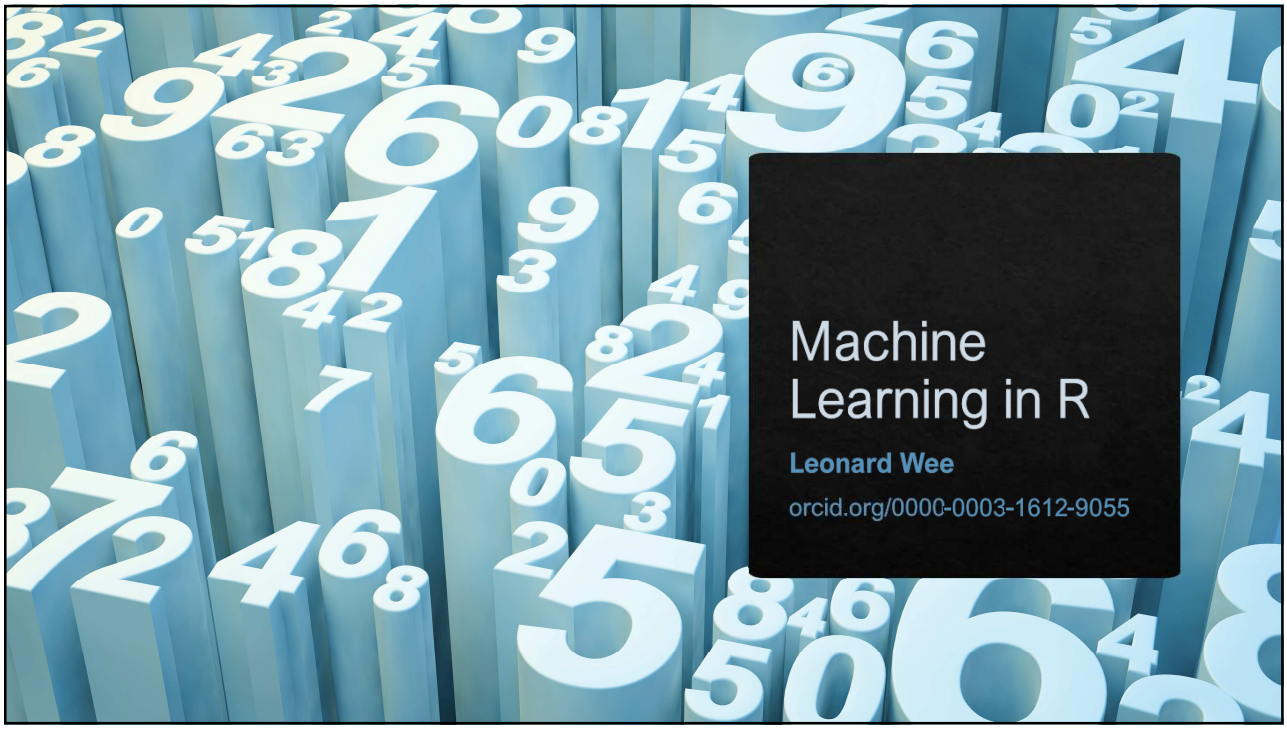
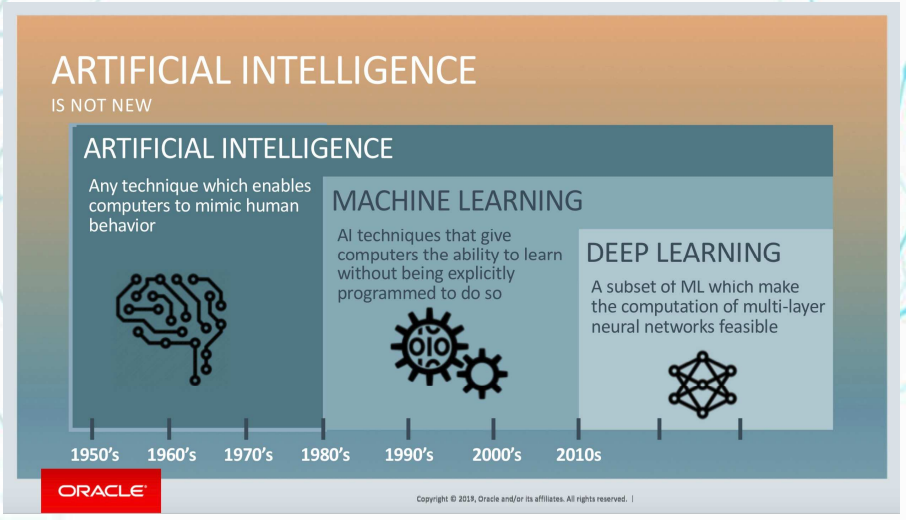


