DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Statistical Methods for Machine Learning

Kim Steenstrup Pedersen

Teachers and Instructors

- Christian Igel, igel@di.ku.dk, DIKU (course responsible)
- Kim Steenstrup Pedersen, kimstp@di.ku.dk, DIKU

Instructors:

- Pengfei Diao, diao@di.ku.dk
- Oswin Krause, oswin.krause@di.ku.dk
- Didac Rodriguez Arbonès, didac@di.ku.dk
- Jan Kremer, ja.kremer@di.ku.dk
- Niklas Kasenburg, niklas.kasenburg@di.ku.dk


Course Goals

- To give you an introduction to stochastic data analysis and modeling
- To introduce you to the most common machine learning and pattern recognition techniques
- To introduce common examples of usage of such techniques within different areas of science
- To introduce the fundamental ideas of some more advanced techniques
- You will gain a practical oriented knowledge of machine learning and pattern recognition theory

Learning Outcome (see course description)

Learning Outcome

At course completion, the successful student will have:

Knowledge of

- the general principles of machine learning;
- basic probability theory for modeling and analyzing data;
- the theoretical concepts underlying classification, regression, and clustering;
- the mathematical foundations of selected machine learning algorithms;
- common pitfalls in machine learning.

Skills in

- applying linear and non-linear techniques for classification and regression;
- performing elementary dimensionality reduction;
- elementary data clustering;
- implementing selected machine learning algorithms;
- visualizing and evaluating results obtained with machine learning techniques;
- using software libraries for solving machine learning problems;
- identifying and handling common pitfalls in machine learning.

Competences in

- recognizing and describing possible applications of machine learning;
- comparing, appraising and selecting machine learning methods of for specific tasks;
- solving real-world data mining and pattern recognition problems by using machine learning techniques.

We assume that you know

- Basic mathematical analysis (high school level and DiMS or MatIntro) and linear algebra (vectors and matrices)
- Assignment 1 includes a math quiz – use it as a guide!
- Probability theory at high school level
- Programming at an introductory level (we will use either Matlab, R, Python, or C/C++ - it is up to you)

Be aware:

- You are a mixed crowd with different backgrounds!
- There might be parts you find trivial and other parts you won’t.

When and where

- Lectures:
  - Tuesday 10:15 - 12:00, Room: DIKU Aud. 4.1.22 (lille UP1)
  - Thursday 13:15 - 15:00, Room: DIKU Aud. 4.1.22 (lille UP1)

- Exercise classes:
  - Thursday 9:15 - 12:00, Rooms:
    - Class 1: DIKU-NC 1.0.04
    - Class 2: DIKU-NC 3.1.25
    - Class 3: DIKU-NC 1.0.37
    - Class 4: DIKU-NC 1.0.26
    - Class 5: DIKU-NC 4.0.17

- You have been assigned to one of these exercise classes (you can see which in Absalon).

Format of exercise classes

- Main purpose: To work on the assignments
- You can get individual help with the assignments while you work on them
- The TAs will sometimes lead a general discussion of the current lectures and assignment as well as provide general feedback on finished assignments
- The exercise rooms have no computer terminals.
- So bring your laptop!

Mandatory assignments

3 mandatory assignments + 1 exam assignment:

- A mix of theoretical and practical problems
- Two weeks to solve each of them
- Necessary theory will be presented at lectures
- The solutions can be made individually or in groups of no more than 3 participants
- Help from the TAs at the exercise class
- Feedback at exercise class
- Use the discussion forum!

How do I pass this course?

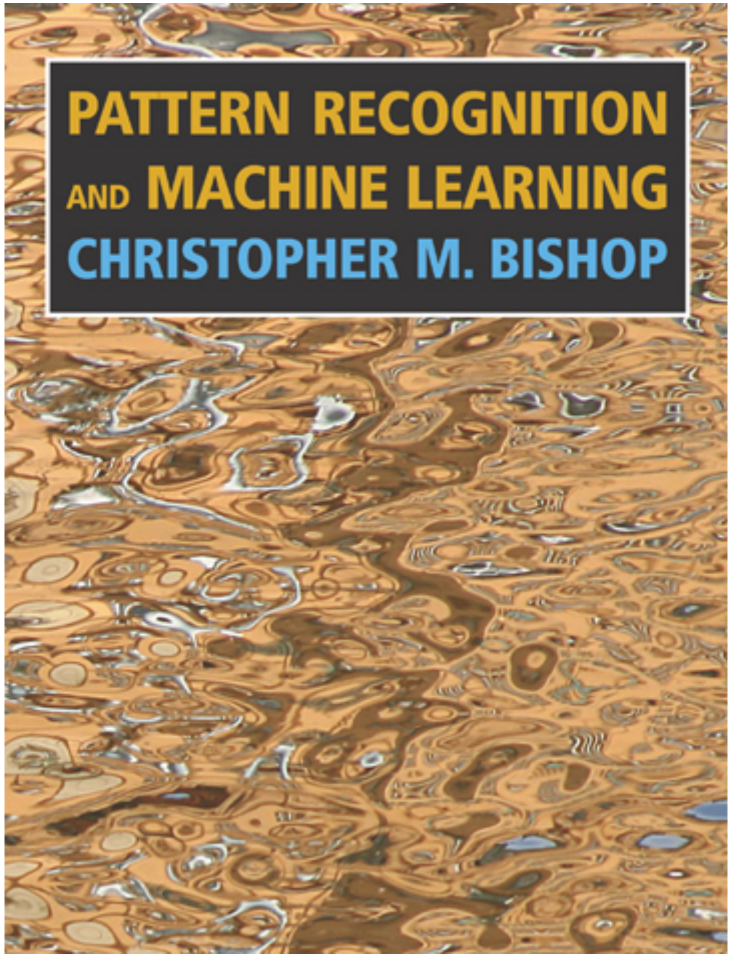
- Must pass the 3 mandatory assignments to be eligible for participating in the exam.
- If you do not pass an assignment the first time you will be given a second chance to submit a new solution (assuming that you have made a SERIOUS attempt the first time).

- **Exam assignment:** Larger written assignment similar to the other mandatory assignments.
- This assignment must be solved individually, but we encourage you to discuss it with your fellow students.
- Final grading for the course is: 7-point grading based on the exam assignment only.

How much time should I spend on this course?

- KU expect that you use ca. 20 hours / week for a 7.5 ECTS course. Approx. 40 hours/wk for full time study. (Yes, it is more than the 37.5 hours/wk common out in real life, i.e. according to Danish union agreements)
- **How should I spend my time?**
  - Lectures and exercise class = 2 + 2 + 3 = 7 hours/wk
  - Reading and assignment = 20 – 7 = 13 hours/wk
- **Recommended:**
  - Prepare by reading the current weeks material at least prior to the lectures (approx. 6 hours/wk). Be prepared for the exercise class – at least read the assignment
  - Work on the assignment at home (approx. 7 hours/wk) (and you have spend 2-3 hours on the assignment in exercise class)

Course Material



- **Challenge:**
  - If you find the book not sufficiently mathematical, write out the proofs yourself.
  - If you find the book too mathematical, draw figures to understand what the math describes.

