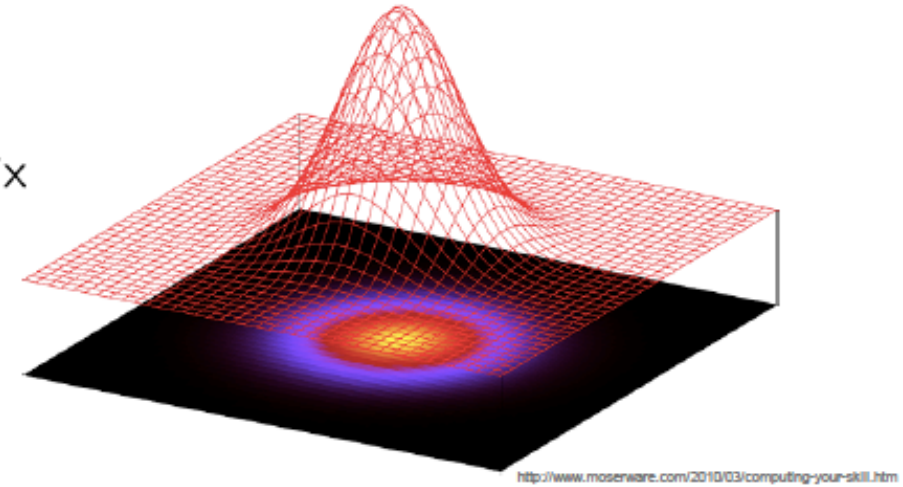
For $A \subset \mathbb{R}^d$, $P([X_1, \ldots, X_d] \in A) = \int_A f(x_1, \ldots, x_d) dx_1 \cdots dx_d$

$$F_X(z_1, \ldots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \ldots, x_d) dx_1 \cdots dx_d$$

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

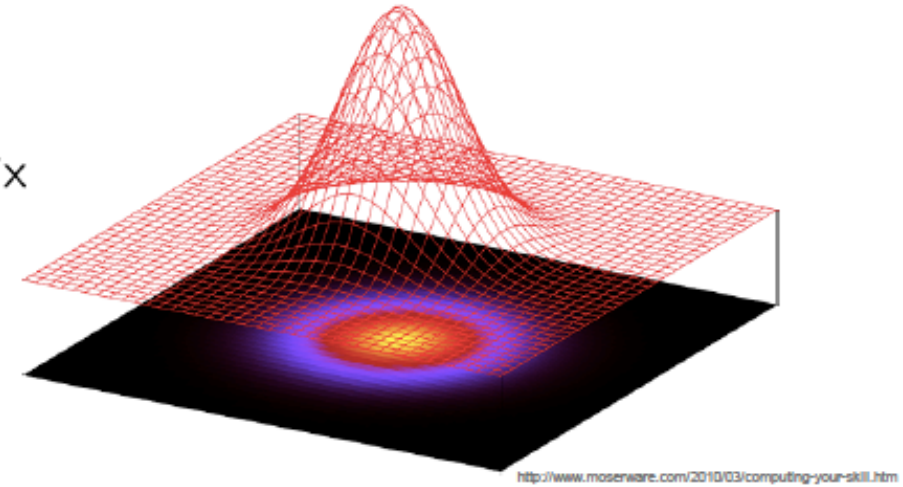
Multivariate Gaussian distribution

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

$$F_X(z_1, \ldots, z_d) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} f(x_1, \ldots, x_d) dx_1 \cdots dx_d$$

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

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

27

Conditional Probability

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

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

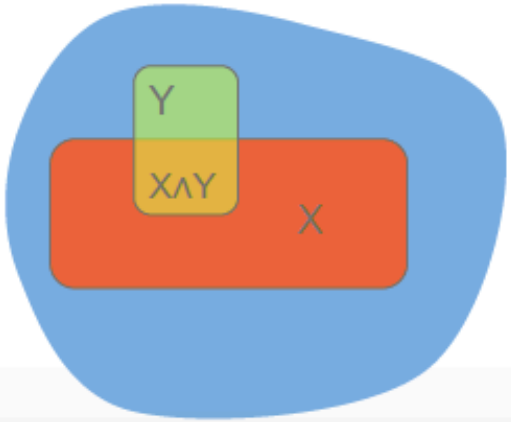
Conditional Probability

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



28

Independence

Independent random variables:

