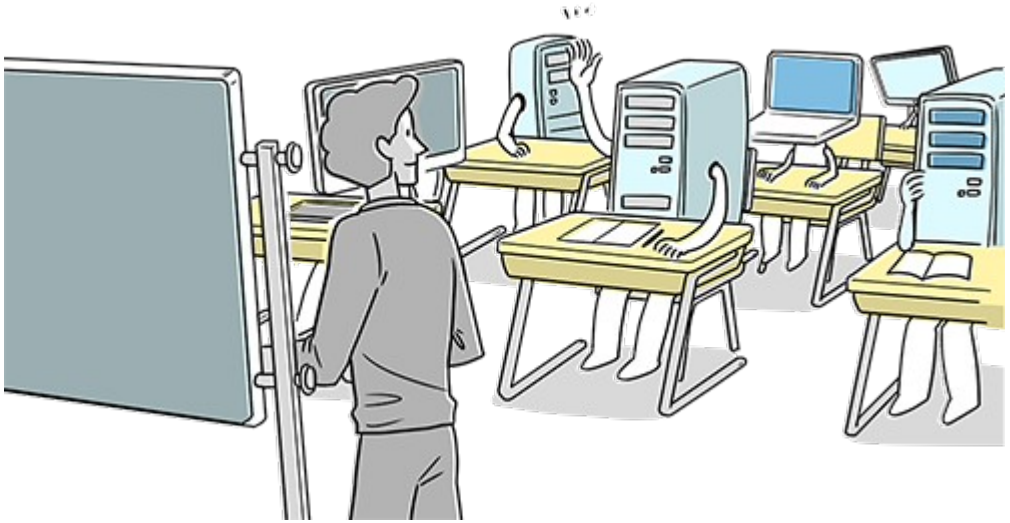
Machine learning
Lecture 1



Marcin Wolter

IFJ PAN

29 November 2019

- Machine learning: what does it mean?
- Software to work with and literature.
- A little bit of mathematics and examples of simple linear classifiers.
- Some examples

All slides will be here: https://indico.ifj.edu.pl/event/289/

Outline of the course

- Introduction: introduction to statistics, what does "Machine Learning" mean? A little bit of mathematics, but also examples of simple linear classifiers.
- Simple non-linear methods like Naive Bayes, k-Nearest Neighbors, Probability Density Estimators and Boosted Decision Trees (BDT).
- Neural Networks and Bayesian Neural Networks.
- Cross-validation and optimization of machine learning algorithms.
- Introduction to Deep Learning.
- Convolution Deep Neural network – classification of images.
- Some regression problems with Deep Neural Networks.
- Generative deep networks – Generative Adversary Networks (GANs)

Recommended books

- M. Krzyśko, Systemy uczące się: rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości. WNT, 2008.
- C. Bishop, Pattern recognition and machine learning. Springer, 2009.
- Internet information about Deep Neural Networks – many things have changed in the recent years.

Programs

- Python programming language

- http://scikit-learn.org
  scikit-learn - Machine Learning in Python
  Simple and efficient tools for data mining and data analysis
  Accessible to everybody and reusable in various contexts
  Built on NumPy, SciPy, and matplotlib

- https://keras.io/
  Keras: The Python Deep Learning library
  Emulates Deep Neural Network, uses google TensorFlow software

- **Google Colaboratory** as a working platform – you need a google ID
  https://colab.research.google.com/
  Platform to run python notebooks, gives free GPU

About the course

- We will have 20 hours of lectures in total.
- I propose to do some exercises, run some python programs on *google colaboratory*. So please bring your laptops!

- We should make some type of exam. I propose, that instead of exam you write small projects with machine learning. This gives a chance to learn how to use machine learning tools.

Statistics

- **Statistics – describes random events.**
- First works – أبو يوسف يعقوب بن إسحاق الصبّاح الكندي Al-Kindi (801-873) used statistical methods to break the Ceasar cipher by investigating the frequency in which particular letters appear.

Caesar cipher

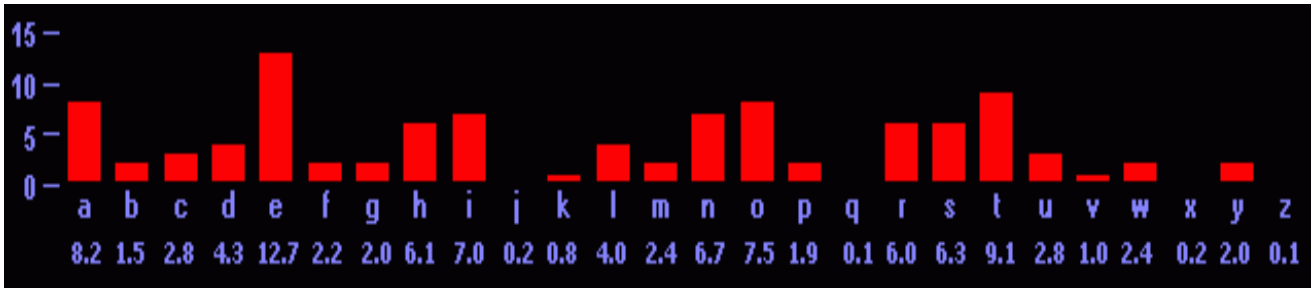
- Each letter of a text is replaced by another, shifted by n letters:

A	B	C	D	E
C	D	E	F	G

- In general Al-Kindi's method can be used to break any replacement cipher (each letter is replaced by another letter, always the same)

A	B	C	D	E
Z	D	P	G	T

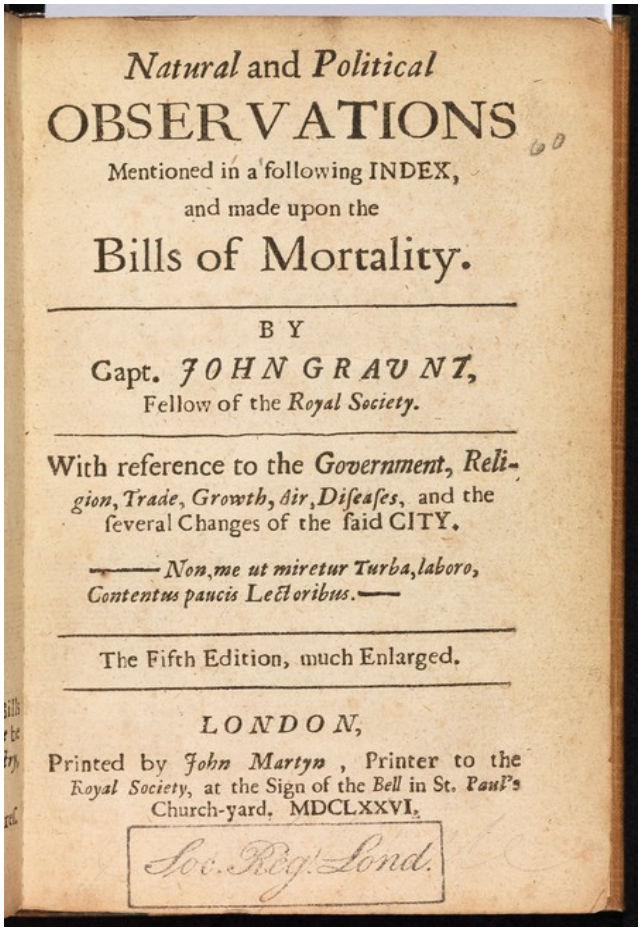
- Frequency analysis – the frequency of appearance of different letters is investigated.



- Frequency of different letters in English.

Statistics

- Important step in the development of statistics were the first studies of demography and the games of chance (1663 John Graunt „*Natural and Political Observations upon the Bills of Mortality*”).