Course Home Page and Information

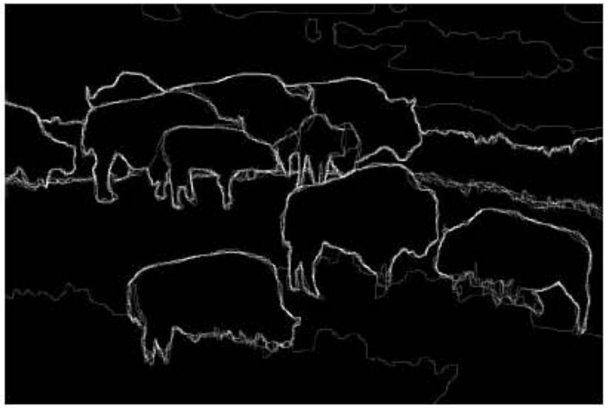
- We use Absalon (access via your KUnet account)
  - You will find latest lecture (we use the Planner) and exercise schedules
  - Links to lecture slides (usually after the lecture)
  - Exercise material
  - Course material (reading material)
  - Links to additional material (reading, programming, etc.)
  - A discussion forum for course related topics

Tentative Lecture and Exercise Plans

- Week 1: Probability theory and estimation
- Week 2: Basic learning theory and regression I
- Week 3: Regression II and Linear Classification
- Week 4: Neural networks
- Week 5: Kernel methods
- Week 6: Unsupervised learning & clustering and PCA
- Week 7: Trees, forests and some more learning theory

Let’s get started

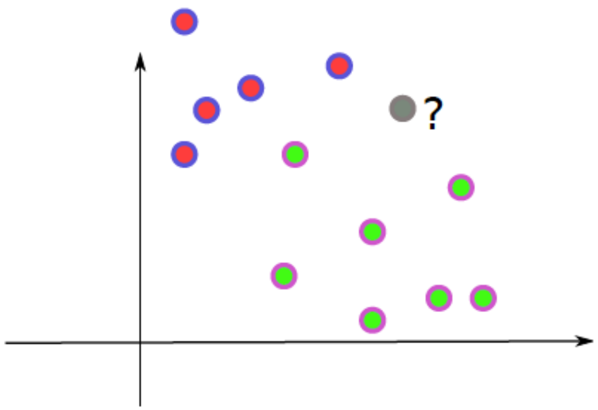
Machine Learning/Data Mining/Pattern Recognition



- **Example 1: Image segmentation**
  - Split the image into "objects" (foreground) and "irrelevant" (background).
- Classification of voxels x into classes:
  - y(x) = 1 (foreground)
  - y(x) = 0 (background)

Images from Berkeley Segmentation Dataset and Benchmark,
https://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
D. Martin and C. Fowlkes and D. Tal and J. Malik, ICCV 2001

Machine Learning/Data Mining/Pattern Recognition



Classification splits data x into a finite number of classes:

- y(x) = 1 (foreground)
- y(x) = 0 (background)

General goal of ML:

- Model a mapping (rule) between data x and some abstract description y(x) of the data.

Supervised learning

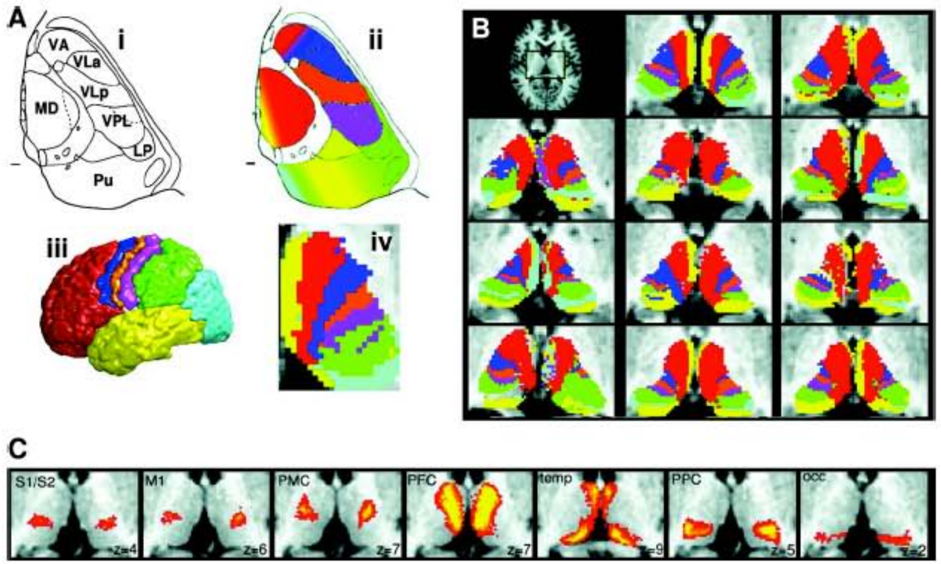
- We know the rule y for a set of data (the training set) and try to learn a general rule y

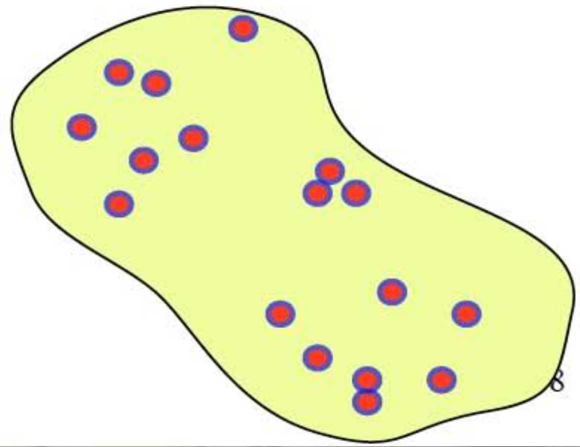
Machine Learning/Data Mining/Pattern Recognition

- **Example 2: Stock market prediction**
- Regression
  - $y(x)$ = stock price
  - Predicting a continuous variable
- Also a case of **supervised** learning

Machine Learning and Pattern Recognition?

- **Example 3: Clustering**
  - Cluster brain MRI voxels with respect to connectivity
- Example of **unsupervised learning:**
  - No known values y(x)
  - Don't know which clusters we are looking for

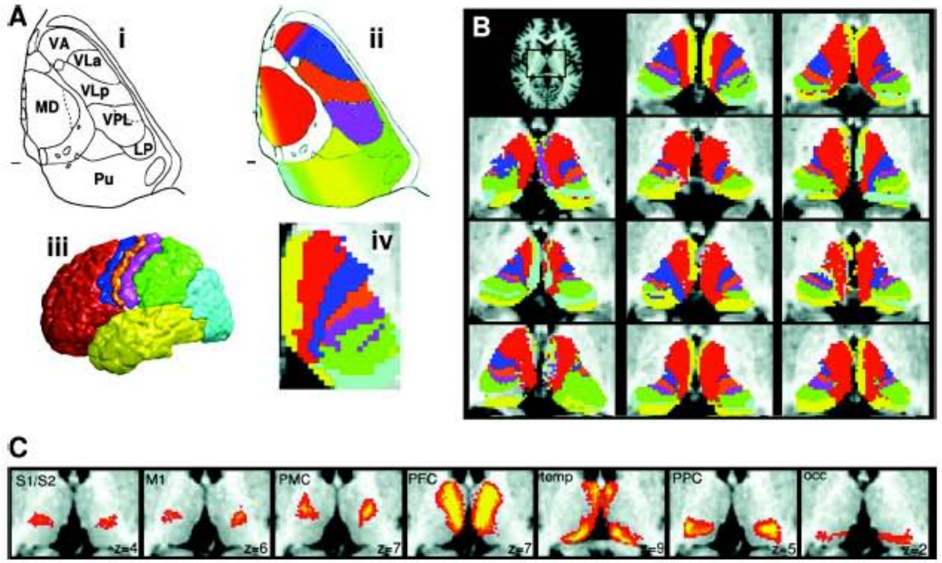


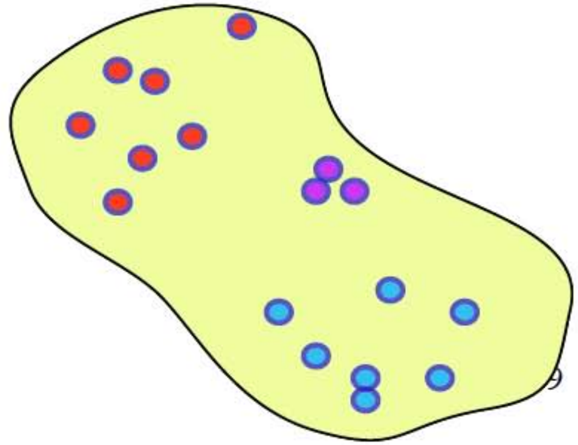


Johansen-Berg et al, Cerebral Cortex 2005

Machine Learning and Pattern Recognition?