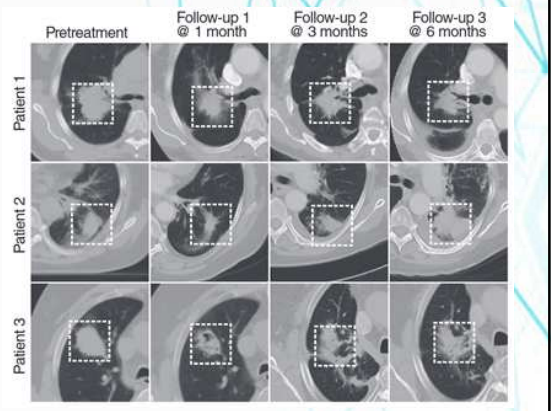
ML and DL grew out from computer science



Possible input sources for ML/DL

rr	oxygen_sat	urea	crp	gcs	age	female	comorbidity	date	set	mortality
33	90	26,2	129,5	13	54	0	2	6-feb-20	dev	0
25	92	21,3	171,2	15	86	0	2	6-feb-20	dev	1
26	89	5,6	164,0	15	86	1	1	6-feb-20	dev	1
14	80	24,4	137,4	15	88	0	2	6-feb-20	dev	1
17	90	6,7	296,0	15	86	0	0	6-feb-20	dev	0
42	92	17,5	127,2	13	55	1	1	6-feb-20	dev	1
19	91	4,1	37,7	15	76	0	1	6-feb-20	dev	1

Cancer (Malignant Neoplasm), Hepatic (Liver)
Assessment: Patient is more lethargic yesterday &
today than he was on Fri ([**2-10**] days ago).
Action: He was made DNR/CMO tonight, per agreement of family.
Assessment: Patient had acute SOB, midsternal chest pain,
feeling that he was going to die @ [**2016**] when he rolled
in bed onto bedpan & had BM. HR increased to low 70s SR.
BP increased to 149/systolic. Desatted to 85%.
Action: Given 100% high flow neb, 0.5 NTP & 0.25mg IV morphine.
EKG done during SOB.
Response: Pain & SOB relieved. No changes on EKG.
Plan: Now that patient is CMO, medicate w/morphine before
rolling patient in bed. Continue to medicate w/Lopressor to
prevent ACS as well as NTP or SL NTG, morphine & O2
during episodes.



Learning Objectives

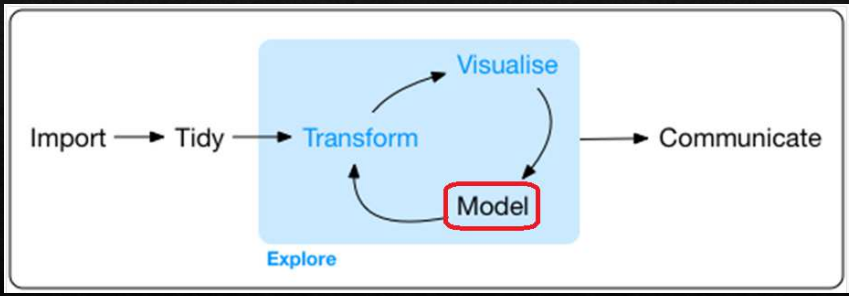
What is ML
Understand the approach
ML work cycle
Interpretation