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

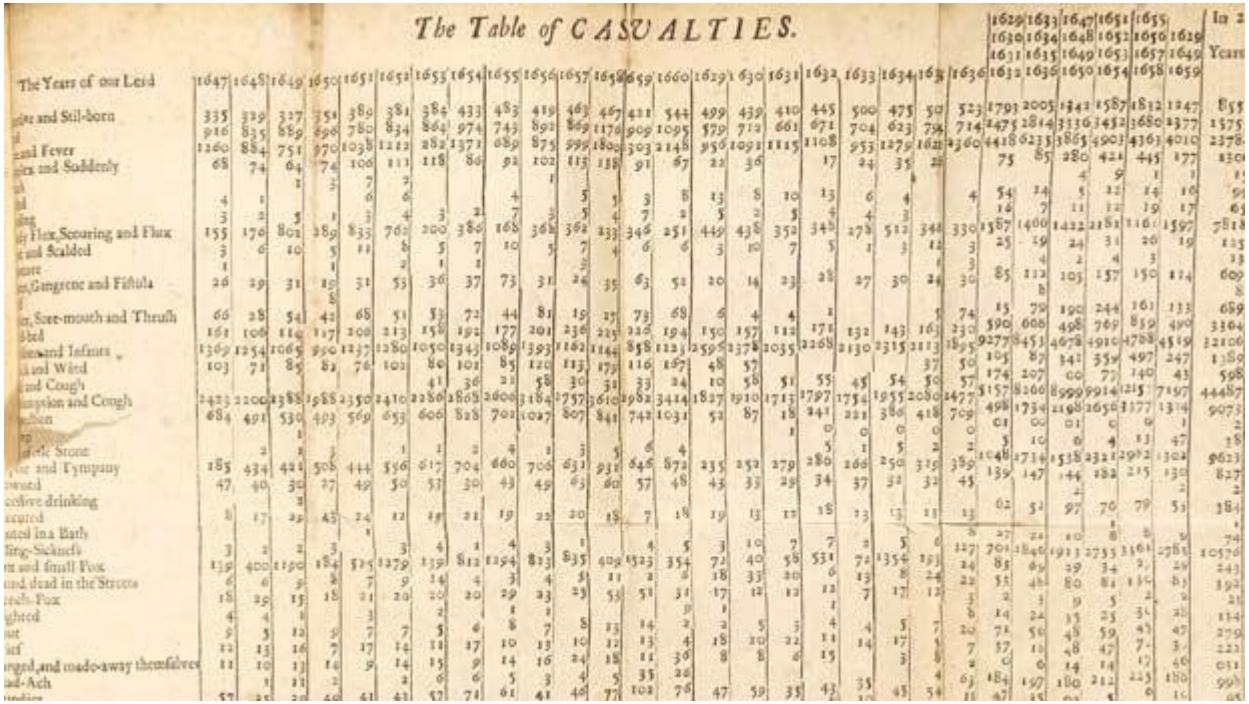
Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,
Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXXVI.



The Table of CASUALTIES.

What is an “event”?

- Elementary event is each result of an experiment (like throwing a dice), which result is random. It contains only a single outcome in the sample space.

The numerical value of a probability may be obtained from its "classical" definition: The probability is equal to the quotient of the number of cases "favouring" a certain event to the total number of "equally possible" cases. (Laplace 1812).

Pierre-Simon Laplace (1749–1827)

- Let’s denote a set of all possible events as Ω. Elements of a set Ω are elementary events ω, so Ω is a set of elementary events. The set of events favoring A is a subset of Ω and:

$$P(A)=\frac{|A|}{|\Omega|}$$

where |A| is a number of elements of a set A, and |Ω| a number of elements of a set Ω.

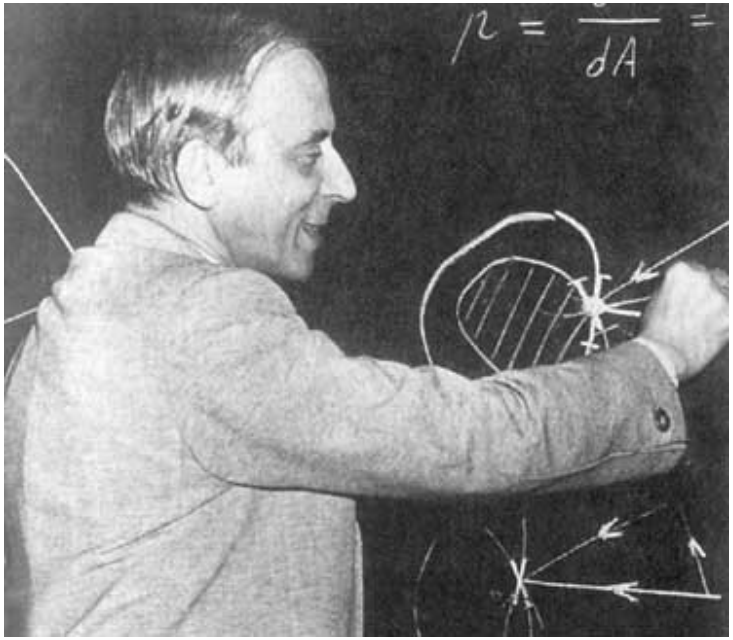
Example: probability of getting 6 while throwing a dice. Set of elementary events Ω = {1, 2, 3, 4, 5, 6}, so the number of elementary events is |Ω| = 6. Set of events favoring A = {6}, their number is |A| = 1. So P(6)=1/6

Frequentist definition of probability

- **Probability (frequentist definition) of an event A is a limit (N approaching infinity) of the ratio of n events when the event A occurred to the total number of trials N:**

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

What is a probability of getting “6” while throwing a dice? It is the relative ratio of obtaining “6” in an infinite series of throws.

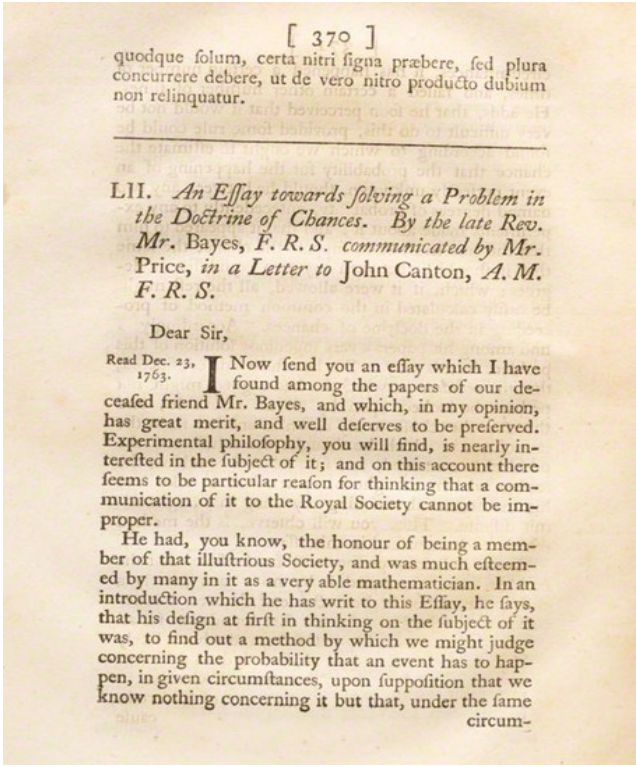


- The definition comes from Richard von Mises (born 19 of April 1883 in Lwów, died 14 July 1953 in Boston) – mathematician, brother of the economist Ludwig von Mises.

This idea of frequentist probability was used earlier, for example by de’Moivre, Bernoulli, Gauss ...

Bayesian definition

- Probability “a priori”, called unconditional, is a measure of belief, based on rational premises, that a given event will occur.
- In the next step we perform an experiment, called observation, and their results allow to modify the probability. We get the probability “a posteriori”, which is again a measure of belief, but modified by the observation.
- Supporters of the Bayesian approach were P. S. Laplace, H. Poincare and the economist John Keynes, arguing, that this is a method we use analyzing the world around us.



[370]

quodque ſolum, certa nitri ſigna præbere, ſed plura concurrere debere, ut de vero nitro producto dubium non relinquatur.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

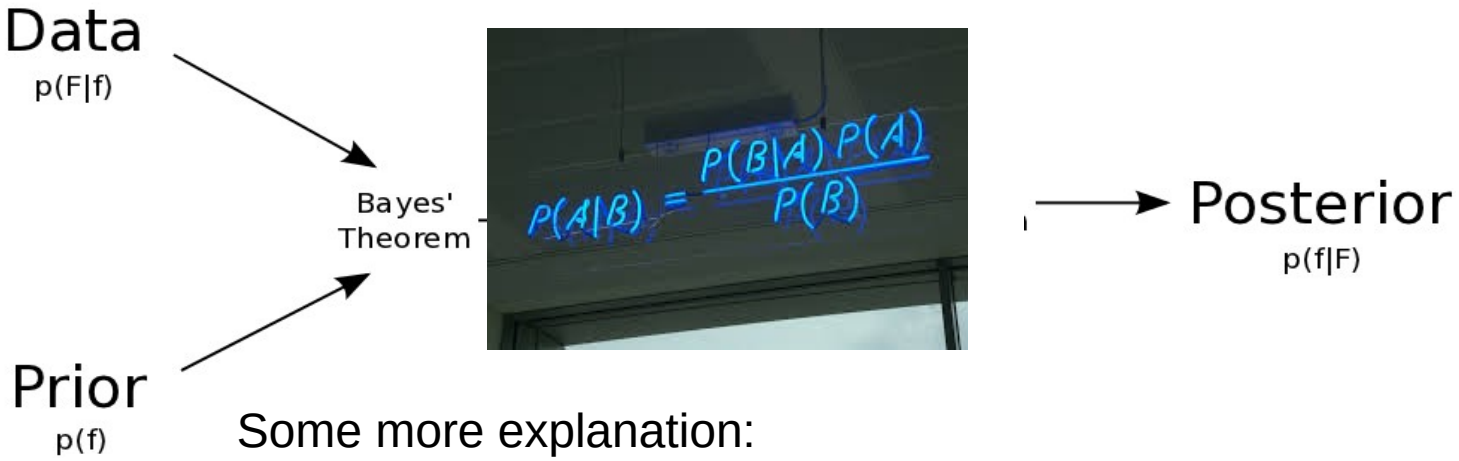
Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Thomas Bayes (1702 - 1761) was an English statistician, philosopher and Presbyterian minister.The most important work: „Essay Towards Solving a Problem in the Doctrine of Chances”.