- **Example 3: Clustering**
  - Cluster brain MRI voxels with respect to connectivity
- Example of **unsupervised learning:**
  - No known values y(x)
  - Don't know which clusters we are looking for

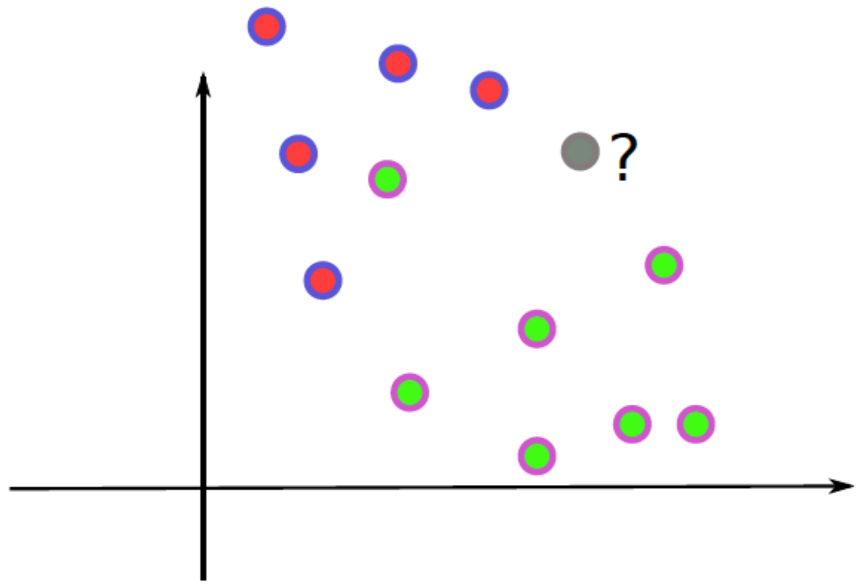




Johansen-Berg et al, Cerebral Cortex 2005

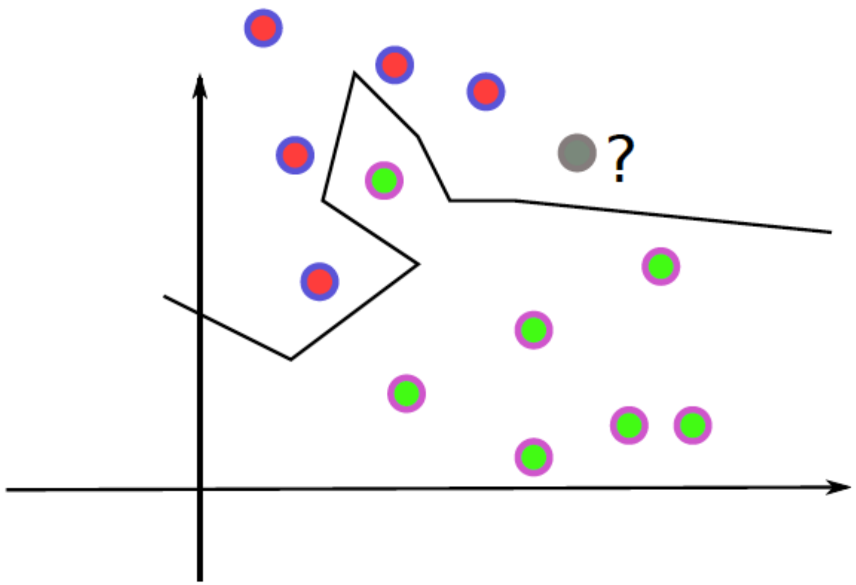
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



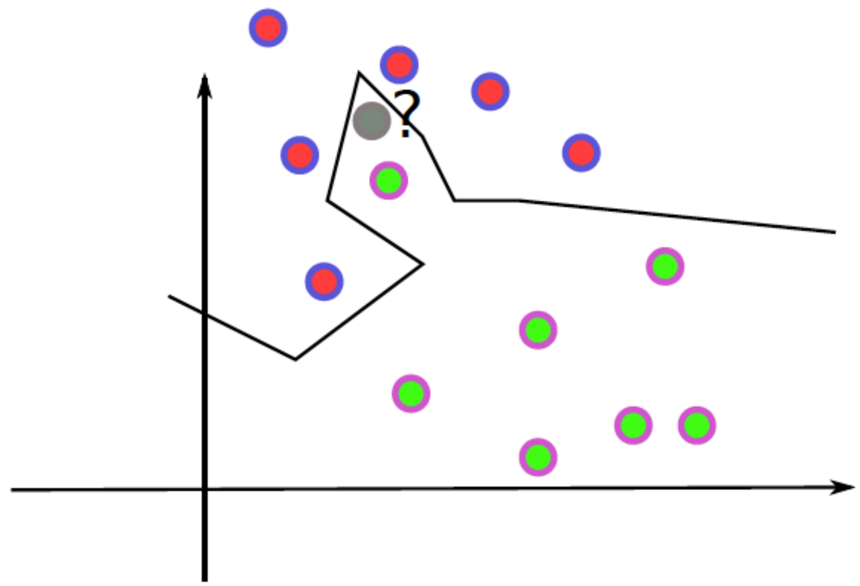
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



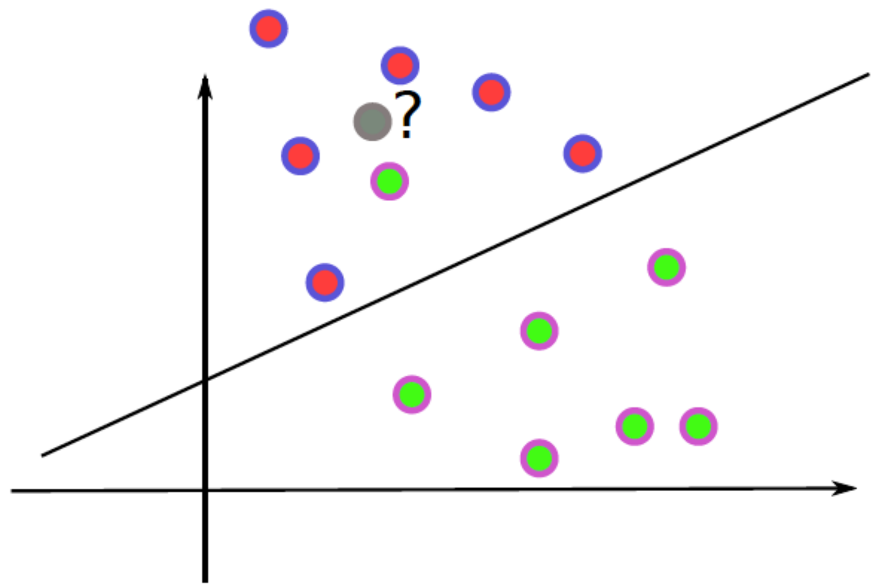
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



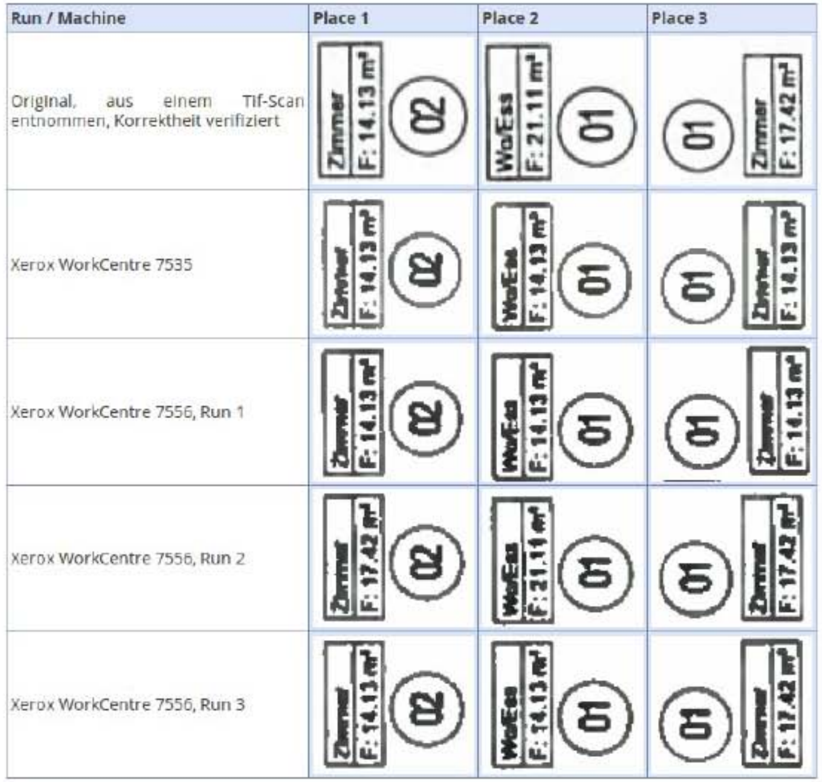
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).
- **Example 3: Xerox**
  - **July 2013:** Xerox scanners were found to mangle numbers in documents
  - **Cause:** JBIG2 compression algorithm replacing image patches by “similar” image patches from a database
  - Model for “similar” did not generalize
  - Unexpected impact of ML: legal documents scanned, etc...



http://www.dkriesel.com/en/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning

Summary of ML principles

- General ML task:

  Learn rule y(x) which predicts a target t from measured data x

- **Unsupervised learning:**
  - No examples of y(x)
  - Examples:
    - Clustering

- **Supervised learning:**
  - Have a set of examples x for which y(x) is known (training set)
  - Learn a function y from the training set
  - Check generalizability to test set
  - Examples:
    - Classification – discrete target t
    - Regression – continuous target t

Plan for lectures on probability and estimation (this and the next lectures)

- Why *Statistical* Machine learning?
- Probability theory and statistics 101 (crash course or reminder)
- Bayesian probabilities
- The Gaussian / Normal distribution
- Parametric and non-parametric estimation of probability distributions
  - Maximum likelihood and maximum a posteriori estimation
  - Histograms as example of non-parametric methods (more to come later in the course)
- Curse of dimensionality

Why *Statistical* Machine learning?

