6th Summer School on INtelligent signal processing for FrontIEr Research and Industry

30th August 2021, University Autónoma de Madrid

INFIERI

Introduction to Machine Learning and Deep Learning (Part I)

Juan Carlos San Miguel

Associate Professor at UAM & Researcher at VPULab &
Director of Master in Deep Learning for Audio and Video Signal Processing

Juancarlos.sanmiguel@uam.es

UAM Universidad Autónoma de Madrid

Escuela Politécnica Superior

VPU Video Processing and Understanding Lab

OUTLINE

- What is Machine Learning?
- Performance evaluation
- Examples of Machine Learning algorithms
- Ensembles
- Conclusions

- **Arthur Lee Samuel** (1901-1990)
  Pioneer of artificial intelligence research
  [IEEE Computer Pioneer Award](#) 1987

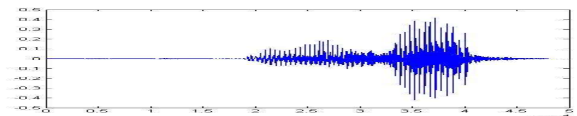
"Field of study that gives computers the ability to learn from data without being explicitly programmed"

Source: https://history-computer.com/people/arthur-samuel-biography-history-and-inventions/

WHAT IS MACHINE LEARNING?

“Field of study that gives computers the **ability** to learn from data without being explicitly programmed”

- Looking for a function to mimic human brain decisions…
  - Speech Recognition

$f($  $)=$ “How are you”

  - Image recognition

$f($  $)=$ “Cat”

  - Playing Go

$f($  $)=$ “5-5” (next move)

WHAT IS MACHINE LEARNING?

"Field of study that gives computers the ability to learn from **data** without being explicitly programmed"

- It can be any **unprocessed digital signal** of any nature like a fact, value, text, sound or picture
- It can have temporal dependency **(time-series)**
- Often transformed to **Numerical and Categorical** types

- Organized as **Datasets**, which are collections of data instances that all share a common attribute
- Requires **annotations** of attributes for each data instance of the dataset **to measure efficiency**

WHAT IS MACHINE LEARNING?

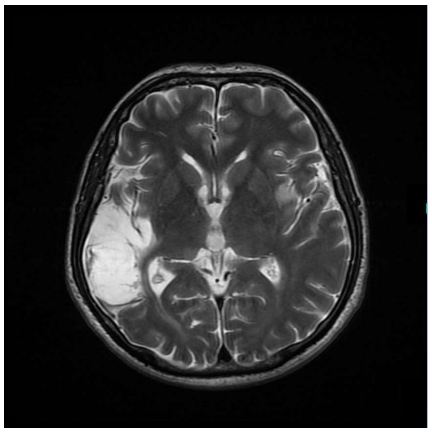
"Field of study that gives computers the ability to **learn** from data without being explicitly programmed"

- **Learning problems** in Machine Learning

Algorithm employs data annotations?

Type of signals		Supervised learning	Unsupervised learning
	Discrete	Classification or categorization	Clustering
	Continuous	Regression	Dimensionality reduction

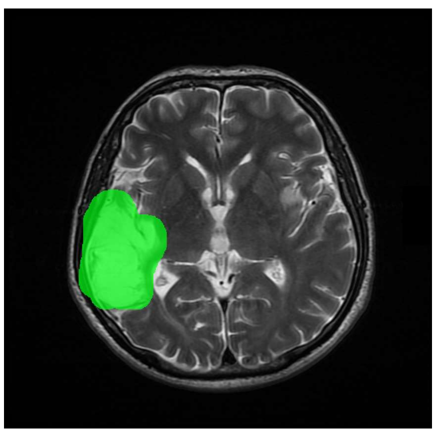
- The **accuracy of ML algorithms** must be evaluated to choose the best one for each specific task



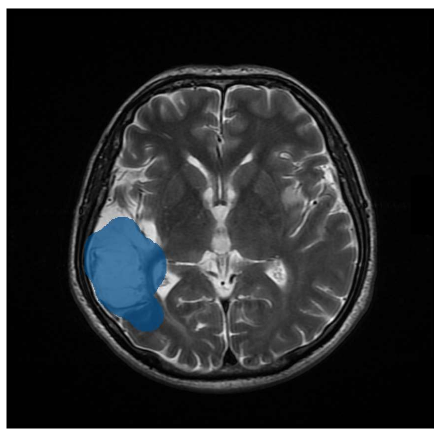
Task

Brain Tumor Segmentation in MRI images

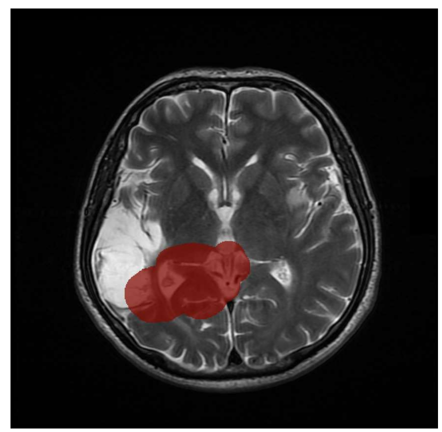
(i.e. identify which image pixels are tumor)