Bayesian probability

- Experiment we can't repeat many times: what is a probability to pass an exam?
- Based on our knowledge we estimate: ½ (probability *a priori*).
- But if all people before us didn't pass an exam (set of experiments) and we know our knowledge is not significantly higher we should verify this estimate (probability *a posteriori).*



Some more explanation:

https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348

Bayes Theorem

- Bayes' theorem relates the conditional (posterior) and marginal (prior) probabilities of events A and B:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- **P(A)** is the prior probability or marginal probability of A. It is a "prior" in the sense that it does not take into account any information about B.
- **P(A|B)** is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

- **Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.**

Bayes Theorem – an example: a cancer test

$$\Pr(\mathrm{A}|\mathrm{X}) = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X})} = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A}) + \Pr(\mathrm{X}|\text{not A})\Pr(\text{not A})}$$

- Pr(A|X) = Chance of having cancer (A) given a positive test (X). This is what we want to know: How likely is it to have cancer with a positive result? .
- Pr(X|A) = Chance of a positive test (X) given that you had cancer (A). This is the chance of a true positive, let say 80% in our case.
- Pr(A) = Chance of having cancer (1%).
- Pr(not A) = Chance of not having cancer (99%).
- Pr(X|not A) = Chance of a positive test (X) given that you didn't have cancer (not A). This is a false positive, 9.6% in our case.
- **In our case Pr(A|X) is 7.8%**

Bayesian vs. Frequentist approach

- **PROBABILITY: degree of belief** (Bayes, Laplace, Gauss, Jeffreys, de Finetti)
- **PROBABILITY: relative frequency** (Venn, Fisher, Neyman, von Mises).
- **Bayesian approach:** probability is degree of belief. Thus the probability *p* is our assessment of the probability of success at each trial, based on our current state of knowledge.

  If our assessment, initially, is incorrect? As our state of knowledge changes, our assessment of the probability of success changes accordingly.
- **Bayesian inference** is statistical inference in which **evidence or observations are used** to update or to newly infer the probability that a hypothesis may be true.
- This allows for a *cleaner* foundation than the frequentist interpretation.

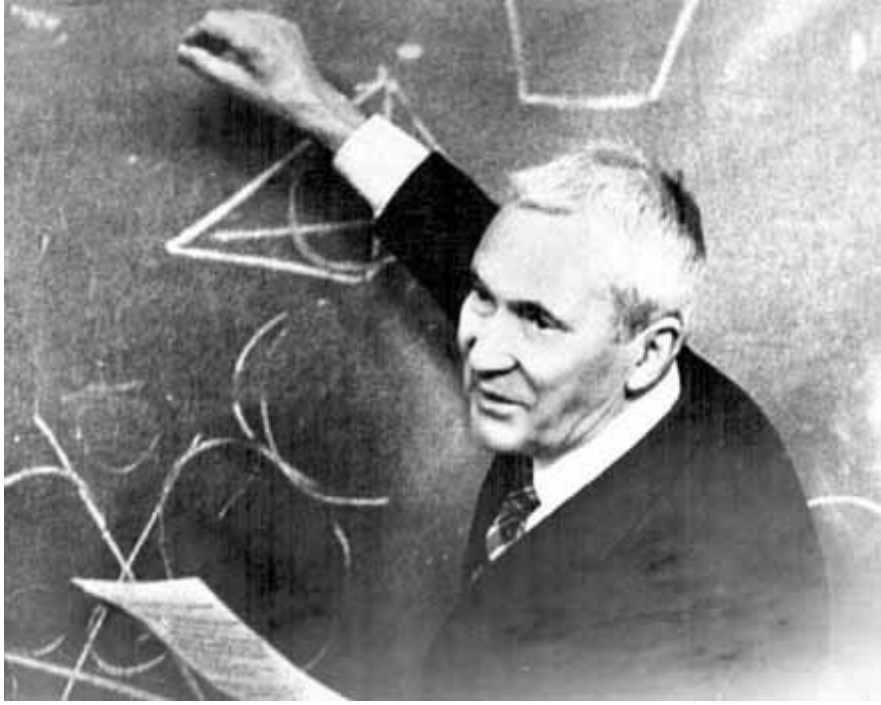
"*We don't know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some a priori postulates. My problem is to get these stated as clearly as possible.*"

Sir Harold Jeffreys, in a letter to Sir Ronald Fisher dated 1 March, 1934

H.B. Prosper,"Bayesian Analysis", arXiv:hep-ph/0006356v1 30 Jun 2000

Axiomatic definition

Probability can be defined in many ways...



Андре́й Никола́евич Колмого́ров (1903-1987)

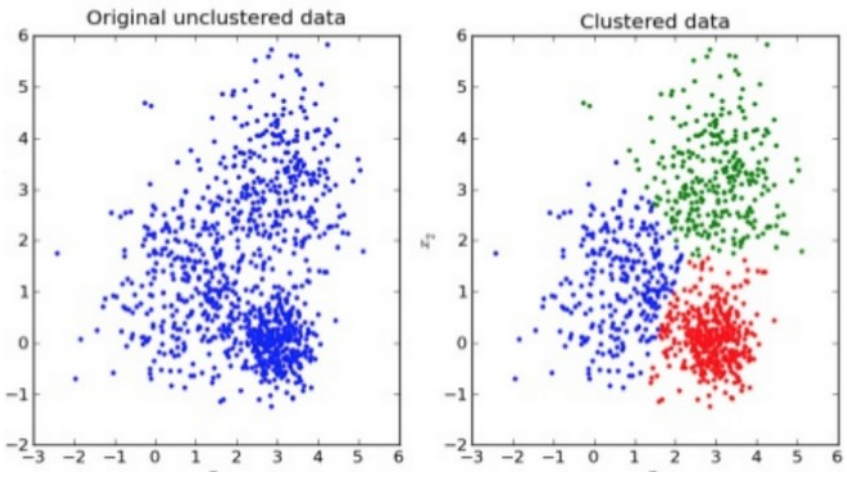
Kolmogorov axiomatic definition:

Let Q_i denote anything subject to weighting by a normalized linear scheme of weights that sum to unity in a set W. The Kolmogorov axioms state that:

1. For every Q_i in W, there is a real number $Q(Q_i)$ (the Kolmogorov weight of Q_i) such that $0<Q(Q_i)<1$.
2. $Q(Q_i)+Q(Q'_i)=1$, where Q'_i denotes the complement of Q_i in W.
3. For the mutually exclusive subsets Q_1, Q_2, ... in W, $Q(Q_1 \cup Q_2 \cup Q_3 \cup ...)=Q(Q_1)+Q(Q_2)+Q(Q_3)+....$

What does “machine learning” mean?

- **Machine learning** is a field of computer science that gives computer systems the ability to "learn" (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed.
- Problems:
  - Supervised learning (classification & regression)
  - Clustering (unsupervised learning)
  - Dimensionality reduction
  - Reinforcement learning
  - Many others….



