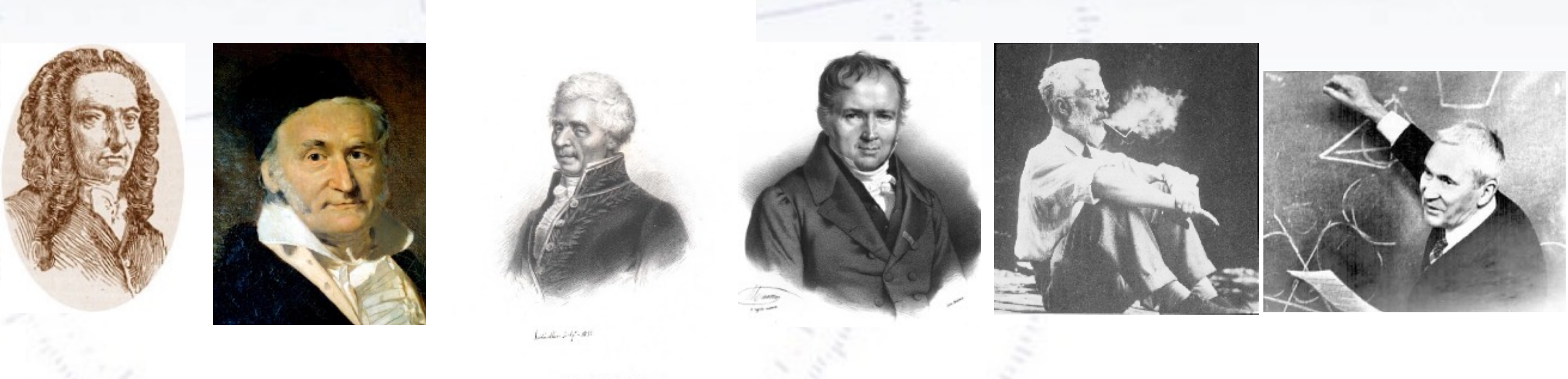
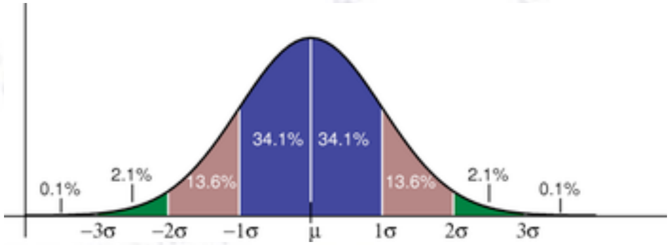
Applied ML

Introduction to Initial Project



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

The motivation

We wanted you to try the very **real challenge** of optimising models, without knowing their performance on the data it is applied to.

We also wanted you to **individually** run ML algorithms, so that you have the machinery in place after the course.

We insisted that you tried **both tree- and NN-based algorithms**, to get a feel for their differences and similarities.

The description file was meant to trigger you to **think about your models**, and what you tried. Also, considerations of size and performance are in place.

Finally, we wanted to **ensure** that you yourself tried all the work and things to consider, to put together ML models and apply them.

The data

The data is (again!) from particle physics, but can (again) be considered to be "anything". The data size is ment to be nice (not too small, not too large), and there are no NaNs put into the samples.

Classification and Regression (from particle physics):

- Training sample (180000 cases).
- Testing sample for classification (60000 cases).
- Testing sample for regression (40000 cases).

Clustering (from astronomy):

- Training and testing sample for clustering (8000 cases).

There are 140 / 24 input variables, but you should NOT use all of them.
Rather, you should determine which are relevant for the task at hand:

- 25 variables for the classification
- 12 variables for the regression
- 7 variables for the clustering

Grading your initial projects

Grading your initial projects

Final Score (example):

You submitted a full solution, from which you get: **64 points**

Your choice of methods based on your description was scored as follows [0, 6]:
Your solution entailed N different algorithms, which gives you a score of [0, 6]:

Your best performance for classification gave: max(0, −log(CrossEntropy − 0.12) × 1.4):
Your variable choice for classification was scored 4 x (VarFreq(you) / VarFreq(top)):
Your classification had 0 penalties, totalling to:

Your best performance for regression gave: max(0, −log(MAD((E-T)/T)/7500−1) × 1.8):
Your variable choice for regression was scored 5 × (VarFreq(you)/VarFreq(top)):
Your regression had 0 penalties, totalling to:

Your best performance for clustering gave: max(0, (Accuracy − 0.75) x 20):
Your variable choice for clustering was scored (VarFreq(you) / VarFreq(top)):
Your clustering had 0 penalties, totalling to:

Thus your total number of points was: …

Overall, focus on delivering good working solutions, and not perfection…

Submitting your solutions

For each solution, you should name your two solutions files as follows:
TypeOfProblemSolved_FirstnameLastname_SolutionName.csv
TypeOfProblemSolved_FirstnameLastname_SolutionName_VariableList.csv

Example:
Classification_TroelsPetersen_SKLearnAlgo1.csv
Classification_TroelsPetersen_SKLearnAlgo1_VariableList.csv

We also want a single overall description:
Description_TroelsPetersen.txt

It is mandatory to run your solutions through the Solution Checker, and surely also a wise thing to do to avoid mistakes.

Submission format:
Your solution file should be in Comma-Separated Values (CSV) format, thus human readable text files. In order to test, if your format is correct, we have produced a file submission checker:
SolutionChecker.ipynb

You should submit your solutions on Absalon by 22:00 on Sunday the 18th of May 2025

Questions?