ML algorithm #1



ML algorithm #2

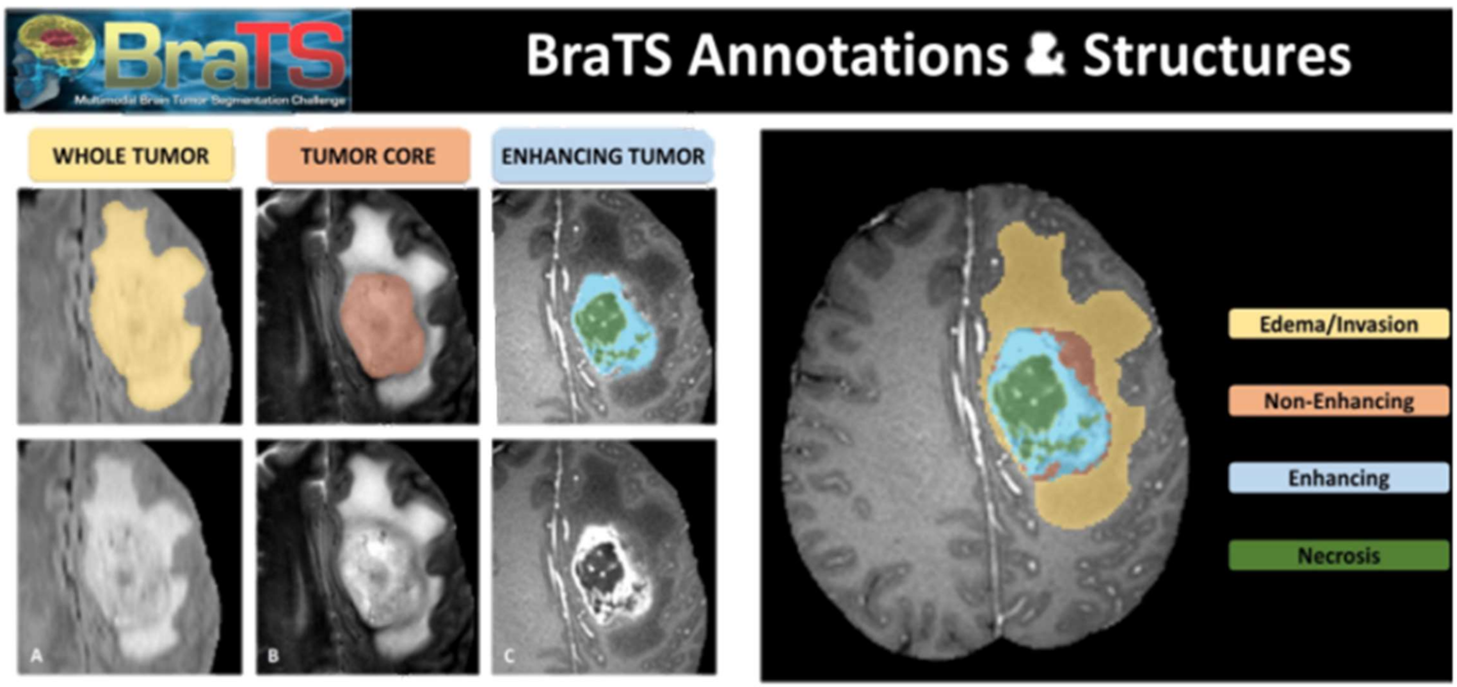


ML algorithm #3

Which one is the best?

Example created with https://htmlsegmentation.s3.eu-north-1.amazonaws.com/index.html

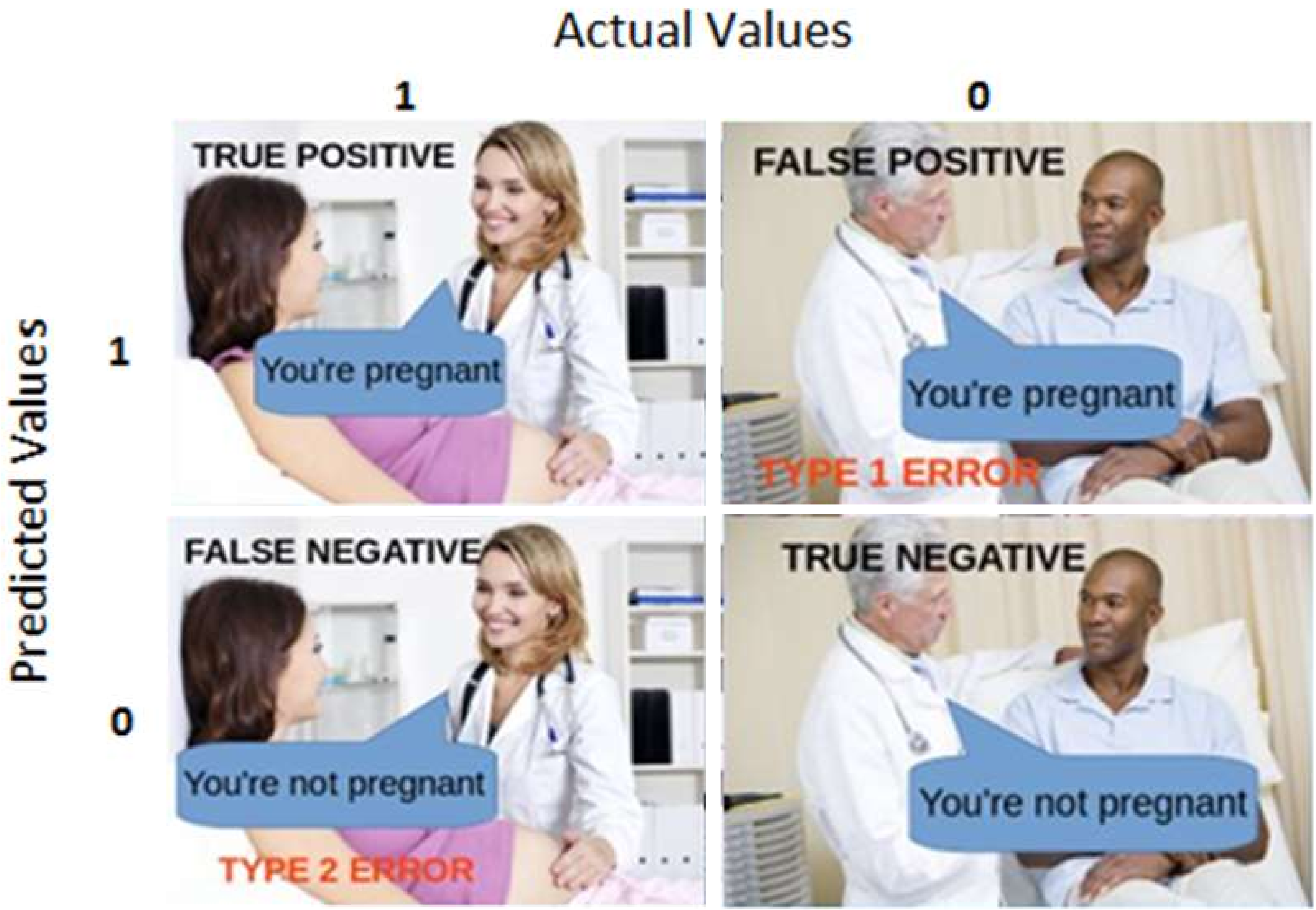
- Three **key elements**
  - **Result**: prediction of the algorithm (e.g. category, scalar value,…)
  - **Ground-truth**: the **knowledge of the truth** for the specific task. (e.g. ideal expected result for the category, scalar value,…)
  - **Metric**: function to compute similarity between result and ground-truth