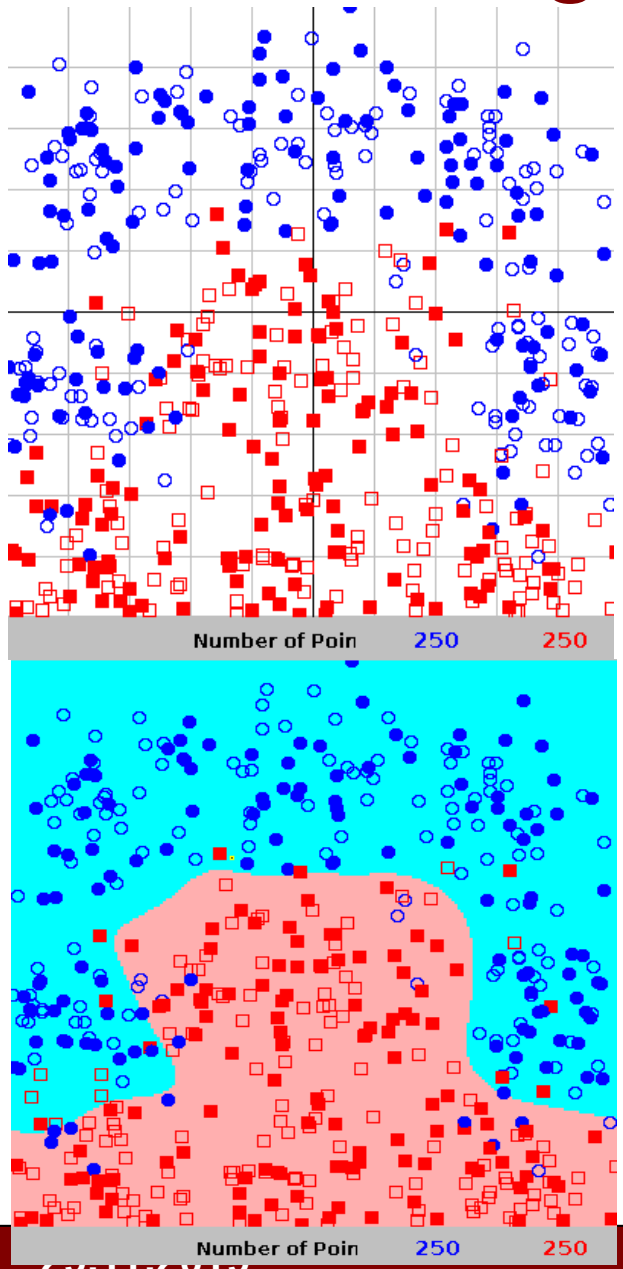
➢Unsupervised Learning

❑Technique of trying to find hidden structure in unlabeled data

➢Supervise Learning

❑Technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs.

How do the (supervised) machine learning algorithms work?

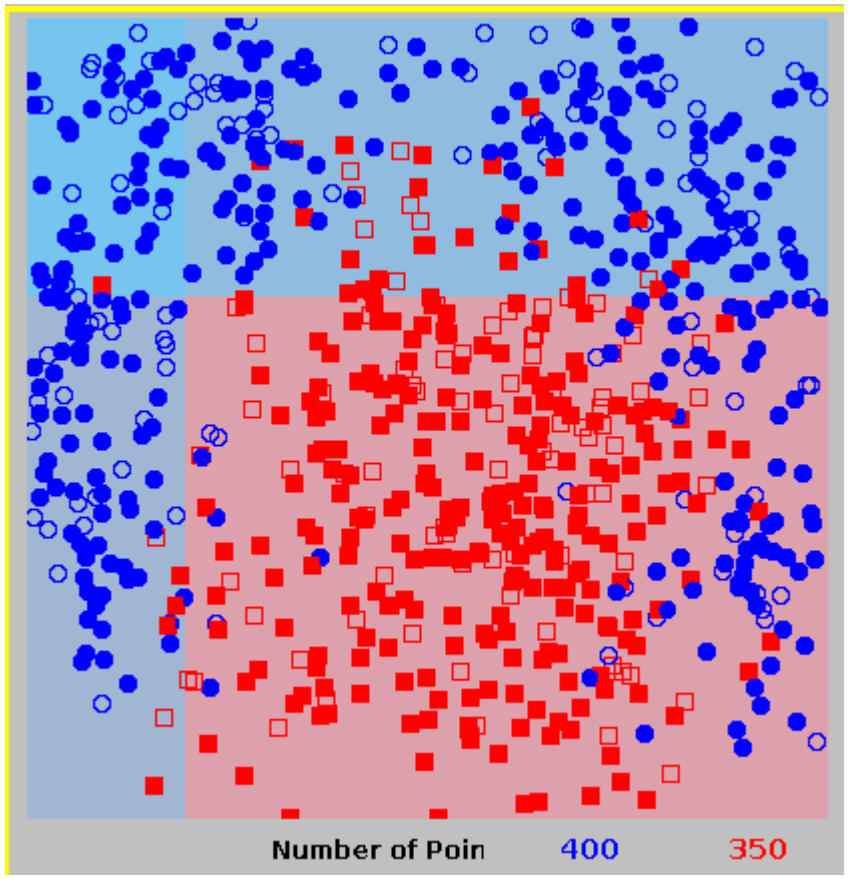


- We need **training data**, for which we know the correct answer, whether it's a signa or background. We divide the data into two samples: training and test.
- We find the best function ***f(x)*** which describes the probability, that a given event belongs to the class "signal". This is done by minimizing the loss function (for example χ^2).
- Different algorithms differ by: the class of function used as ***f(x)*** (linear, non-linear etc), loss function and the way it's minimized.
- All these algorithms try to approximate the unknown *Bayessian Decisive Function* (BDF) relying on the finit training sample.

BDF -an ideal classification function given by the unknown probability densities of signal and background.

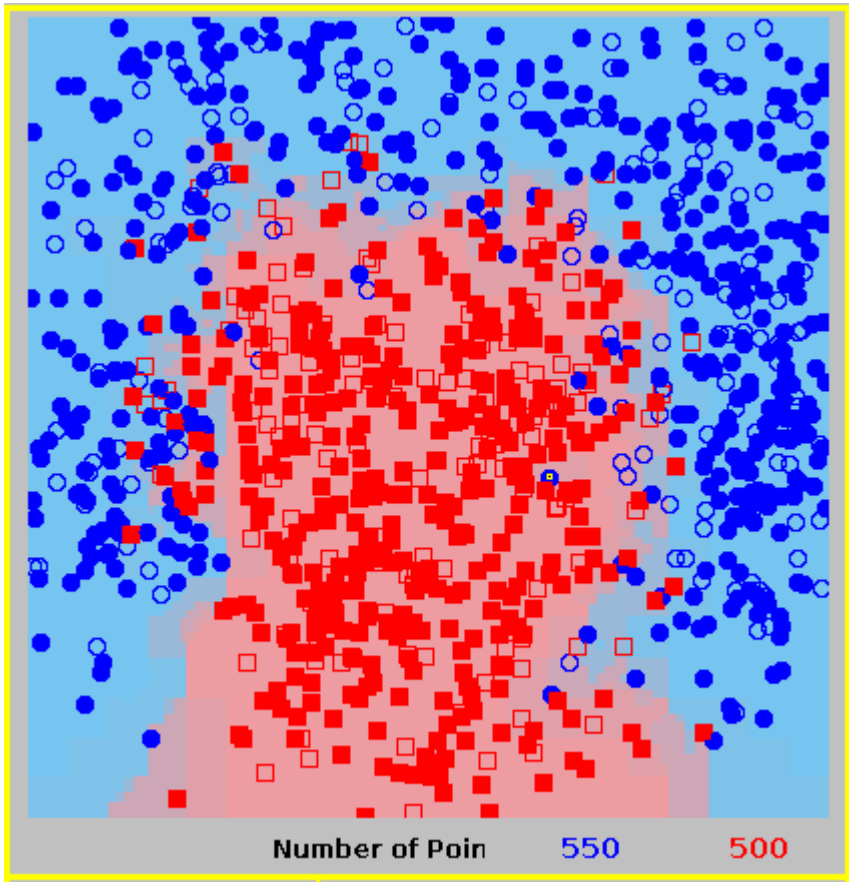
Cuts vs non-linear separation

Cuts



?

Non-linear separation



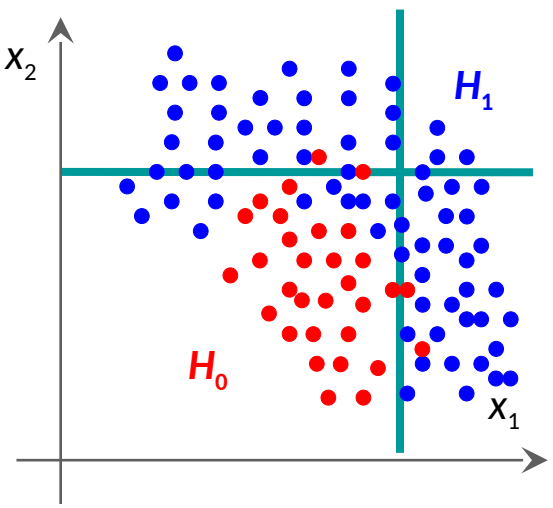
Neural Networks, boosted decision trees, and so on….