Why do we need probabilistic descriptions?

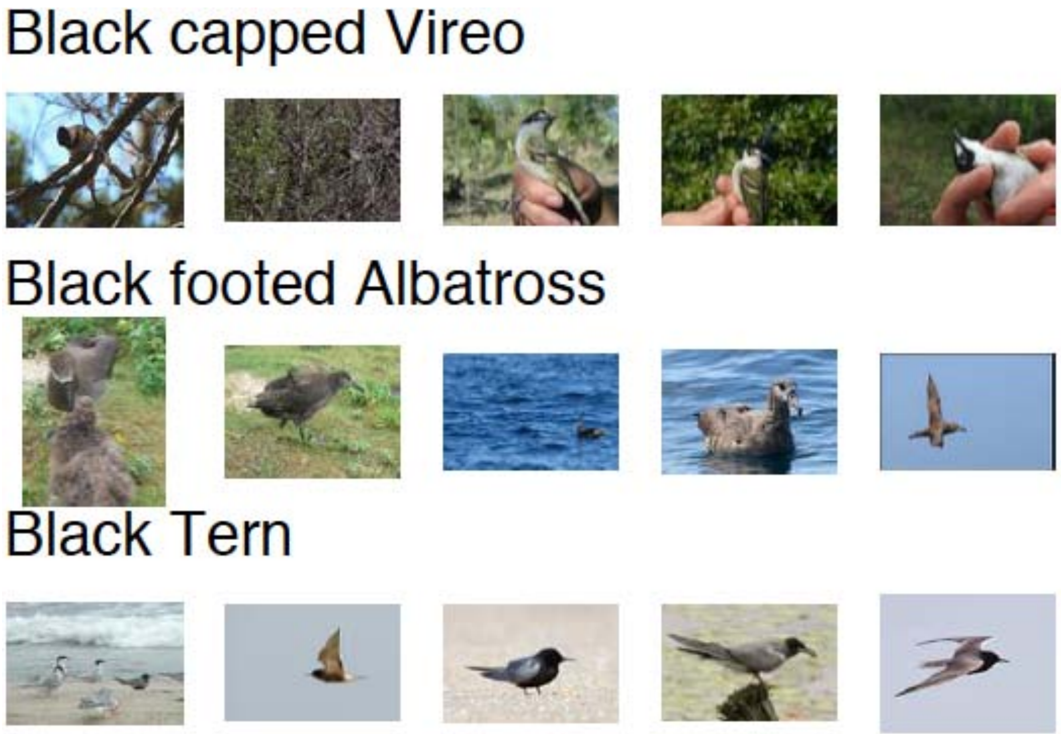
- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules

- Example variability in handwritting

http://yann.lecun.com/exdb/mnist/index.html

Why do we need probabilistic descriptions?

- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules (e.g. biological variation)
  - Are inherently stochastic (e.g. view point)
  - Are noisy (e.g. caused by sensory noise)



Why do we need probabilistic descriptions?

- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules (e.g. biological variation)
  - Are inherently stochastic (e.g. view point)
  - Are noisy (e.g. caused by sensory noise)

Hence a probabilistic description is most often needed.

- For probabilistic models we need to be able to represent and estimate probability distributions either:
  - Parametric
  - Non-parametric

Probability Theory 101

Probabilities (a quick brush-up)

- Random variables: X, Y
- Realizations / instances of the variables: $X = x, Y = y$

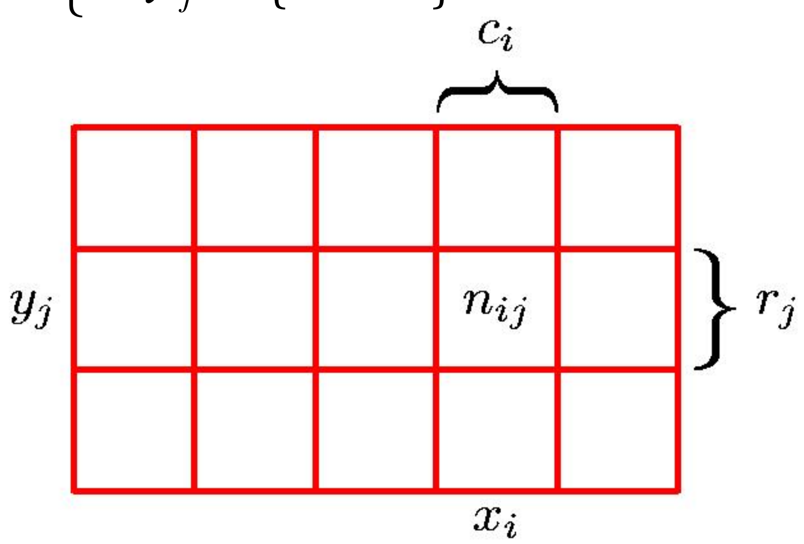
- Frequentists example: Make an experiment with *N* trials. Each trial gives two discrete outputs (two discrete random variables): $X = x_i \ , \ i = 1, \ldots, 5$ $\quad Y = y_j \ , \ j = 1, \ldots, 3$

Example: $\begin{cases} x_i \in \{1,\ 2,\ 3,\ 4,\ 5\} \\ y_j \in \{1,\ 2,\ 3\} \end{cases}$

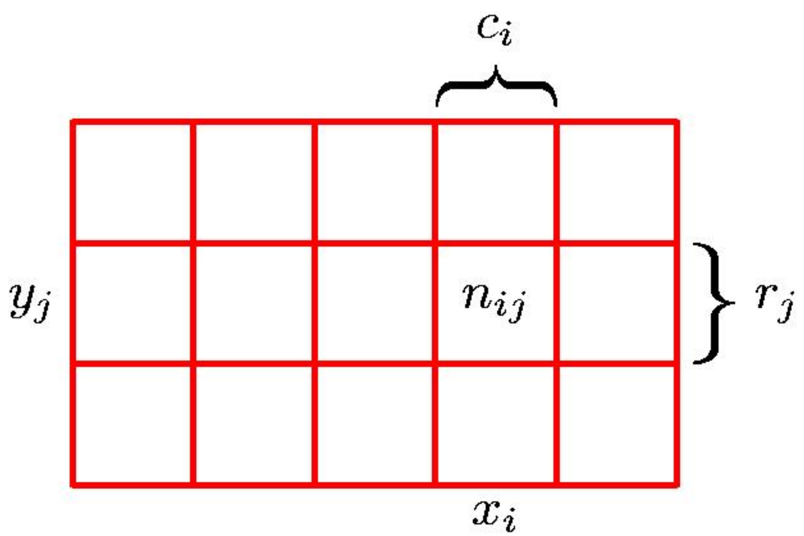
n_{ij} Number of trials with $(X = x_i \ , \ Y = y_j)$

$c_i = \sum_j n_{ij}$ number of $X = x_i$

$r_j = \sum_i n_{ij}$ number of $Y = y_j$



Probabilities (a quick brush-up)



Joint Probability:
Probability of the occurrence of (x_i, y_j)

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- **Marginal Probability**

$$p(X = x_i) = \frac{c_i}{N}$$

$$p(Y = y_j) = \frac{r_j}{N}$$

- **Conditional Probability**

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$$p(X = x_i \mid Y = y_j) = \frac{n_{ij}}{r_j}$$