Source: https://www.med.upenn.edu/cbica/brats2020/data.html

PERFORMANCE EVALUATION

- **Metrics** for **binary classifier** evaluation
  (can be also applied to classify data instances into multiple classes)



Source: http://www.info.univ-angers.fr/

PERFORMANCE EVALUATION

- Classification **accuracy** and **error**

$$\text{Accuracy} = \frac{TP+T}{TP+TN+FP+F} = \frac{\#\, correct\ predictions}{\#\, total\ predictions}$$

$$\text{Error rate} = \frac{FP+F}{TP+TN+FP+FN} = \frac{\#\, wrong\ predictions}{\#\, total\ predictions}$$

- **Confusion** matrix
  - Performance visualization and summary of results
  - Diagonal are correct predictions
  - Allows to focus on errors

Predicted class / Actual class	Cat	Dog
Cat	6	2
Dog	1	3

Extended description at
https://en.wikipedia.org/wiki/Confusion_matrix

- *Many more metrics… (see suggested readings)*

PERFORMANCE EVALUATION

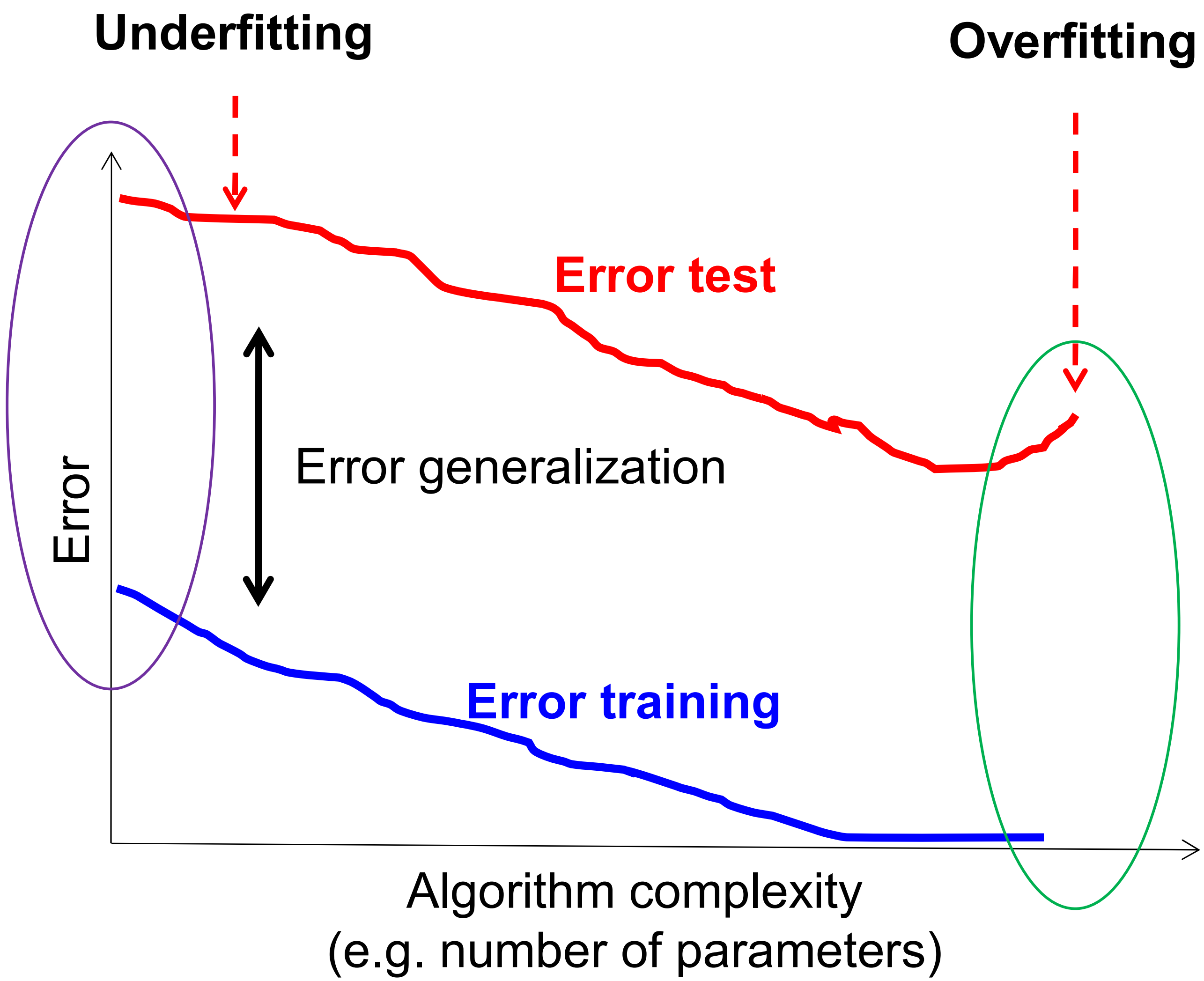
- If dataset is large (i.e. millions of samples), split in two sets:
  - Train (85-98%): algorithm fitting (i.e. adjust) parameters for best performance
  - Test (15-2%): validate the algorithm trained with different data
- If dataset is not large (i.e. thousands of samples), then dataset is randomly split into **“k” folds** (often k=5 so 20% each)

Iteration 1	Train (20%)	Train (20%)	Train (20%)	Train (20%)	Test (20%)	→ Error or acc
Iteration 2	Train (20%)	Train (20%)	Train (20%)	Test (20%)	Train (20%)	→ Error or acc
Iteration 3	Train (20%)	Train (20%)	Test (20%)	Train (20%)	Train (20%)	→ Error or acc
Iteration 4	Train (20%)	Test (20%)	Train (20%)	Train (20%)	Train (20%)	→ Error or acc
Iteration 5	Test (20%)	Train (20%)	Train (20%)	Train (20%)	Train (20%)	→ Error or acc