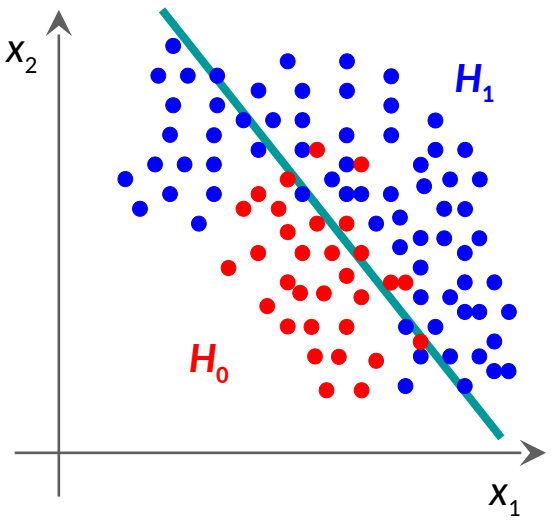
Types of algorithms

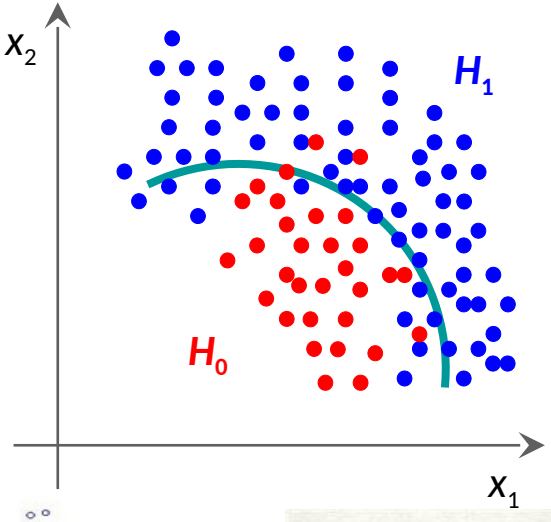
How to use the information available

Classification: find a function ***f(x1,x2)*** giving the probability, that a given data point belongs to a given class (signal vs background).

Simple cuts (easy and intuitive)	Linear (fast and stable)	Non-linear (most effective)

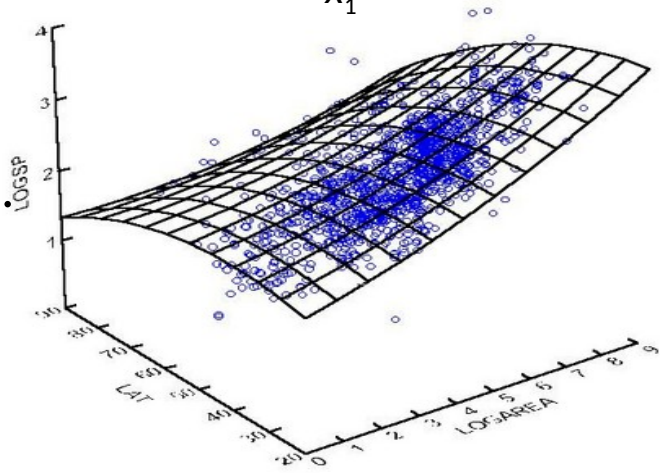


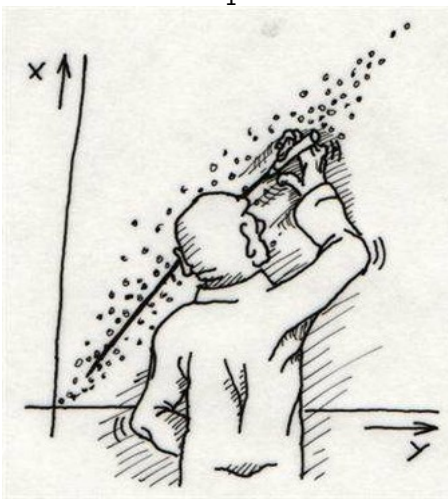




Regression: fit a continuous function

(find particle energy from calo readouts).





Classification

A Bayes classifier (optimal classifier):

$$p(S|x) = \frac{p(x|S)\,p(S)}{p(x|S)\,p(S) + p(x|B)\,p(B)}$$

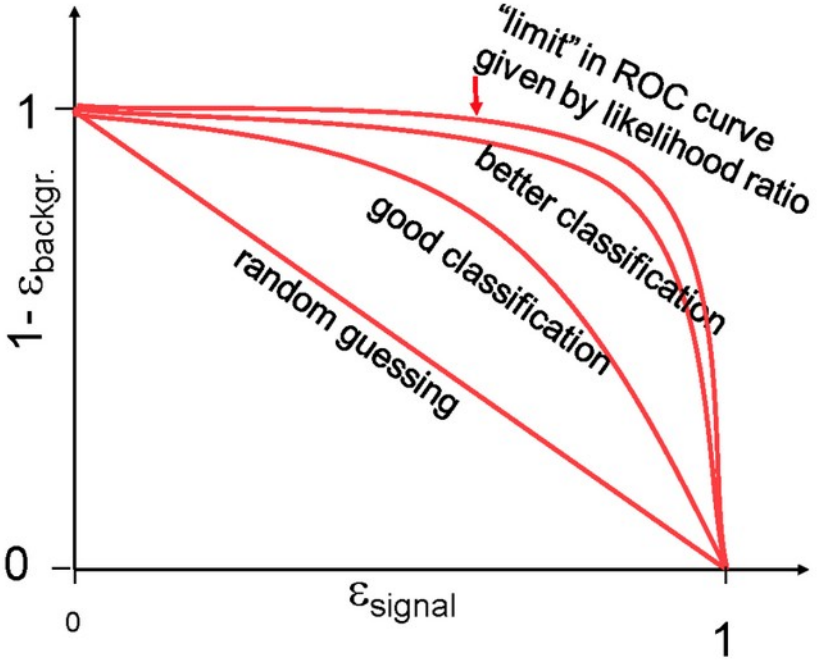
where ***S*** is associated with *y* = **1** and ***B*** with *y* = **0**. **Bayes classifier** *accepts events x if p(**S**|x) > **cut** as belonging to **S**.*

We need to approximate probability distributions *P(**x**|**S**)* and *P(**x**|**B**).*

- *If your goal is to **classify objects** with the fewest errors, then the **Bayes classifier** is the **optimal** solution.*
- Consequently, if you have a classifier known to be ***close*** to the **Bayes limit**, then *any* other classifier, *however sophisticated*, can ***at best*** be only marginally better than the one you have.
  - =>If your problem is **linear** you don't gain anything by using sophisticated **Neural Network**
- *All* classification methods, such as all we will be talking about, are different numerical approximations of the Bayes classifier.

ROC curve

- ROC (Receiver Operation Characteristic) curve was first used to calibrate radars.
- Shows the background rejection ($1-\varepsilon_B$) vs signal efficiency ε_B. Shows how good the classifier is.
- The integral of ROC could be a measure of the classifier quality:



Integral(ROC) = ½ – random

Integral(ROC) = 1 - ideal

Practical applications

A Short List of Multivariate Methods

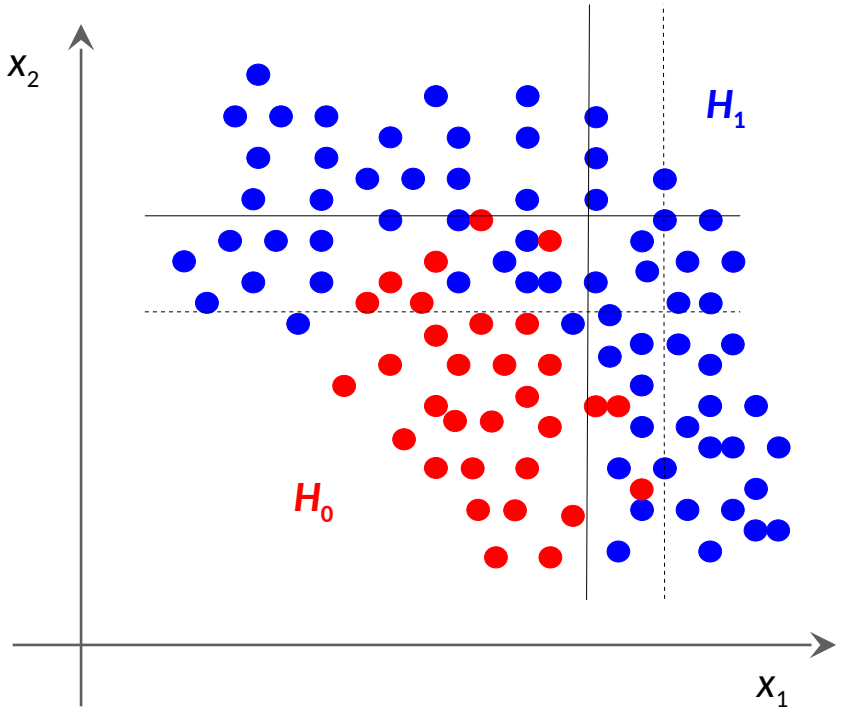
- Cuts
- Linear Discriminants (like Fisher)
- Naive Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- Decision Trees
- Neural Networks
- Bayesian Neural Networks
- Genetic Algorithms
- Support Vector Machines

- And many, many others..... I want to present briefly just few of them.