Classification
Regularized logistic
Trees and forests
Support vector models

Using R
Caret package
Train 3 models
Introduction to SHAP

Online Resources

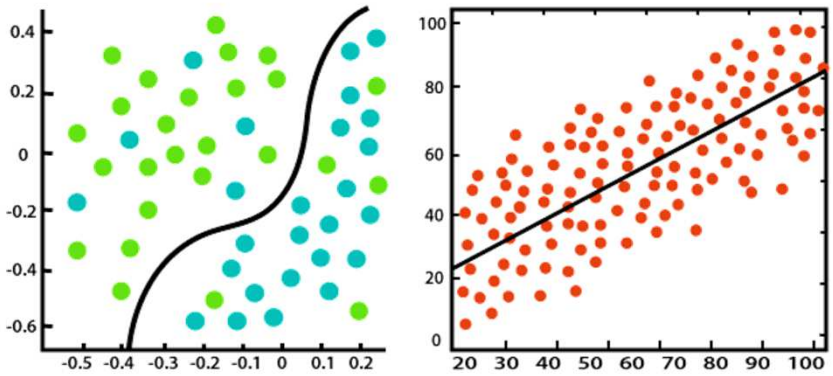
- R for Data Science : https://r4ds.had.co.nz/ (Hadley Wickham)
- The caret package : https://topepo.github.io/caret/ (Max Kuhn)



Some important terminology

- **Data PRE-PROCESSING**
- FEATURE SELECTION
- HYPER-PARAMETERS
- EXPLAINABILITY

Regression compared to classification

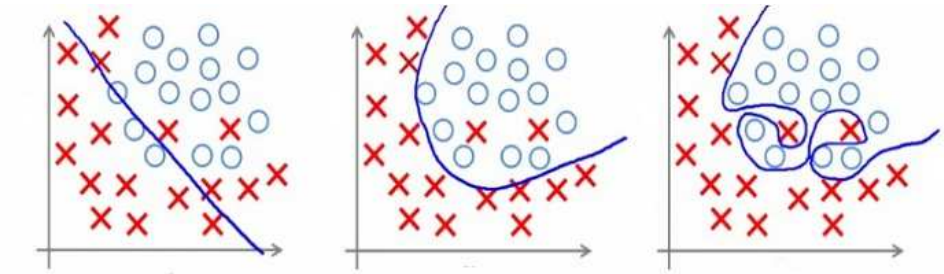


Classification Regression

Maastro

Maastricht University | Maastricht UMC+

Classification - CAUTION



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

Maastro

Maastricht University | Maastricht UMC+

Parameter estimation compared to prediction

One of the key aims of Statistical Modelling :

"Obtain the most accurate estimation of the parameters describing the true relationship between the control variable(s) and the response variable(s)".

e.g. the true Hazard Ratio / Odds Ratio of serum cholesterol for heart attack

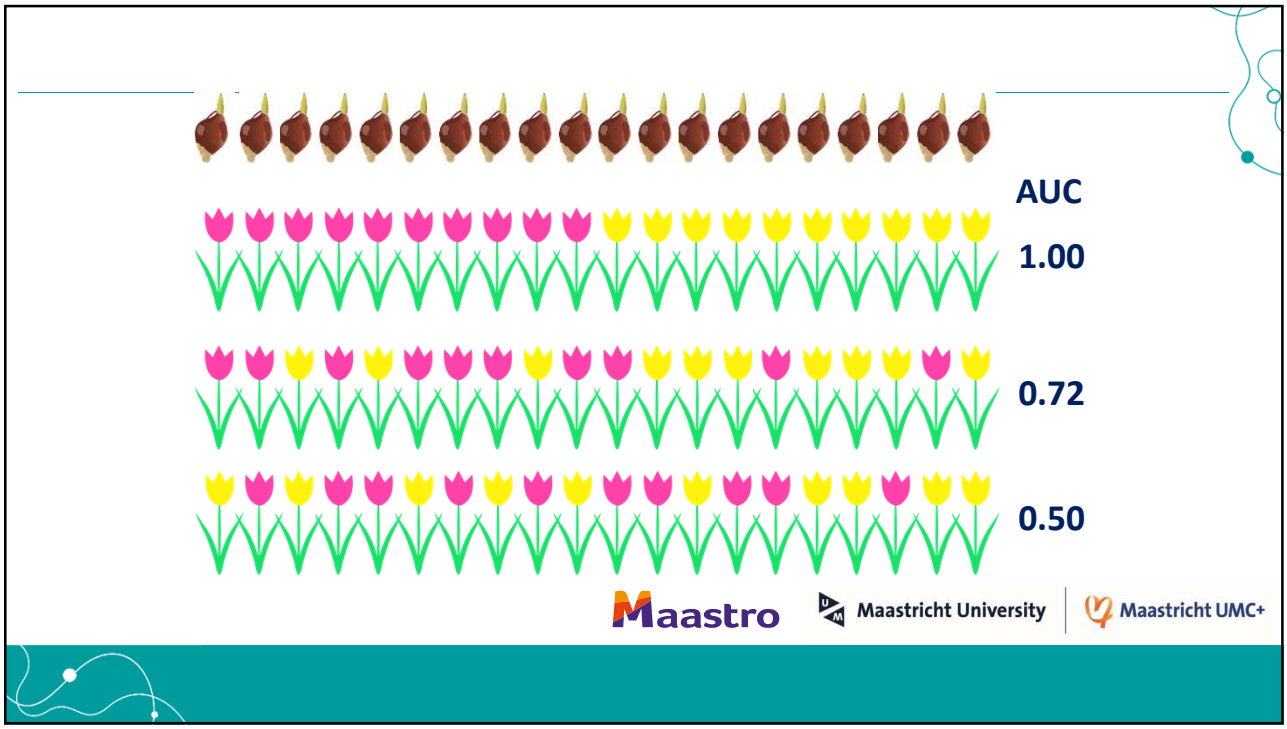
One of the key aims of Machine Learning :