Average error or accuracy

- Moreover, a **validation set** is often added to add fairness in evaluation
  - **Train set** used for algorithm fitting (resulting in a learned model)
  - **Validation set** used to estimate prediction error for selecting the best model
  - **Test set** used to assess the generalization error of the final chosen model

- Error generalization: algorithm complexity for a given dataset

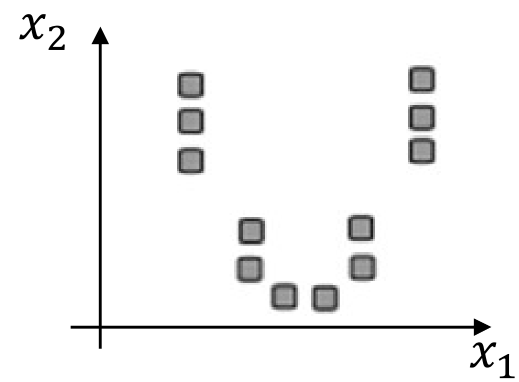


Adapted from D. Hoiem

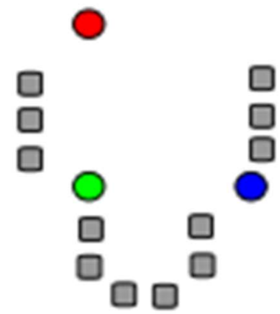
MACHINE LEARNING ALGORITHMS

Source: https://www.deepmarketer.com/blog/2017/1/30/machine-learning-algorithm-taxonomy

- Unsupervised learning: K-means[1]
  - Iterative algorithm for clustering



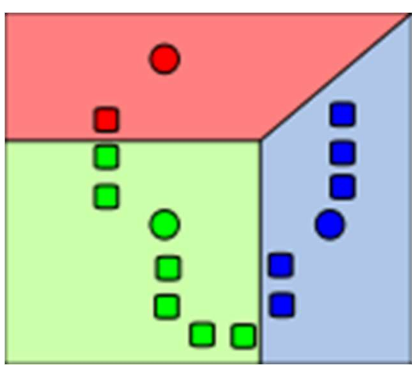
Step 1: Select the number of clusters and set randomly a cluster center (i.e. representative)

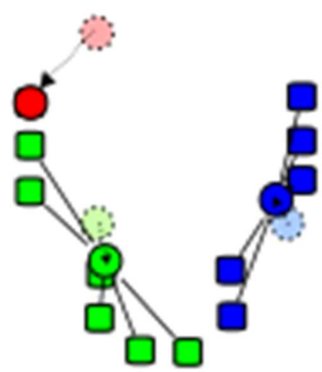


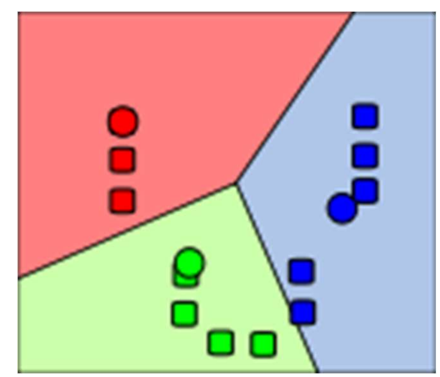
Step 2: associate each data to clusters by minimum distance with cluster centers

Step 3: update cluster centers with the mean of new data associated to each cluster in step 1

Step 4: repeat step 2 and 3 until convergence of cluster centers







[1]Lloyd, Stuart P. "Least Squares Quantization in PCM."
IEEE Transactions on Information Theory. Vol. 28, 1982, pp. 129–137.

Credit images: https://en.wikipedia.org/

- Supervised learning: Support Vector Machines[1]

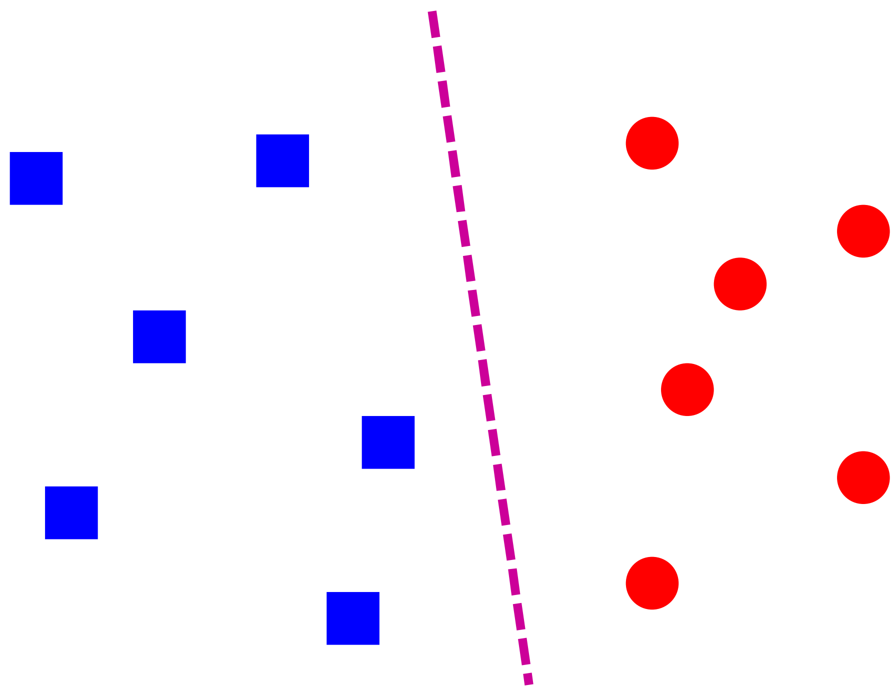
A linear classifier learns a lineal function to determine the classification boundaries

Training data sample ith

$$\hat{y}^i = f\left(\vec{\omega} \cdot \overrightarrow{x^i}\right) = f\left(\sum_{j=0}^{N_d} \omega_j \cdot x_j^i\right)$$