We will talk in the first lectures about:

- Simple ML linear methods:
  - Cuts
  - Fisher linear discriminant
  - Naive Bayes
  - Principal Component Analysis, PCA
  - Independent Component Analysis, ICA

Cuts



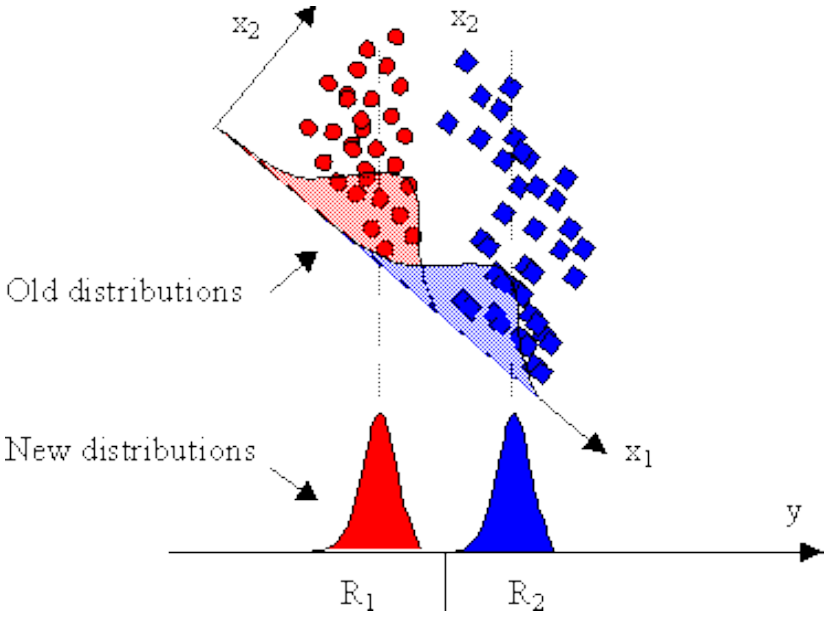
Optimization of cuts:

- Move cuts as long as we get the optimal signal vs. background selection. For a given signal efficiency we find the best background rejection → we get the entire ROC curve.
- Optimization methods:
  - Brute force
  - Genetic algorithms
  - Many others...

Fisher discriminants

LDA, Linear Discriminat Analysis

Projection to one dimension, than discrimination



Equivalent to linear separation

We choose a projection vector in such a way, that the separation is maximized.

Assumptions for new basis:

- Maximize distance between projected class means
- Minimize projected class variance

Method introduced by Fisher in 1936.
Optimal separation for Gaussian distributions.

Fisher Linear DiscriminantAnalysis

Objective

$$argmax_w J(w) = \frac{w^T S_B w}{w^T S_W w}$$

$$m_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$ Variance Between classes

$$S_W = \sum_j^2 \sum_{x \in C_j} (x - m_j)(x - m_j)^T$$ Variance Within class

Algorithm

1. Compute class means
2. Compute $w = S_W^{-1}(m_2 - m_1)$
3. Project data $y = w^T x$

Fisher's linear discriminant

The terms *Fisher's linear discriminant* and *LDA* are often used interchangeably, although Fisher's original article *The Use of Multiple Measures in Taxonomic Problems* (1936) actually describes a slightly different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances.

Suppose two classes of observations have means $\vec{\mu}_{y=0}, \vec{\mu}_{y=1}$ and covariances $\Sigma_{y=0}, \Sigma_{y=1}$. Then the linear combination of features $\vec{w} \cdot \vec{x}$ will have means $\vec{w} \cdot \vec{\mu}_{y=i}$ and variances $\vec{w}^T \Sigma_{y=i} \vec{w}$ for $i = 0,1$. Fisher defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\vec{w} \cdot \vec{\mu}_{y=1} - \vec{w} \cdot \vec{\mu}_{y=0})^2}{\vec{w}^T \Sigma_{y=1} \vec{w} + \vec{w}^T \Sigma_{y=0} \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_{y=1} - \vec{\mu}_{y=0}))^2}{\vec{w}^T (\Sigma_{y=0} + \Sigma_{y=1}) \vec{w}}$$

This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling. It can be shown that the maximum separation occurs when

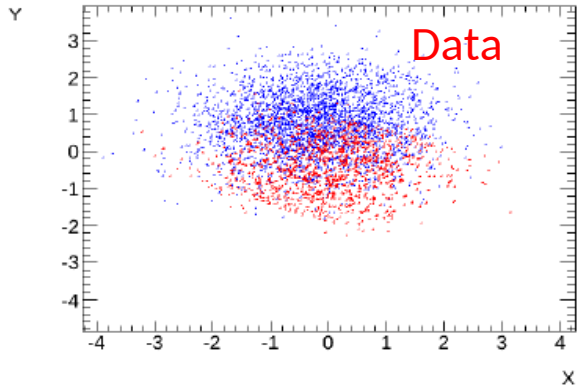
$$\vec{w} = (\Sigma_{y=0} + \Sigma_{y=1})^{-1} (\vec{\mu}_{y=1} - \vec{\mu}_{y=0})$$

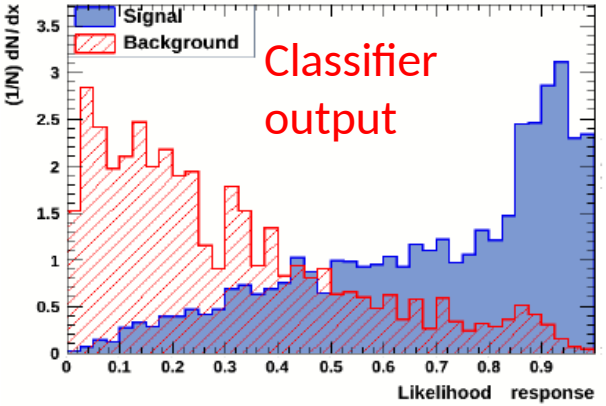
When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

Be sure to note that the vector $\vec{w}$ is the normal to the discriminant hyperplane. As an example, in a two dimensional problem, the line that best divides the two groups is perpendicular to $\vec{w}$.

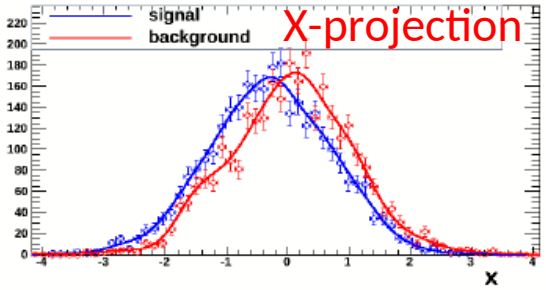
Generally, the data points are projected onto $\vec{w}$. However, to find the actual plane that best separates the data, one must solve for the bias term b in $w^T \mu_1 + b = -(w^T \mu_2 + b)$.

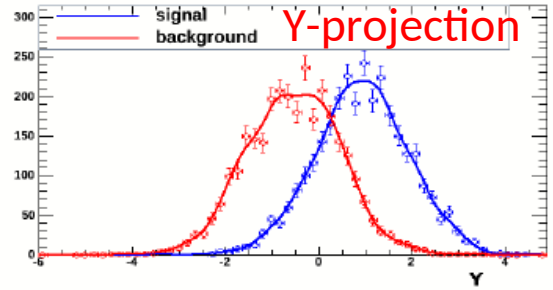
Naive Bayes classifier





Frequently called **“projected likelihood" (TMVA).**





- Based on the assumption, that variables are independent (so „naive"):

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

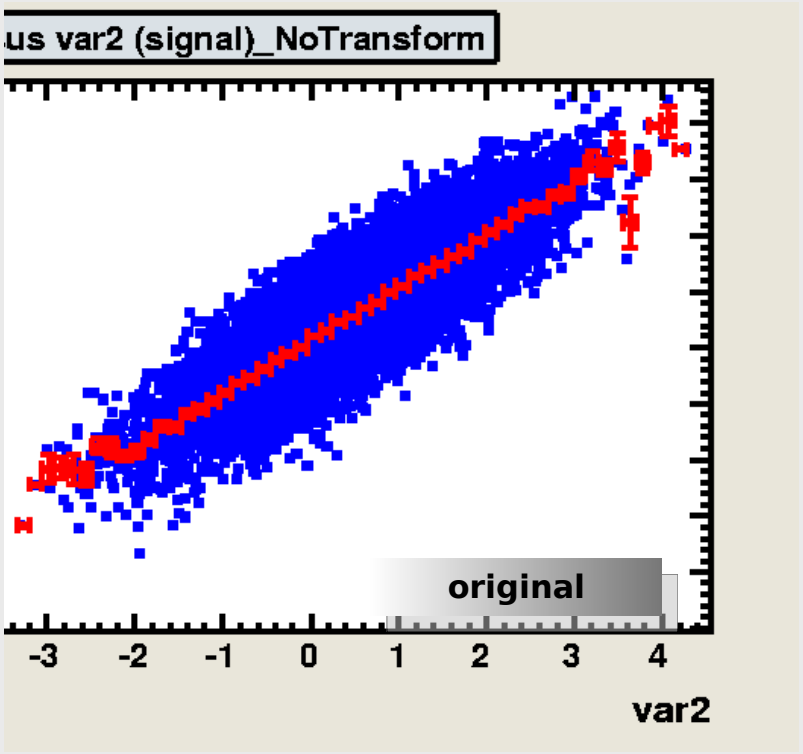
“Naive” assumption: $P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$