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

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

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

29

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

Conditionally Independent

Conditionally independent:

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

Knowing Z makes X and Y independent

Examples:

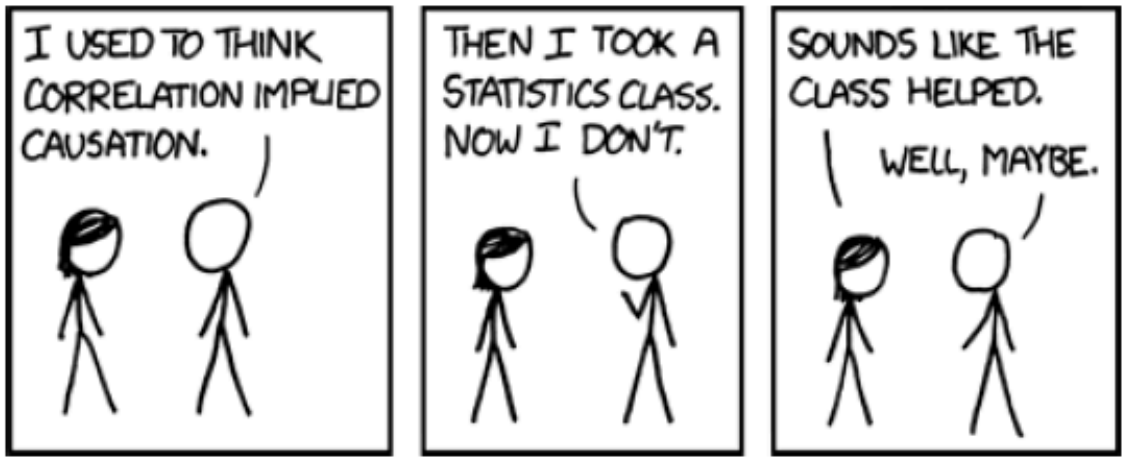
Dependent: show size and reading skills

Conditionally independent: show size and reading skills given age

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…



xkcd.com

Conditional Independence

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

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

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

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

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:

$$P(a=1|t=1) = \frac{P(t=1|a=1)P(a=1)}{P(t=1)}$$

$$= \frac{P(t=1|a=1)P(a=1)}{P(t=1|a=1)P(a=1) + P(t=1|a=0)P(a=0)}$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Only 9%!...

35

Improving the diagnosis

Use a follow-up test!

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

Improving the diagnosis

Use a follow-up test!

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

$$P(a=0|t_1=1,t_2=1) = \frac{P(t_1=1,t_2=1|a=0)P(a=0)}{P(t_1=1,t_2=1|a=1)P(a=1)+P(t_1=1,t_2=1|a=0)P(a=0)}$$

$$= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

$$P(a=1|t_1=1,t_2=1) = 0.643$$

Why can't we use Test 1 twice?

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

$$p(t_1,t_2|a) = p(t_1|a) \cdot p(t_2|a)$$

36

Naïve Bayes Assumption

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

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

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

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 $(2-1)dK$

39

Naïve Bayes Classifier

Given:

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

Naïve Bayes Classifier

Given:

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

Decision rule:

$$f_{NB}(\mathbf{x}) = \arg\max_y P(x_1,\ldots,x_d \mid y)P(y)$$

$$= \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$

40

Naïve Bayes Algorithm for discrete features

Training Data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

n d dimensional features + class labels

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

We need to estimate these probabilities!

Naïve Bayes Algorithm for discrete features

Training Data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

n d dimensional features + class labels

$$f_{NB}(\mathbf{x}) = \arg\max_y \prod_{i=1}^{d} P(x_i|y)P(y)$$ **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:

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

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

42

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='Earn' when Y='SpamEmail' in our dataset.

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

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='Earn' when Y='SpamEmail' in our dataset.

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

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

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

$$P(Y = b \mid X_1 = a, X_2, \ldots, X_d) = 0$$

What now???

43

That’s all!