"Obtain the most accurate prediction of the expected outcome for an individual having certain initial characteristics."

e.g. whose cancer is going to return within 2 years after having surgery

Maastro | Maastricht University | Maastricht UMC+

Maastro | Maastricht University | Maastricht UMC+



Complementary approaches

Classical modelling

- Most accurate parameter estimation possible
- Hypothesis driven
- Mostly (but not all) parametric models
- Assumptions (re model) can be strong
- Prominent (pre-eminent) role of pre-existing knowledge and domain expertise
- Usually directly intuitive / interpretable

Machine learning

- Most accurate individual-level prediction possible
- Data driven
- Parameterization is useful but not essential
- Assumptions (re model) are somewhat weaker
- Pattern recognition and phenomenological
- Can be explainable but generally much less interpretable

Examples of questions

Classical modelling

- If I reduce the number of hospital beds by 10%, what is the expected increase in mortality?
- After correcting for age and sex, is there a relationship between serum cholesterol and heart disease?
- If I wish to raise my market share by 10%, should I spend money on advertising the existing credit card or launch a new credit card?
- Will a social-media based public health campaign reduce smoking rates among young women by 25%?

Machine learning

- If I give this particular Br Ca patient in front of me a higher radiation dose to the breast, does her cancer recur within the next 2 years or not?
- Should I put this person with these specific blood level readings on a "surveillance" plan for heart attack?
- This customer is right now asking for a home loan of $200k, shall I approve it or not?
- Which order of ads should I present on the smartphone of this young woman to encourage her to quit smoking?

Some important terminology

- **Data PRE-PROCESSING**
- **Dimensionality REDUCTION**
- Model tuning HYPER-PARAMETERS
- Model EXPLAINABILITY

Pre-processing

Step 1 : Are there any columns with little or no variation?

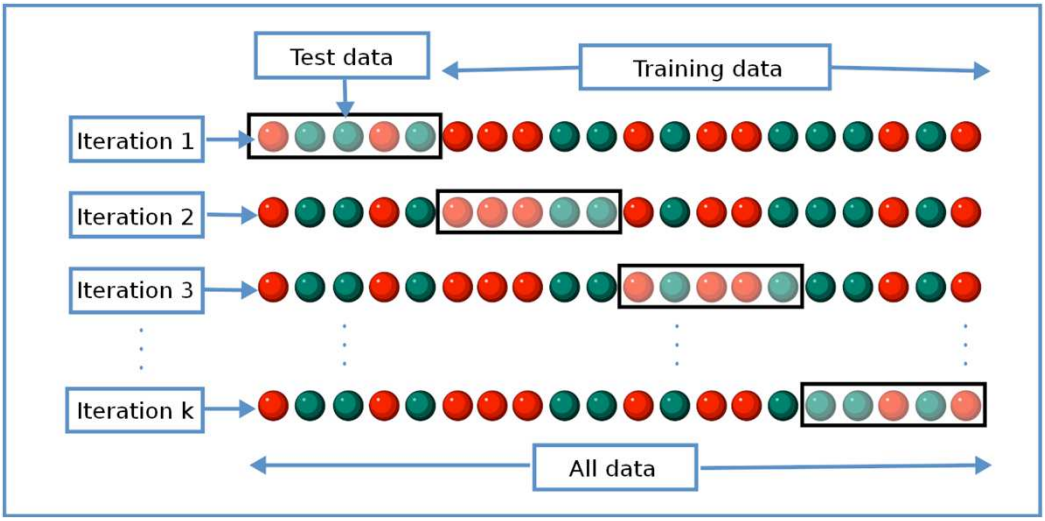
Step 2 : Are there missing values? Should we impute or not?