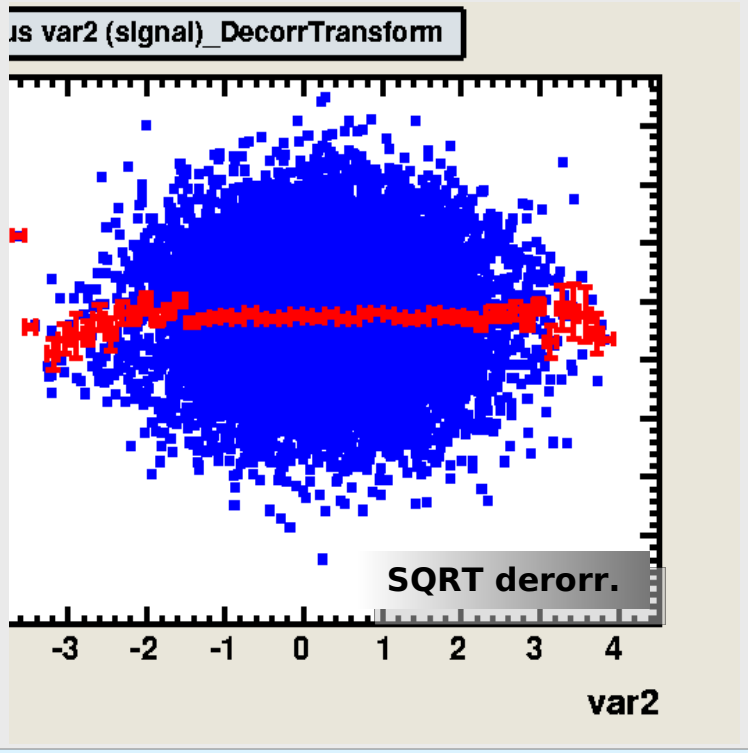
$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

- Output probability is **a product of probabilities for all variables.**
- Fast and stable, not optimal, but in many cases sufficient.

Decorrelation

- **Removes correlation between variables by a rotation in the space of variables**



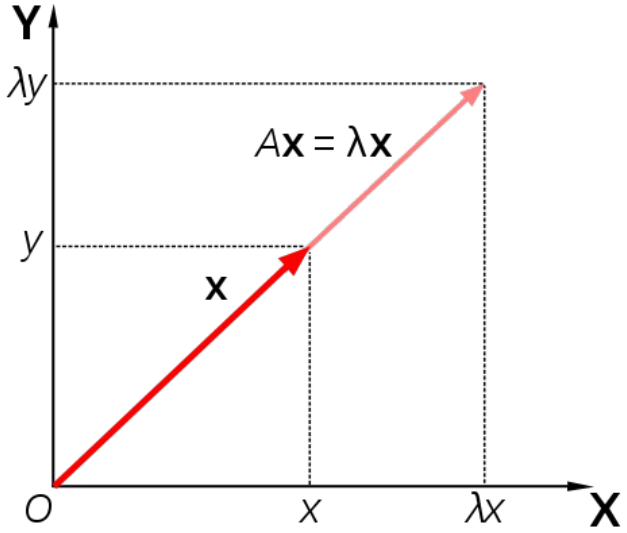


Eigenvalues and eigenvectors

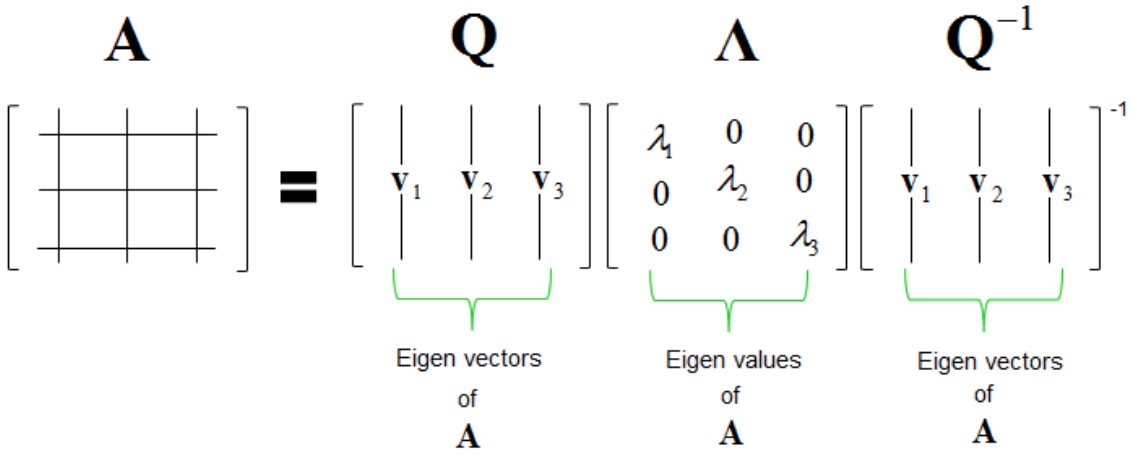
In essence, an eigenvector v of a linear transformation T is a non-zero vector that, when T is applied to it, does not change direction. Applying T to the eigenvector only scales the eigenvector by the scalar value λ, called an eigenvalue. This condition can be written as the equation

$$T(\mathbf{v}) = \lambda \mathbf{v}$$

referred to as the eigenvalue equation or eigenequation. In general, λ may be any scalar. For example, λ may be negative, in which case the eigenvector reverses direction as part of the scaling, or it may be zero or complex.



Matrix A acts by stretching the vector x, not changing its direction, so x is an eigenvector of A.



Eigendecomposition of a matrix

Covariance Matrix

- Let X be a p-variate random vector. The covariance matrix of X is defined as:

$$\Sigma_{XX} = Var(X) = E\{(X-\mu)^T(X-\mu)\}$$

$$= \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_p) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p,X_1) & Cov(X_p,X_2) & \dots & Var(X_p) \end{pmatrix}.$$

Where:

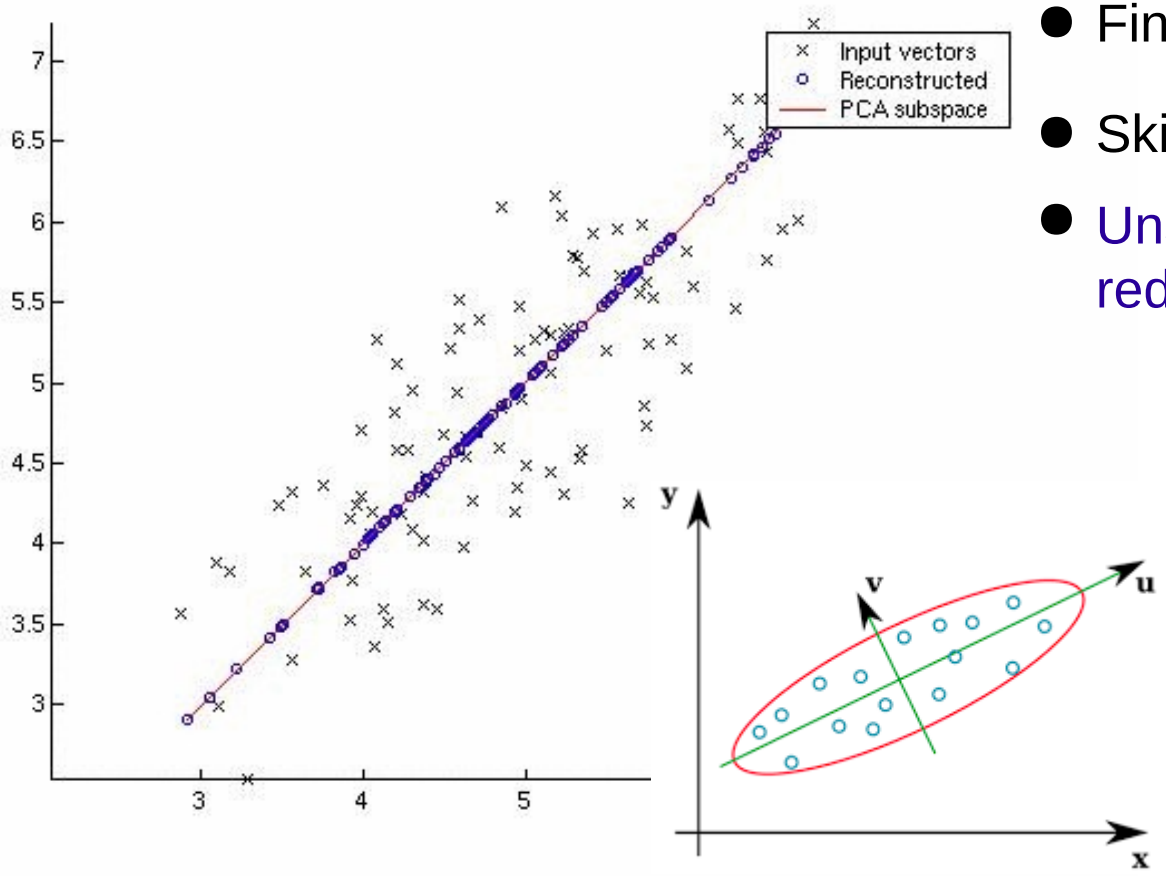
$$Var(X_i) = E\{(X_i - \mu(X_i)) \cdot (X_i - \mu(X_i))\}$$
$$Cov(X_i,Y_j) = E\{(X_i - \mu(X_i)) \cdot (Y_j - \mu(Y_j))\}$$