Algorithm params

Training data for class A



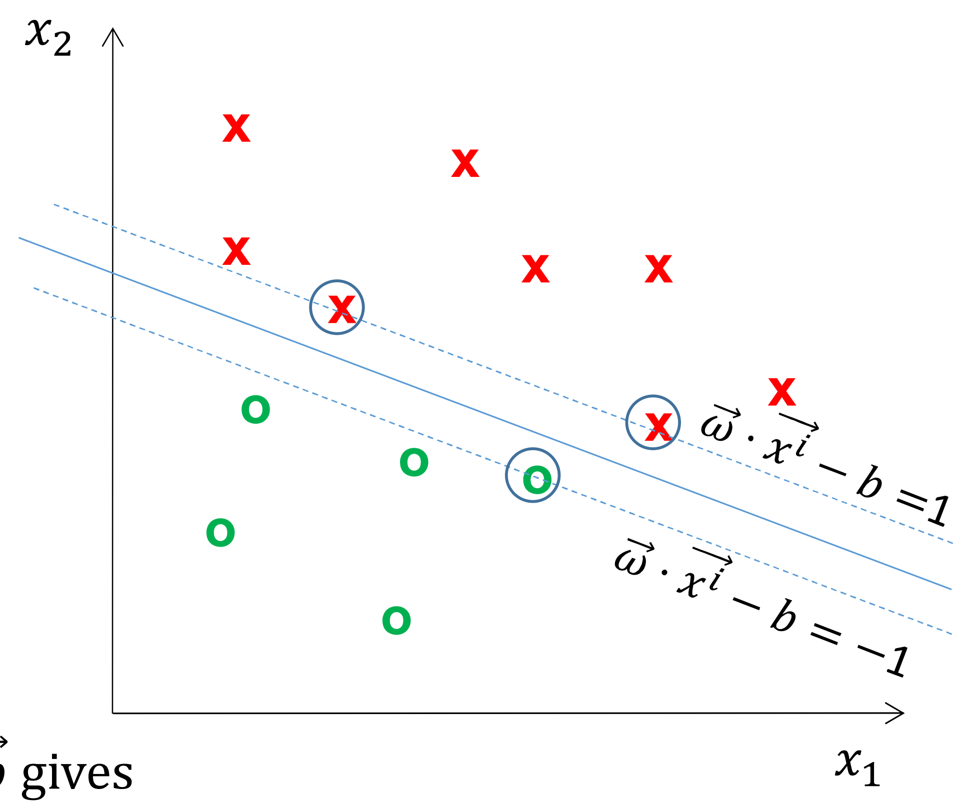
Training data for class B

[1]C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, 1998

- Supervised learning: Support Vector Machines
  - Defines a hyperplane $\vec{\omega} \cdot \overrightarrow{x^i} - \vec{b} = 0$ for binary classification

$$\hat{y}^i = sgn\left(\vec{\omega} \cdot \overrightarrow{x^i} + \vec{b}\right)$$

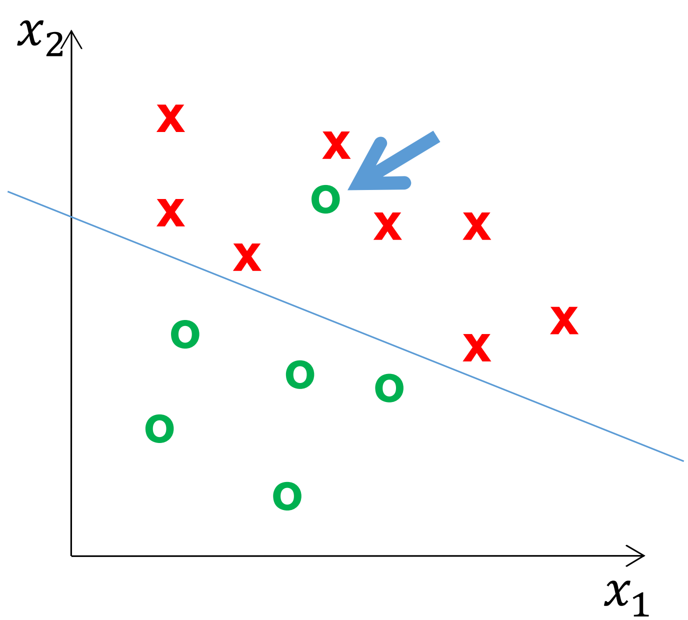
$$\hat{y}^i = sgn\left(\sum_{j=1}^{N_d} \omega_j \cdot x_j^i + \vec{b}\right)$$



  - Prediction
    - **Class x** (+1) if $\vec{\omega} \cdot \overrightarrow{x^i} - \vec{b} > 0$
    - **Class o** (-1) if $\vec{\omega} \cdot \overrightarrow{x^i} - \vec{b} < 0$
  - Training
    - $minimize \|\vec{\omega}\|$ subject to $\vec{\omega} \cdot \overrightarrow{x^i}$ - $\vec{b}$ gives the correct classification y^i for all data samples $\overrightarrow{x^i}$
    - Optimal solution $\overrightarrow{\omega_{opt}} = \sum_k \alpha^k y^k \overrightarrow{x^k}$

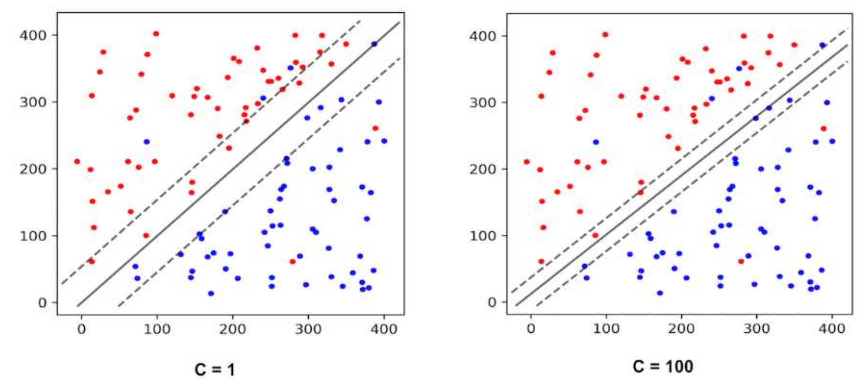
- Supervised learning: Support Vector Machines

What if data is non-linear?