Step 3 : (Multi)-collinearity – does anything need to be done about it?

Step 4 : Categorical variables – recode these. Numerical values – consider if scaling/transform/centering is needed

Step 5 : Which subjects to "train" on and which to validate on?

K-fold sampling

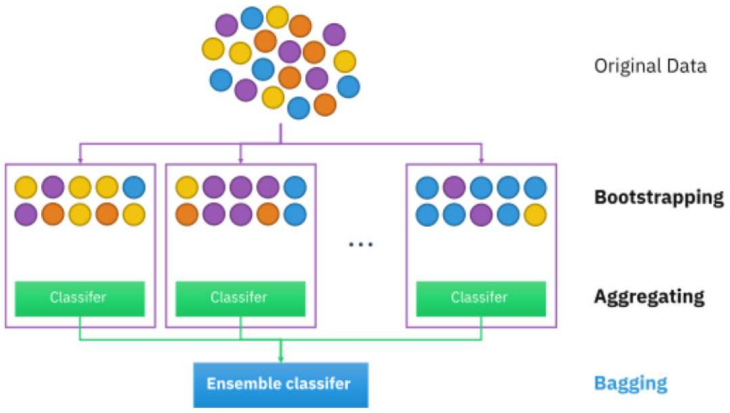


Maastro
Maastricht University | Maastricht UMC+

Bootstrap sampling

- Independent estimates of model performance in a large "true" population
- But based on a finite one
- Assumptions?
- Confidence intervals
- Over-optimism estimation
- Variations include:

- Minority oversampling
- Majority undersampling



Maastro

Maastricht University | Maastricht UMC+

Regularized (logistic) regression

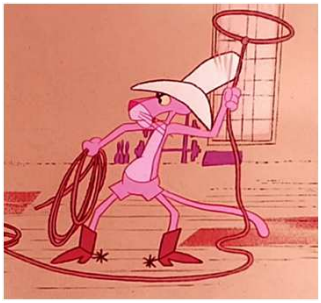
We frequently use **regularization** to make a machine learning model less sensitive to data that is not in the training set.

But how do we know what data is NOT in the training set?!?

Answer : K-folds and/or bootstraps for cross-validation, to estimate the distribution of data outside training

Least **A**bsolute – **S**hrinkage and **S**election **O**perator (LASSO) :

- Lasso penalty shrinks or reduces the coefficient value towards zero.
- Least-contributing variables are gradually "shrunk" …
- Towards zero or almost-zero coefficients, leaving strongly-contributing ones.
- Hyper-parameter **LAMBDA 1** ($\lambda 1$) which is the "cost" of a non-zero coefficient
- The model optimizers job is to drive the cost to the minimum.



Maastro

Maastricht University | Maastricht UMC+

Regularized (logistic) regression

RIDGE regularization applies lambda to the SUM OF SQUARES of coefficients
(cf lasso penalty on the absolute value of coefficients)

- Ridge penalizes relatively large or relatively small coefficients.
- Coefficients are gradually "shrunk" …
- Towards the same magnitude of coefficients!
- Hyper-parameter **LAMBDA 2** ($\lambda 2$) which is the "cost" of differences.
- The model optimizers job is to drive the cost to the minimum.

ELASTIC NET == linear mixing of $\lambda 1$ and $\lambda 2$ penalties.

Hyper-parameter ALPHA (α) : $\alpha = 0$ means pure RIDGE, $\alpha = 1$ means pure LASSO.