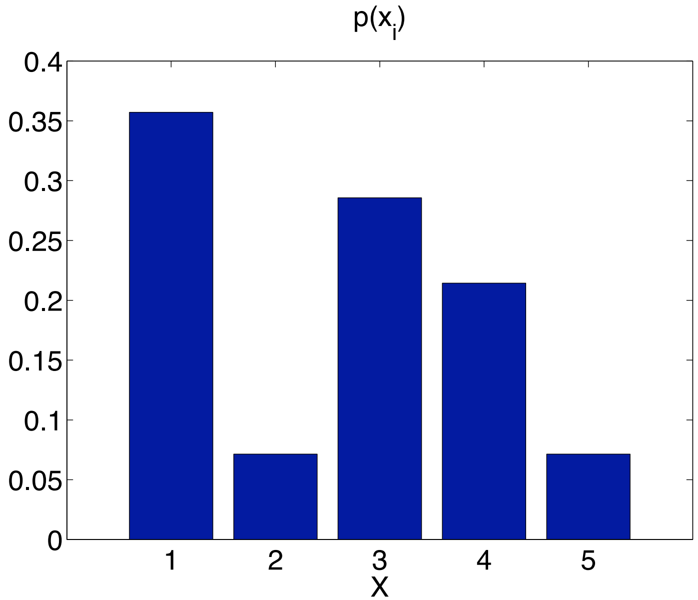
(These definitions are valid in the limit $N \rightarrow \infty$)

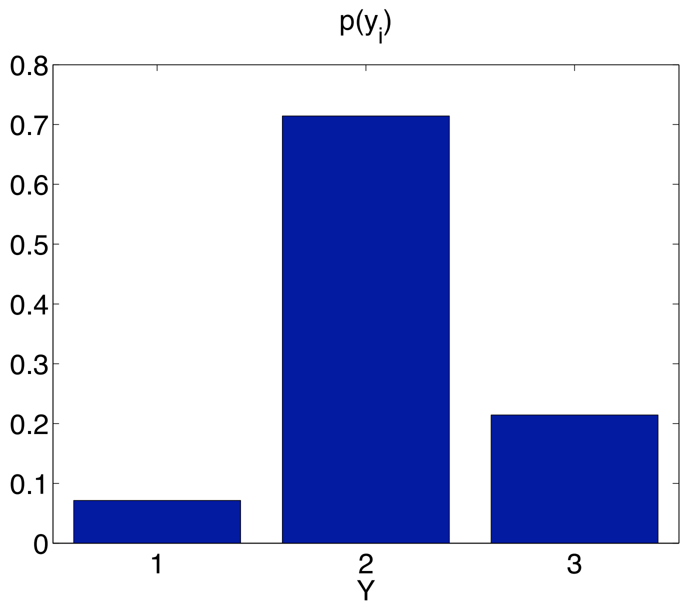
Probability mass functions (discrete variables)

For discrete random variables $p(x_i)$ is referred to as the probability mass function. It must fulfill these conditions:

$$0 \leq p(x_i) \leq 1$$

$$\sum_i p(x_i) = 1$$





Probability density (continuous variables)

Assume that X is a real random variable, $X \in \mathrm{R}$.

Now $p(x)$ is called the probability density of X.

The probability of X falling in the interval (a,b) is

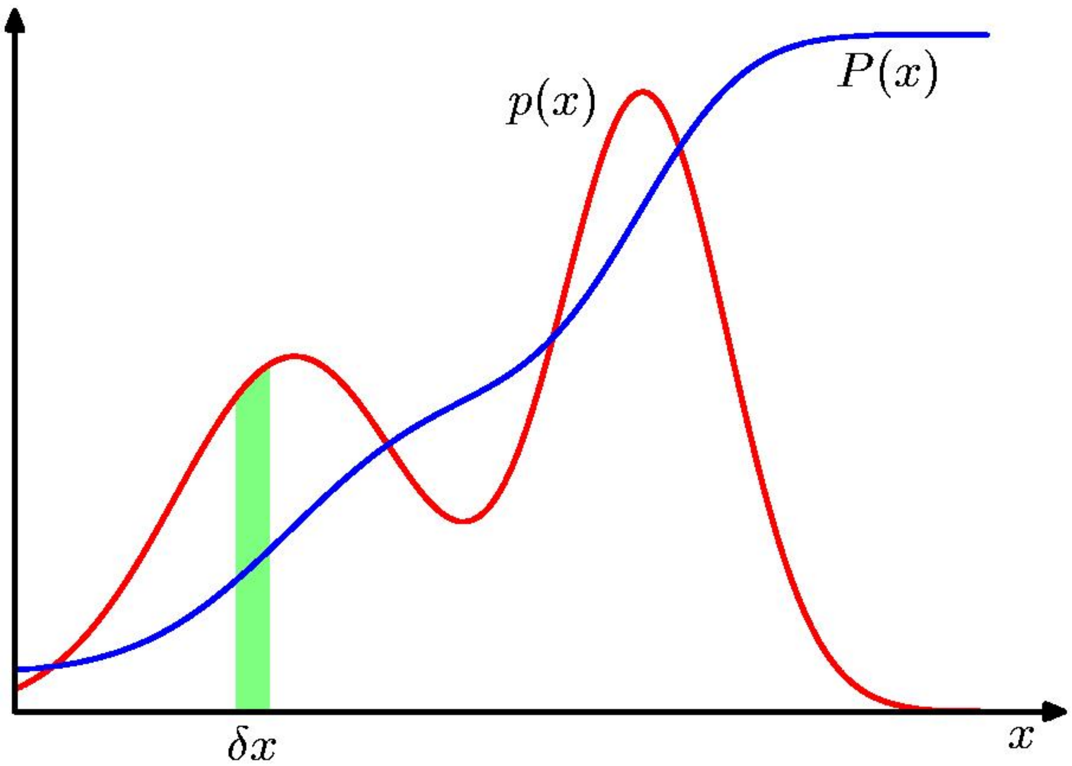
$$p(X \in (a,b)) = \int_a^b p(x)dx$$

Probability density conditions:

$$p(x) \geq 0 \;,\; \int_{-\infty}^{\infty} p(x)dx = 1$$

Aside: The cumulative distribution function:

$$P(z) \;=\; \int_{-\infty}^{z} p(x)dx$$



The Gaussian (a.k.a. Normal) Distribution
(A parametric representation)

The 1-dimensional Gaussian probability density:

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Fulfills the density conditions:

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

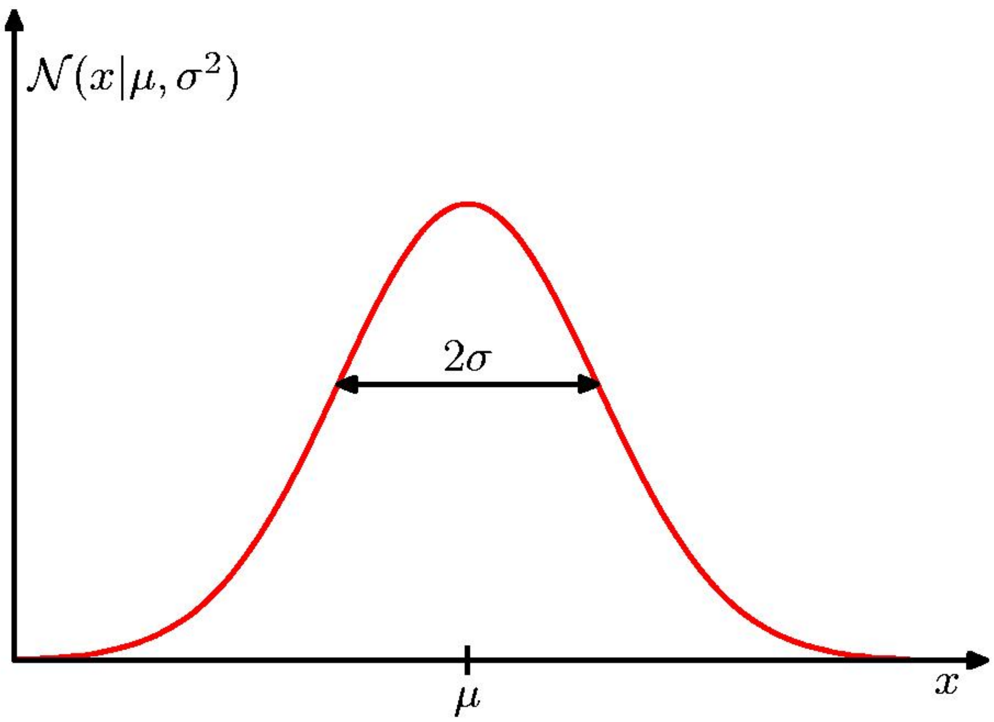
$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