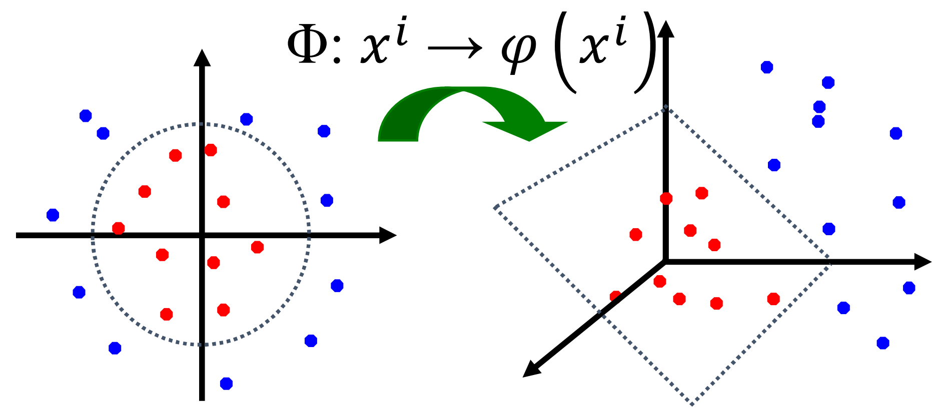
Option A

Employ a linear SVM and tolerate errors (i.e. add a C regularization term)



Option B

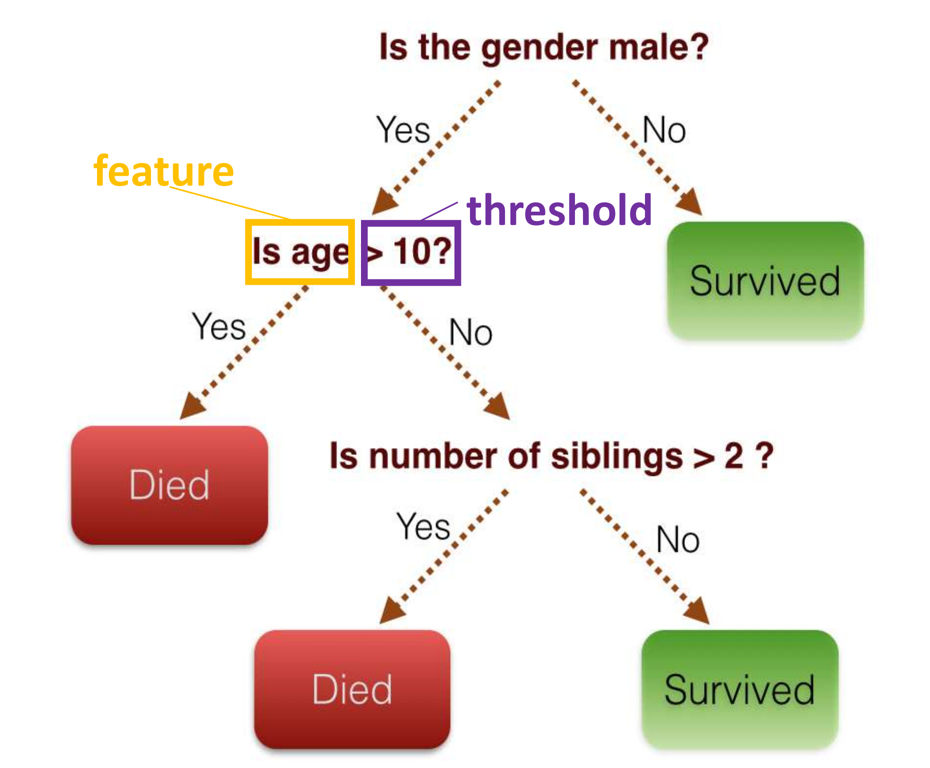
Map data instances to a higher dimensional space where data is linearly separable



Credit images: https://www.learnopencv.com/svm-using-scikit-learn-in-python/ & Andrew Moore

- Decision Trees[1]:
  - Very popular algorithm due to their intelligibility and simplicity
  - **Classification or regression**
  - Structure:
    - Root Node
    - Intermediate Nodes
    - **Leaf nodes → predictions**
  - Tree structure built sequentially by:
    - Splitting data into subsets (i.e. for each available feature)
    - Measuring feature performance
    - Finding the optimal threshold