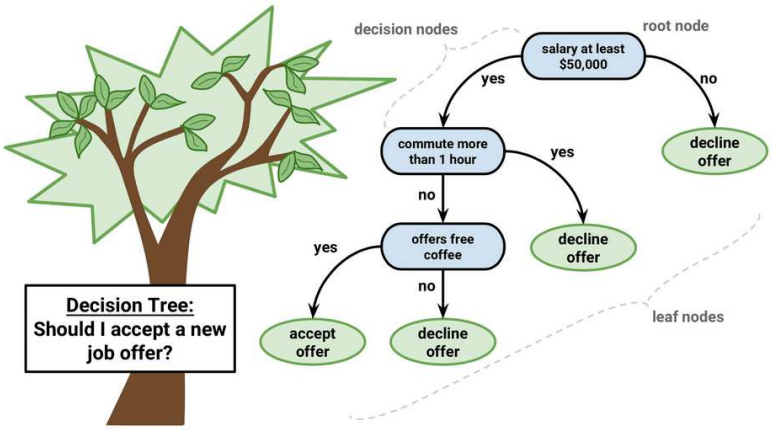
Maastro

Maastricht University | Maastricht UMC+

Tree-based classifiers

Based on the classical decision tree :

- **Applying many consecutive if-then rules**
- ML trees **do not require expert guidance** to select the nodes and decision values.
- **Usually classification** (but some trees can be used to predict values at the output nodes i.e. regression).



Maastro

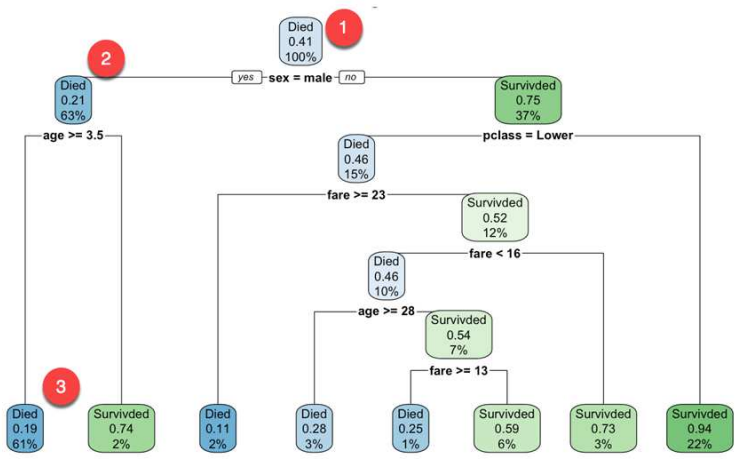
Maastricht University | Maastricht UMC+

Interpretation of tree-based classifiers

Assuming the Titanic is representative of ship sinking events, then **predict** - given certain passenger characteristics - whether he/she will die or survive.

e.g. 17 year-old female in 1st class cf 20-year old male in 3rd class?

- ***Interpretability?***
- ***Explainability?***
- ***Generalizability?***
- ***What is your opinion?***



Maastro

Maastricht University | Maastricht UMC+

Construction of tree-based models

Any similarities with UNSUPERVISED CLUSTERING?

Main principle - in any given tree and each of its sub-tree(s),

- Minimize entropy within subgroups, **and**
- Maximize information gain ie biggest change (in entropy) when choosing a threshold to split to subgroups

Therefore we need an entropy metric -> Gini index (class heterogeneity) is default in R tree package

(Hyper-parameter) Human needs to decide where to CUT THE TREE to avoid fitting model on singular (or extreme) events

Maastro

Maastricht University | Maastricht UMC+

Random forest model (extending tree classifier)

Random Forests (RFs) == very large ensemble model containing very many independent tree models.

(Hyper-parameter) Human expert can decide how many variables per tree?

(Hyper-parameter) Human expert can decide how many trees in total?

The output of random forest (nb – there are many variations on the random forest theme) :

(passenger characteristics) == "died" **(i.e., most popular vote among all the trees)**

- ***Interpretability? Explainability? Generalizability? What is your opinion?***

SHapley's Additive exPlanations (SHAP) in R

Suppose a group of you are doing a joint assignment.

If nobody did anything, you would all get a joint "1" score ie you wrote your name and submitted it ☹

You all did some (but varying) amounts of work and the group grade was "8" ☺

Behavioural economist : "Can we calculate the contribution of each member towards the group net profit of 7?