μ : Mean

σ : Standard deviation

$\mathrm{var}[x] = \sigma^2$: Variance

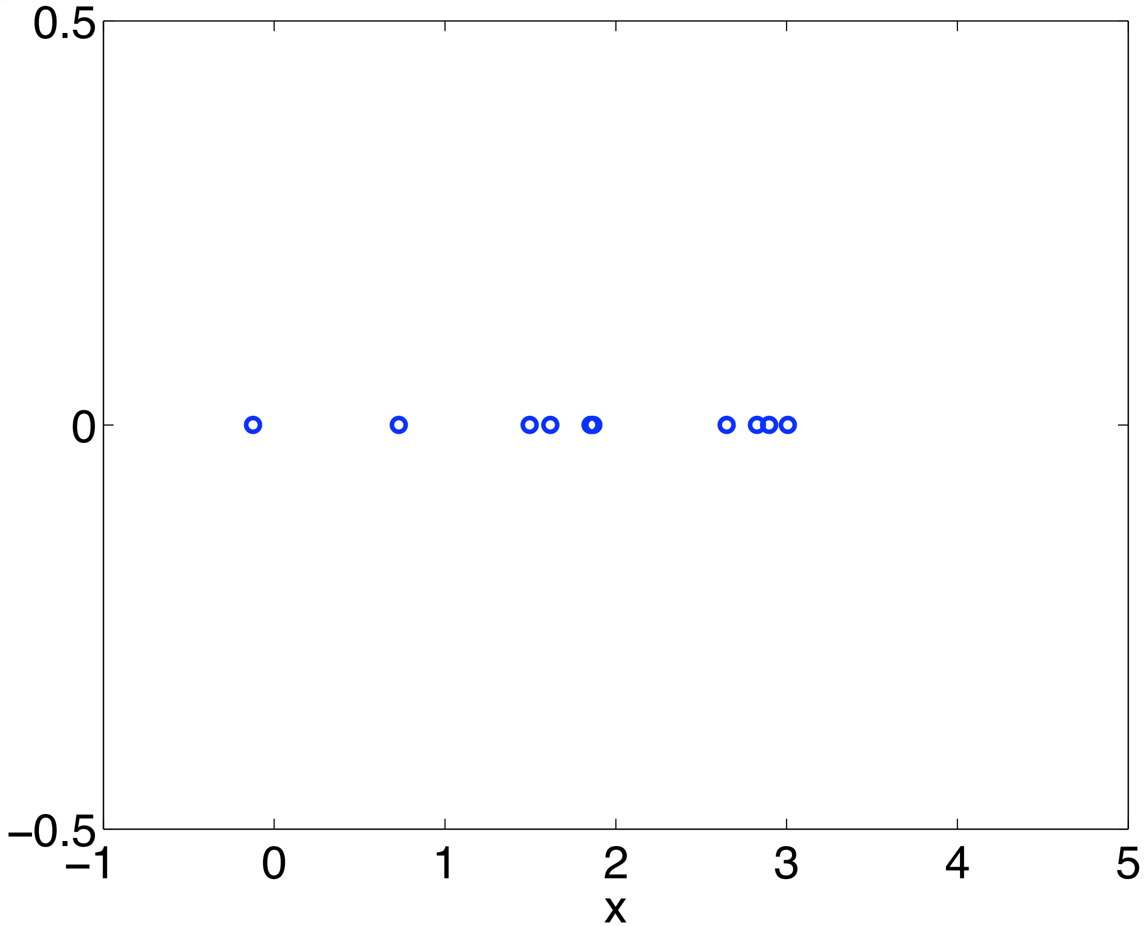
$\beta = 1/\sigma^2$: Precision



Statistics 101

Averages and variances

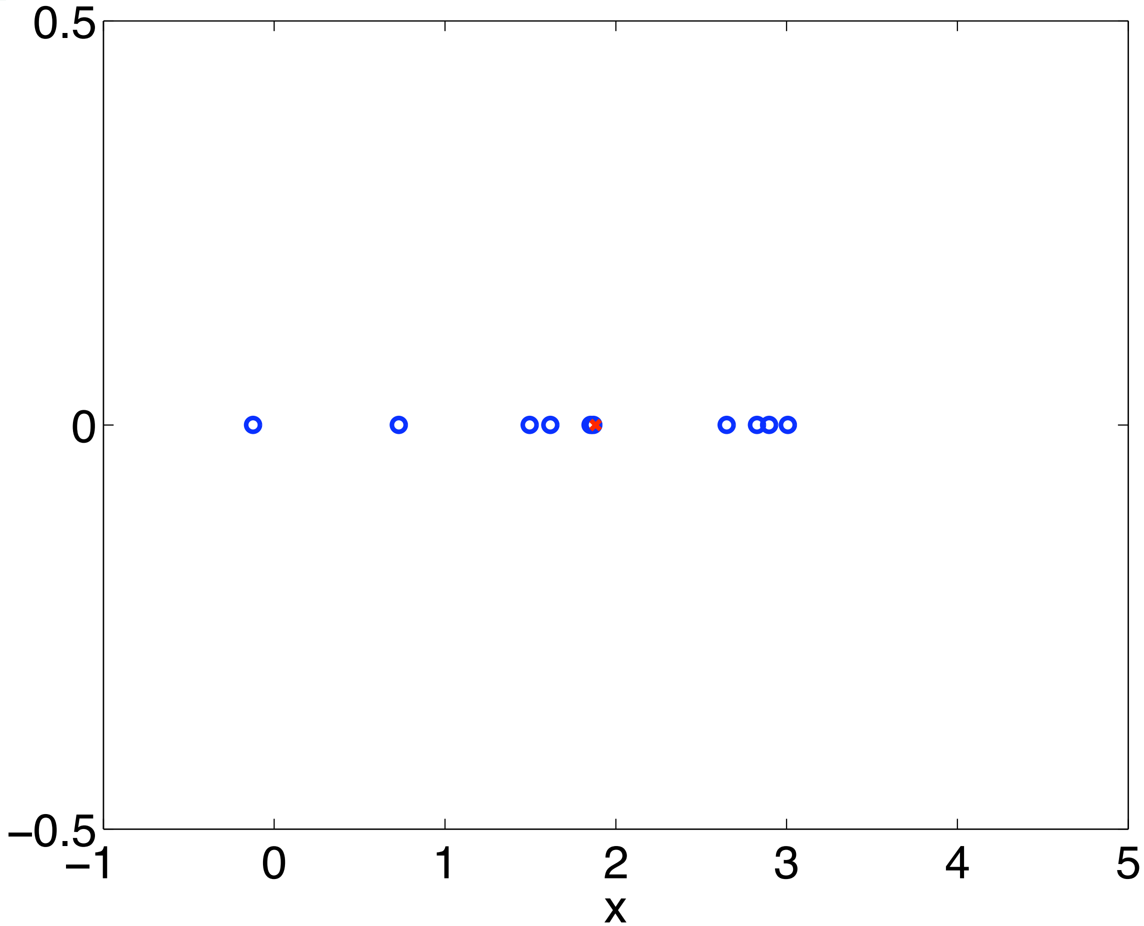
Here is a 1-D data set (*N*=10)