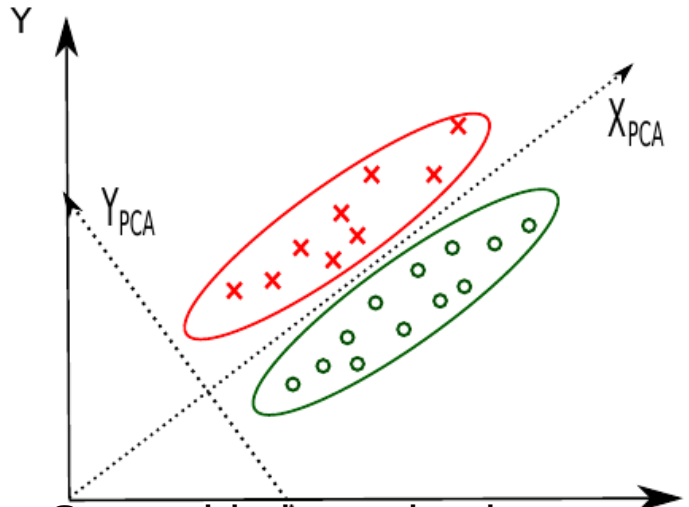
Principal Component Analysis - PCA

- Task: reduce the number of dimensions minimizing the loss of information
- Finds the orthogonal base of the covariance matrix, the eigenvectors with the smallest eigenvalues might be skipped

Procedure:

- Find the covariance matrix Cov(X)
- Find eigenvalues λ_i and eigenvectors v_i
- Skip smallest λ_i
- Unsupervised learning & dimensionality reduction





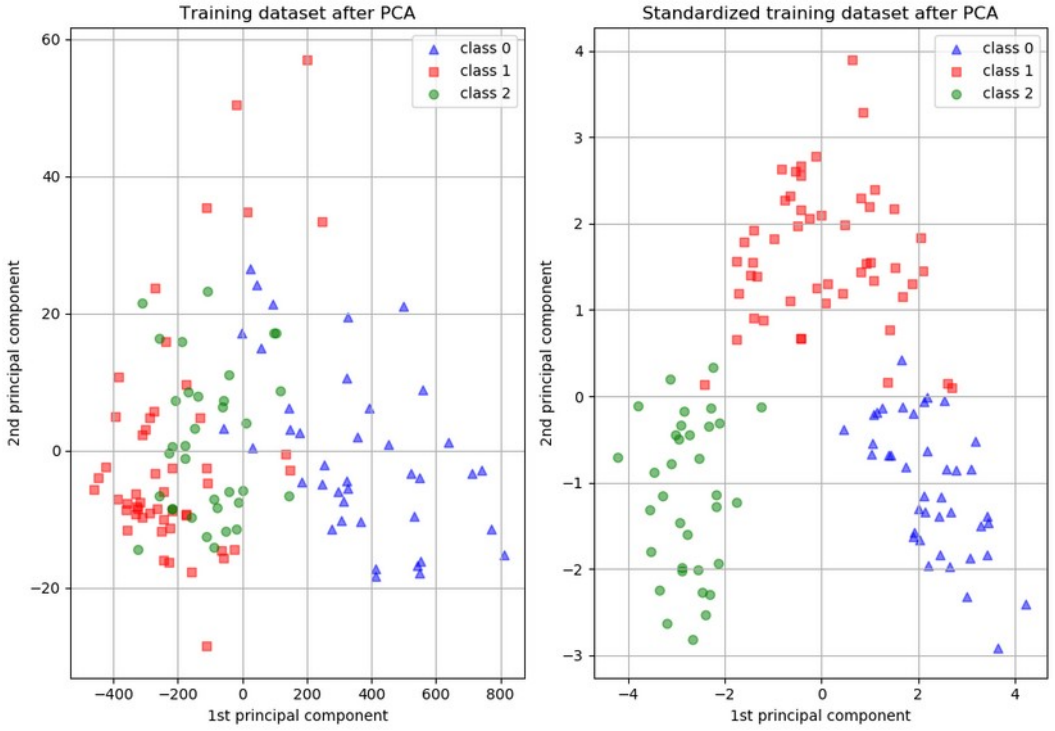
Separable by a simple cut, but PCA doesn't help...

Applications

- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Noise Reduction
- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

What would happen If I did PCA without normalization? Why do we normalize data?

In PCA we are interested in the components that maximize the variance. If one component (e.g. human height) varies less than another (e.g. weight) because of their respective scales (meters vs. kilos), PCA might determine that the direction of maximal variance more closely corresponds with the 'weight' axis, if those features are not scaled. As a change in height of one meter can be considered much more important than the change in weight of one kilogram, this is clearly incorrect.



The dataset used is the Wine Dataset available at UCI. This dataset has continuous features that are heterogeneous in scale due to differing properties that they measure (i.e alcohol content, and malic acid).

Correlation matrix

The correlation matrix refers to the symmetric array of numbers

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

where

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

is the Pearson correlation coefficient between variables $\mathbf{x}_j$ and $\mathbf{x}_k$.

Correlation or covariance matrix?

- Mean-centering is unnecessary if performing a principal components analysis on a correlation matrix, as the data are already centered after calculating correlations.
- We tend to use the covariance matrix when the variable scales are similar and the correlation matrix when variables are on different scales.

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation

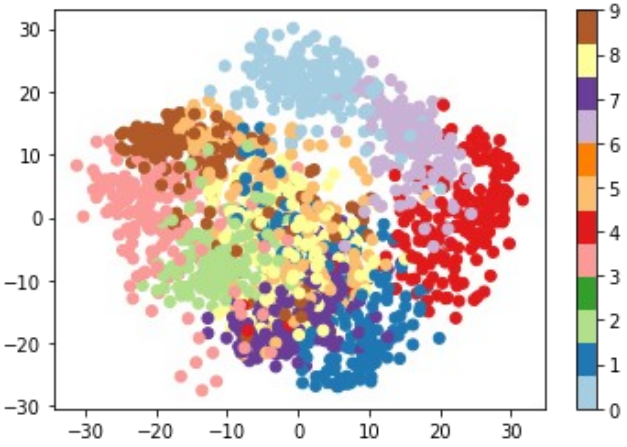
Correlation between X and Y

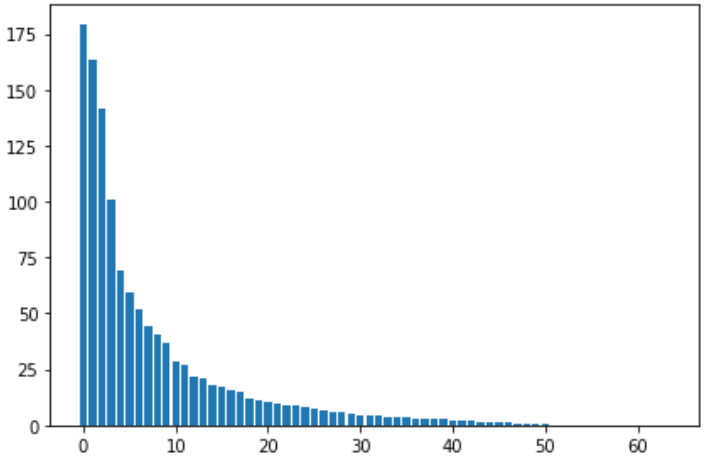
Standard deviation of X

Standard deviation of Y

Example

- All examples will be available here: https://github.com/marcinwolter/MachineLearnin2019
- plot_digits_classif.ipynb - iPython notebook prepared to run on Google Colaboratory https://colab.research.google.com/ (for example)
  - Reads handwritten digits
  - Performs PCA
  - Displays two first principal components:





  - Classification using Naive Bayes and LDA

Summary

- What is machine learning
- Simple unsupervised methods – PCA and decorellation
- Classifiers:
  - Cuts
  - Naive bayes
  - Fisher Linear Discriminants