If you are interested in the game-theoretic modelling and answers, look up :

Shapley, Lloyd S. (August 21, 1951). "Notes on the n-Person Game -- II: The Value of an n-Person Game" (PDF). Santa Monica, Calif.: RAND Corporation.

Shortish answer : We try out every possible combination of variables, and rank variables one-by-one according to which brings us closest to the final overall predicted class probability.

SHapley's Additive exPlanations ("SHAPPER" in R)

Main objective is not to calculate ourselves but interpret the output :

SHAP

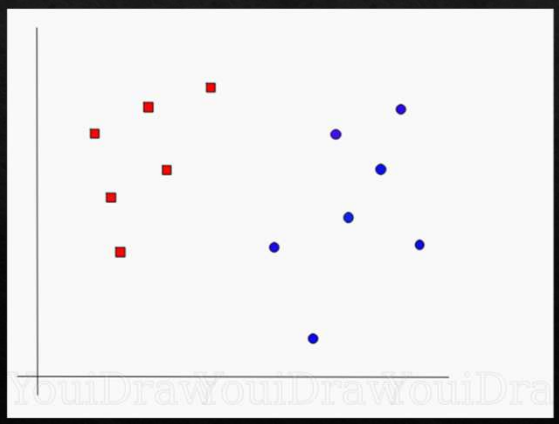
1. Train a model first
2. Feed in a new "patient"
3. Extract prior and posterior class probabilities
4. Calculate the Shapley values for each variable
5. Rank and display

Maastro

Maastricht University | Maastricht UMC+

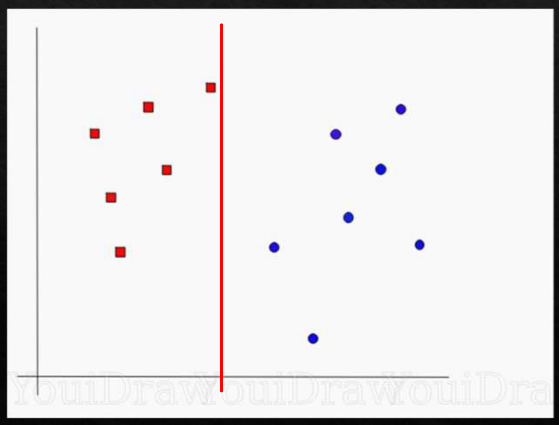
Support vector models – linear space

- Example : Draw a classification line to divide the "red" outcomes from the "blue" outcomes.