We can use statistics to describe the data set

Averages and variances

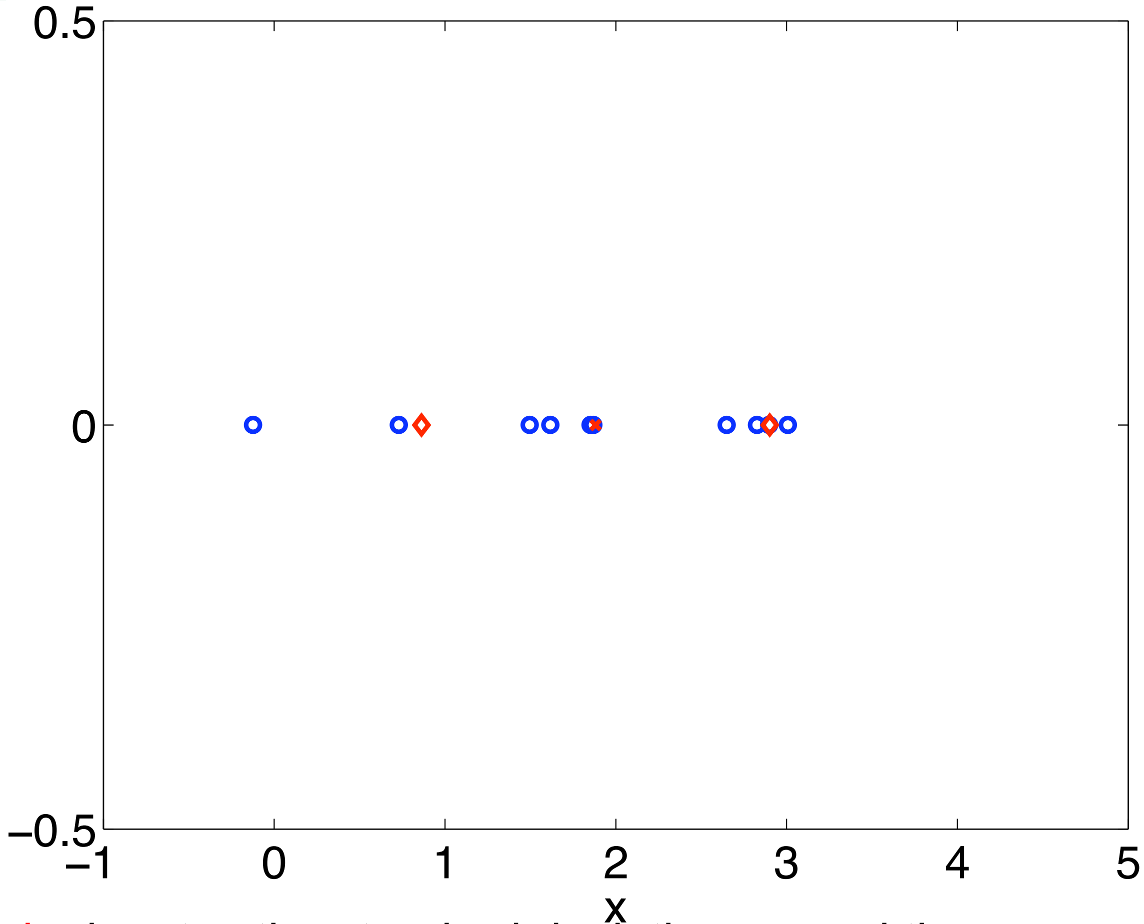
Describing the data by the central value



Red cross denotes the arithmetic mean (average): $\hat{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

Averages and variances

How much is data spread around the central value?

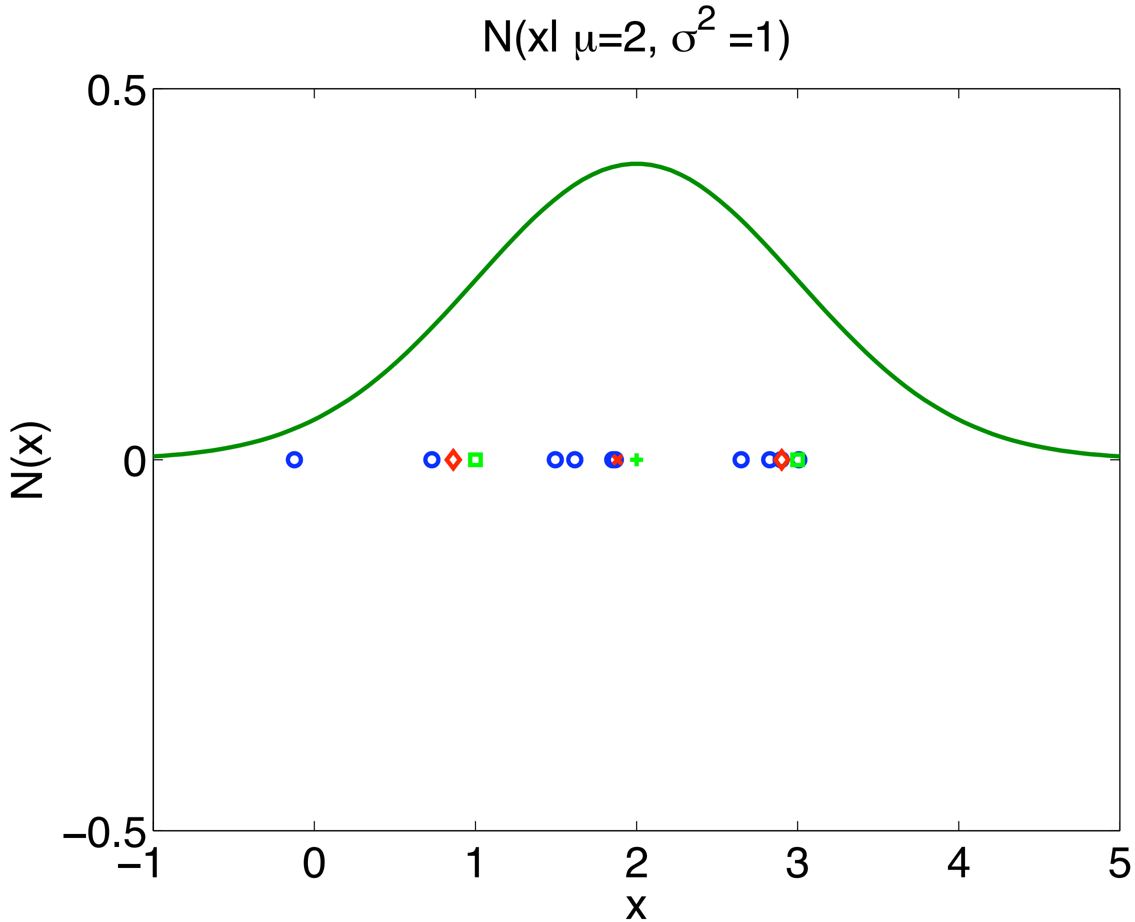


Red diamonds denotes the standard deviation around the mean: $\hat{x} \pm \mathrm{std}[x]$

Variance: $\mathrm{Var}[x] = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_n - \hat{x}\right)^2$ Standard deviation: $\mathrm{std}[x] = \sqrt{\mathrm{Var}[x]}$

Averages and variances

Data points was sampled from this Gaussian distribution



Green + denotes the mean parameter μ

Green squares denotes the standard deviation around the mean $\mu \pm \sigma$

Can be described by a Gaussian by learning the model parameters.

Sum and product rules

Common lazy notation (used when clear from context):

$p(X)$: Distribution of X

$p(x)$ or $p(X = x)$: Probability resp. density of value x

Discrete variables

Sum rule:

$$p(x_i) = p(X = x_i) = \sum_j p(x_i, y_j)$$

Product rule:

$$p(X,Y) = p(Y \mid X)\,p(X)$$
$$= p(X \mid Y)\,p(Y)$$

Continuous variables

Sum rule:

$$p(x) = \int p(x,y)\,dy$$

Product rule:

$$p(x,y) = p(y \mid x)\,p(x)$$
$$= p(x \mid y)\,p(y)$$

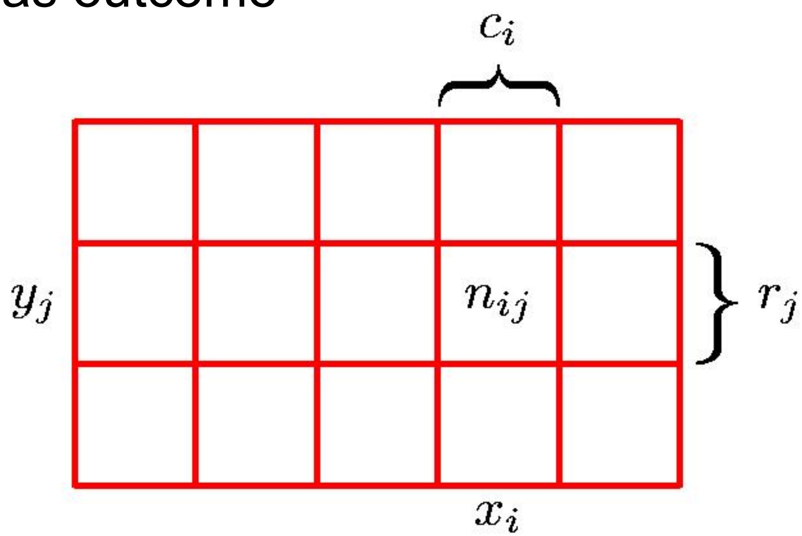
Returning to the Frequentists example

- Recall example with *N* trials:

n_{ij} Number of trials with $(X = x_i \, , \, Y = y_j)$ as outcome

$c_i = \sum_j n_{ij}$ number of $X = x_i$

$r_j = \sum_i n_{ij}$ number of $Y = y_j$



Joint probability:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability from the sum rule:

$$p(X = x_i) = \frac{c_i}{N} = \sum_j p(X = x_i, Y = y_j)$$

$$p(Y = y_j) = \frac{r_j}{N} = \sum_i p(X = x_i, Y = y_j)$$

Explanation for the conditional probability

- From the product rule we get:

$$p(Y \mid X) = \frac{p(X,Y)}{p(X)}$$

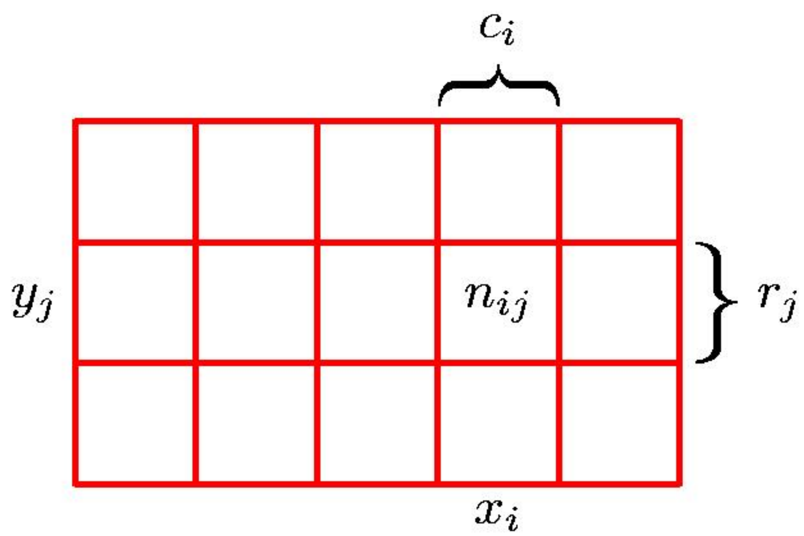
$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$$

- Recall example again:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$$p(Y = y_j) = \frac{r_j}{N} \qquad p(X = x_i \mid Y = y_j) = \frac{n_{ij}}{r_j}$$



Independence

- Two random variables *x* and *y* are said to be *independent* if their joint distribution factorizes into

$$p(x,y) = p(x)\,p(y)$$

- Independence leads to

$$p(y \mid x) = p(y)$$

- **Are-you-awake exercise:** Prove this by assuming independence and using the product rule.

Definition of expectation value

- Expectation value: Weighted average of functions *f(x)* of random variables

$$E[f(x)] = \sum_x f(x)p(x) \quad \text{(discrete)}$$

$$E[f(x)] = \int f(x)p(x)dx \quad \text{(continuous)}$$

- Conditional and marginal expectation:

$$E_x[f(x) \mid y] = \int f(x)p(x \mid y)dx$$

$$E_{x,y}[f(x,y)] = \int f(x,y)p(x,y)dxdy$$

Properties of Expectation

- Expectation has linear properties:

$$E[X+c]=E[X]+c$$

$$E[X+Y]=E[X]+E[Y]$$

$$E[cX]=cE[X]$$

where *c* is an arbitrary constant.

- Connection with statistics:
  In the limit of infinite many observations, we have

$$E[f(x)]=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(x_n)$$

Definition of variance and mean

- Mean of *x*: $E[x]$
- Variance of *x* (variability around the mean):

$$\mathrm{var}[x] = E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

- Covariance of *x* and *y*:

$$\mathrm{cov}[x,y] = E_{x,y}[(x - E[x])(y - E[y])] = E_{x,y}[xy] - E[x]E[y]$$

If *x* and *y* are independent then $\mathrm{cov}[x,y] = 0$

The Gaussian (a.k.a. Normal) Distribution

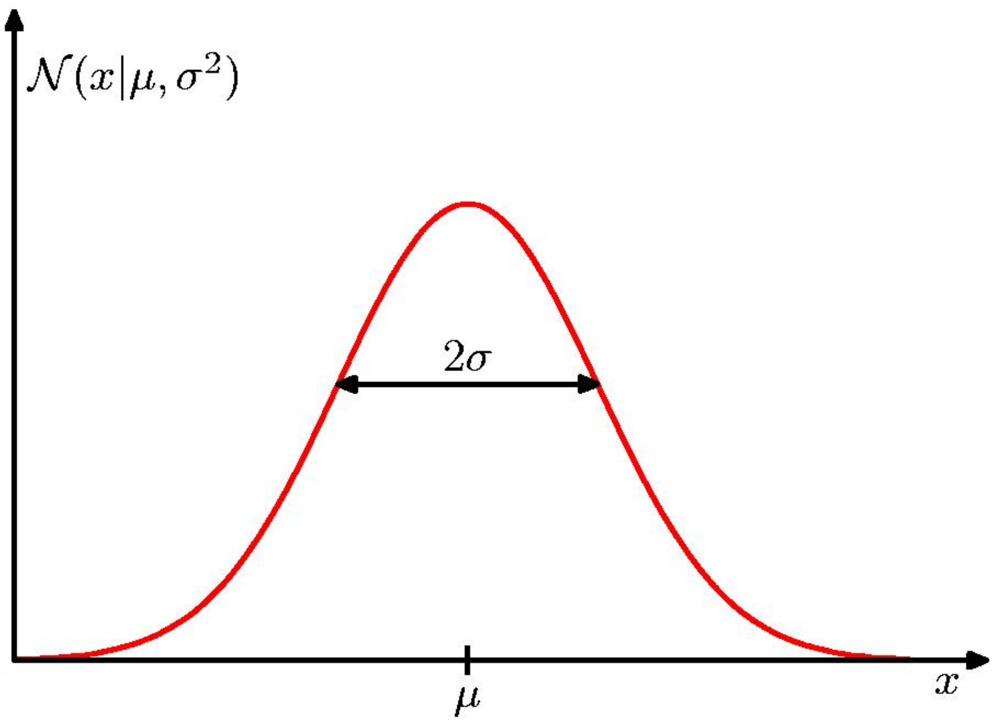
The 1-dimensional Gaussian probability density:

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Fulfills the density conditions:

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$



The expectation of Gaussian variables

$$\text{Mean} \ : \ E[x] = \int_{-\infty}^{\infty} x \mathcal{N}(x \mid \mu, \sigma^2) dx = \mu$$

$$\text{2nd moment} \ : \ E[x^2] = \int_{-\infty}^{\infty} x^2 \mathcal{N}(x \mid \mu, \sigma^2) dx = \mu^2 + \sigma^2$$

$$\text{Variance} \ : \ \text{var}[x] = E\left[x^2\right] - \left(E\left[x\right]\right)^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

σ : Standard deviation

$\beta = 1/\sigma^2$: Precision

Summary

- Goal of ML:
  - Model and learn the mapping between data x and some description y(x) (on training data)
  - Generalization to unseen data (on test data)
- Machine learning: Model and learn the mapping y(x)
- Pattern recognition: Find patterns in data using y(x)
- Common problems:
  - Regression: y(x) is a continuous quantity
  - Classification: y(x) is a discrete quantity (class labels)
- Supervised vs. unsupervised learning
- Probability theory and statistics 101

And there is much more to come on all these points!

Literature

- Introductory material: Bishop Sec. 1.-1.1, 1.2-1.2.4
- Additional material: Read about Matlab, R, Python, Shark, ...
  Also see the math reviews under the Additional course material menu item in Absalon.