Survival of passengers on the Titanic

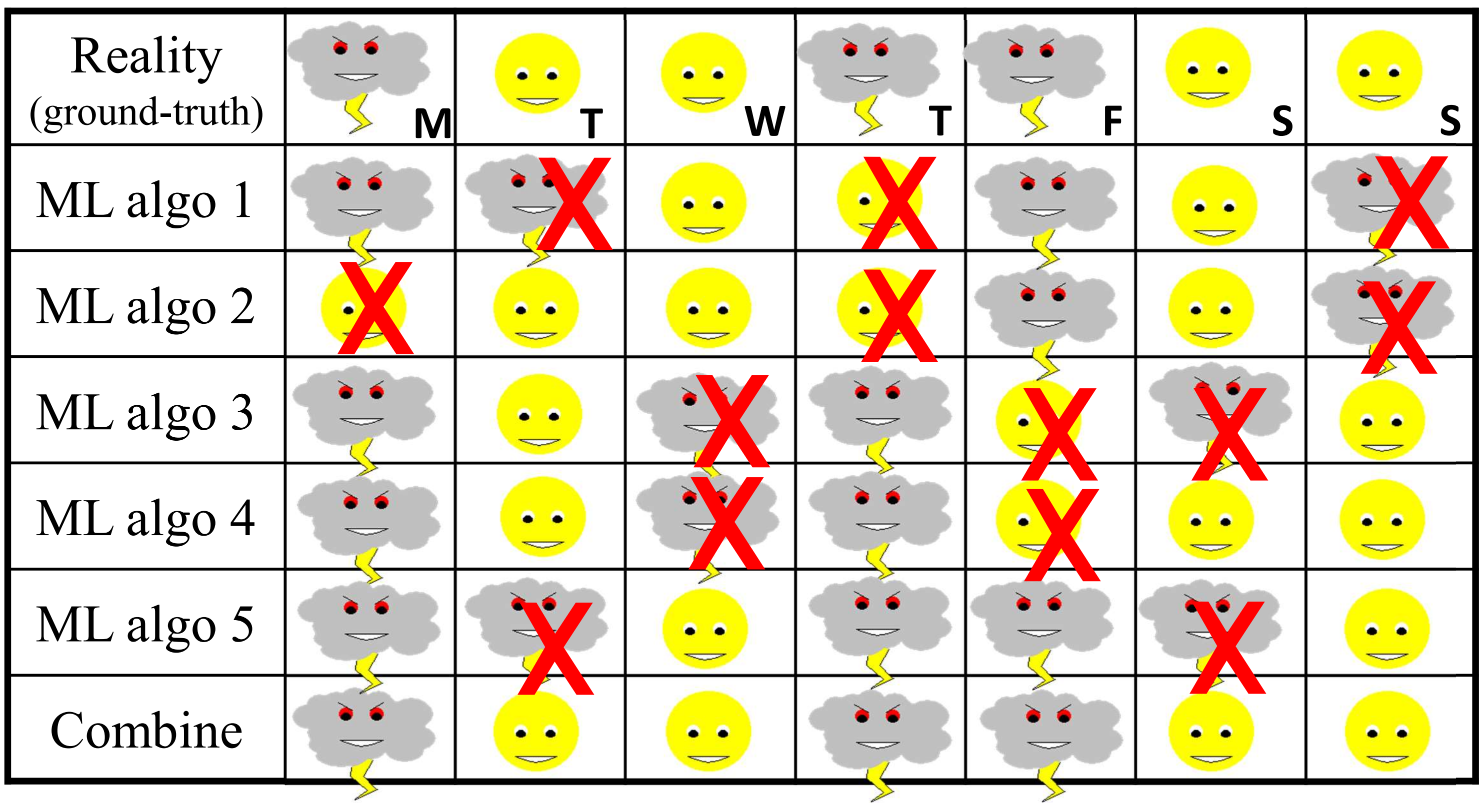


Source: https://towardsdatascience.com/

[1]X. Wu et al. "Top 10 algorithms in data mining". Knowledge and information systems, 14(1), 1-37. 2008.

- Combine **multiple algorithms** applied to the same data to get one high-accuracy meta-algorithm
  - "No Free Lunch" Theorem - No single algorithm wins all the time!

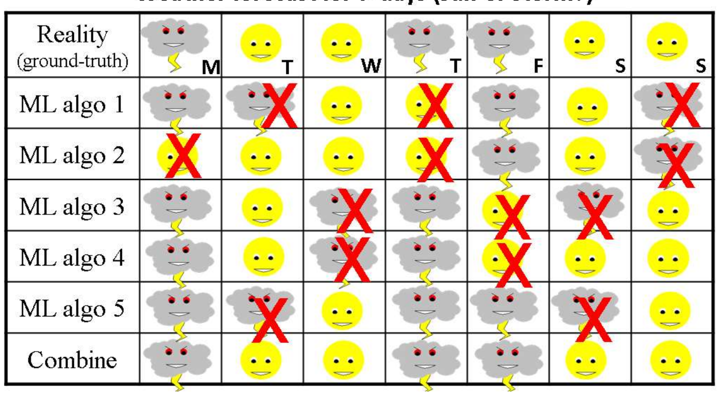
Weather forecast for 7-days (sun or storm?)



Example based on Dr. Carla P. Gomes

- When combing multiple **independent** and **diverse** predictions which are at least more accurate than random guessing, random errors cancel each other, correct predictions are **reinforced**.
- Often **weak learners** are employed in the ensemble (low-accuracy but very fast time for training and prediction)

Weather forecast for 7-days (sun or storm?)



Each **classifier has 70% accuracy** for the task and it is independent to other classifiers