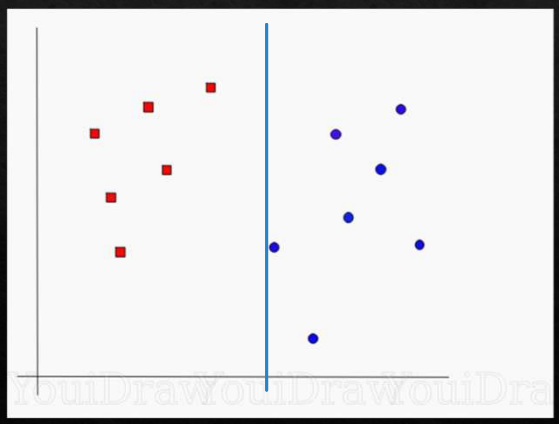
Support vector models – linear space

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



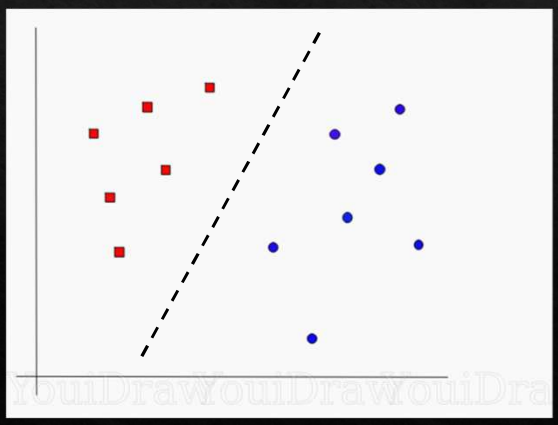
Support vector models – linear space

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



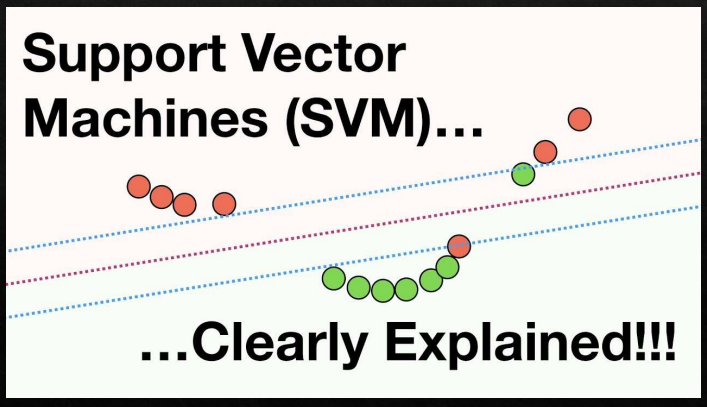
Support vector models – linear space

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



Support vector models – linear space

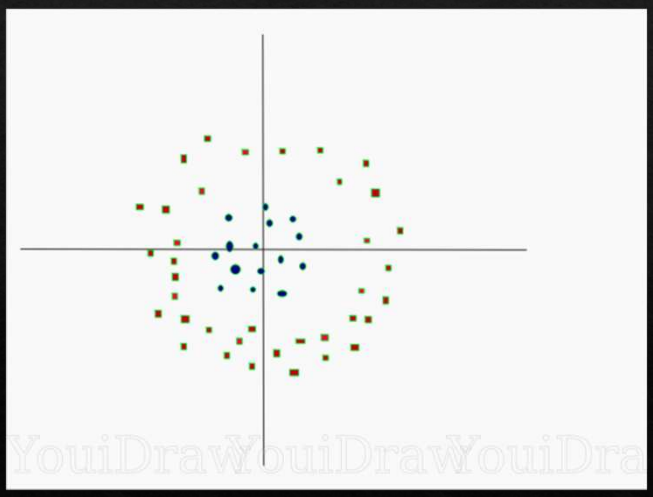
- What happens if the boundary is noisy?



https://www.youtube.com/watch?v=efR1C6CvhmE

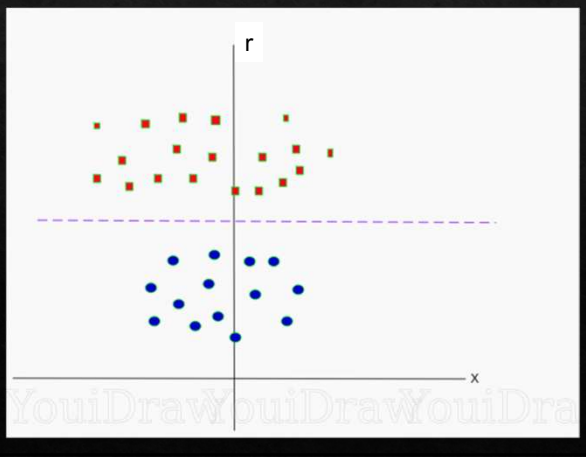
Support vector models - nonlinear

- Now try this one …



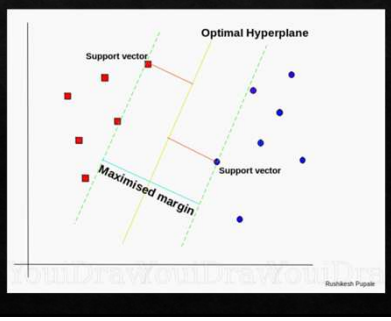
Support vector models – non-linear

- Effect of a transforming from a Cartesian to a Radial coordinate system …



Support vector models

- Hyperplane : “In n-dimensional space, the hyperplane is a (n-1) dimensional surface that cuts the space into two distinct regions.”
- An SVM **optimal hyperplane** is the one that divides the data points so that the labels have minimum overlap and maximum distance from the hyperplane.
- **Linear SVM** = n-dimensional Cartesian space
- **Radial SVM** = n-dimensional Spherical space



Coming up in next practice session :

- Look in detail at 3 machine learning models
- Covid-19 dataset
- **Please install “caret” R library in advance of the practice session**
- **caret documentation : https://topepo.github.io/caret/**

Learning Objectives

What is ML

Understand the approach

ML work cycle

Interpretation

Classification

Regularized logistic

Trees and forests

Support vector models

Using R

Caret package

Train 3 models

Introduction to SHAP