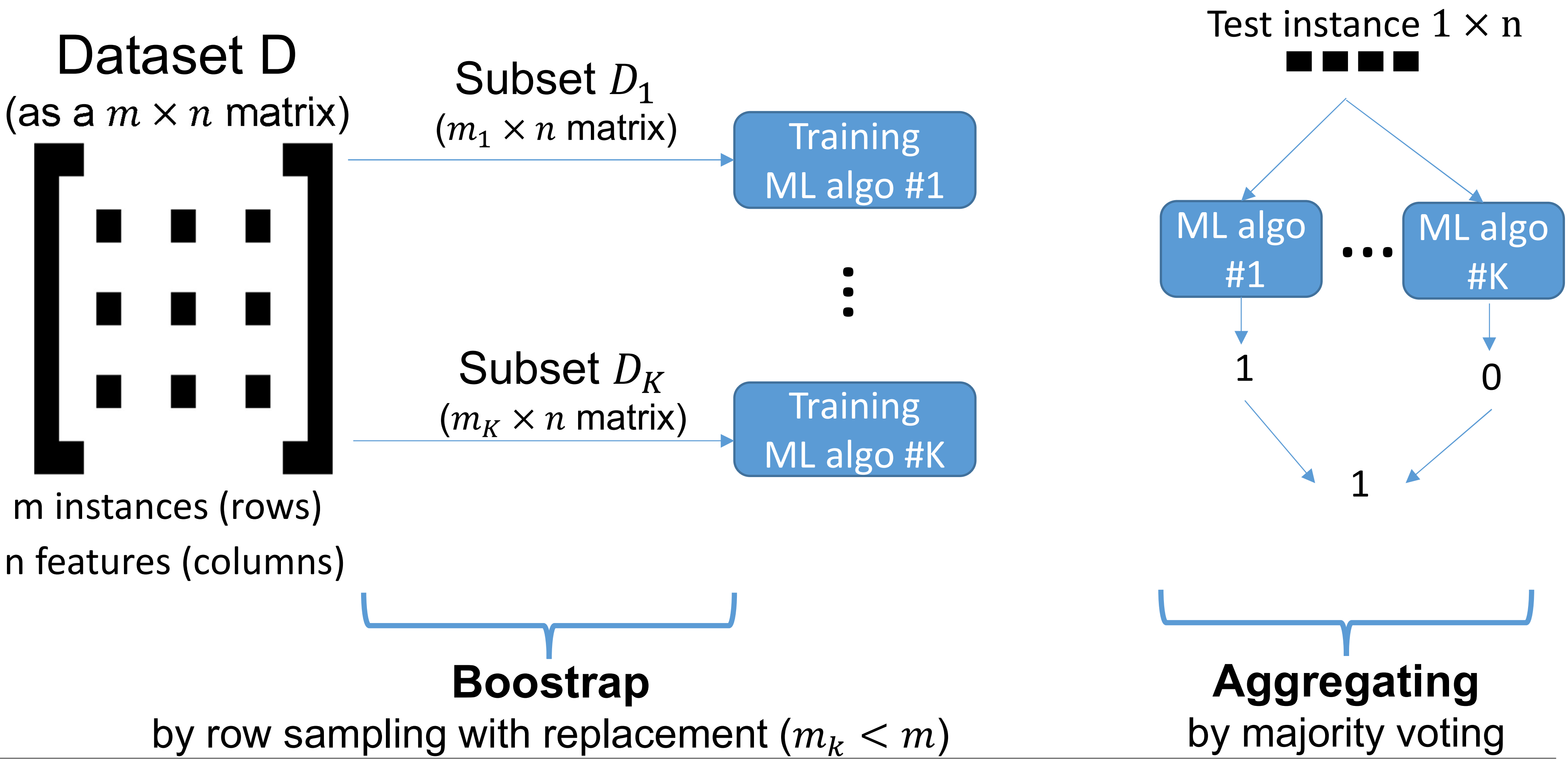
Majority vote accuracy

- 5 classifiers - **83.7% accuracy**
- 101 classifiers - **99.9% accuracy**

Hint: Probability that k out of n independent trials of a random experiment are successful, with success probability p is $\binom{n}{k}p^k(1-p)^{n-k}$

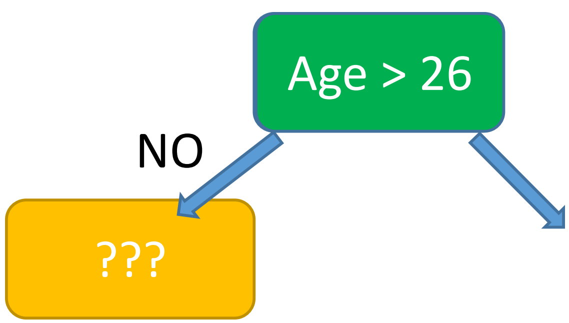
- Two main design choices
  - **Combining strategies**: averaging, majority vote, stacking,…
  - **Learning paradigm**: bagging, boosting.,…

Goal: reduce the variance (i.e. low test accuracy) of combining weak learners by parallel training each algorithm with a subset composed of random selection of data instances



- Random Forest[1]
  - Widely used ensemble method that employs decision trees.
  - However, decision trees alone tend to overfit when becoming deep (overfitting ≡ high variance ≡ high train accuracy and low test accuracy)
  - To overcome this limitation, features are randomly selected for each node of the tree, so to avoid dependency on "dominant" features

Decision Tree DT_1



For each intermediate node, take remaining features and repeat the random selection & choosing the best feature

Subset D_1
($m_1 \times n$ matrix)

#children	Age	Commute time	Salary	Change job?
0	27	30min	32K	**No**
2	30	15min	35k	**No**
0	22	30min	20k	**Yes**
1	27	15min	25k	**Yes**
1	26	15min	25k	**Yes**

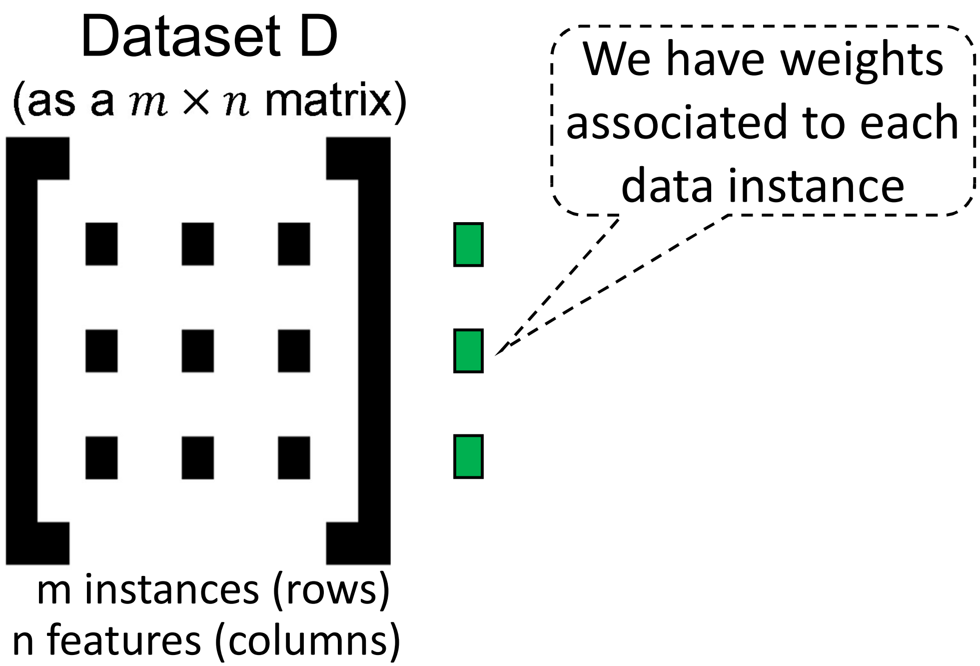
[1]L. Breiman. "Random Forests", Machine Learning 45 (1): 5-32. ,Springer, 2001

Goal: learn a cascade of algorithms (weak learners), where each algorithm attempts to correct the previous errors

Bagging / Random Forest	Boosting / Adaboost[1]
- All data samples have equal weight - Parallel algorithms - All algorithms have equal say - Fully grown trees (each tree may have different depth)	- Data samples have adaptive weights - Sequential algorithms - All algorithms have different say - Weak learners: stumps (trees with one node and two leaves)

[1]Y. Freund & R. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*. 55: 119–139, 1997

- Adaboost[1]



Training stage

Initialization:

- Set equal weights to all data instances
- Define the number of algorithms

…

Recursively do for each ML algorithm ith

1. Training
   a) Randomly select a subset of data based on data instance weights
   b) Train ML algorithm *ith*
   c) Evaluate accuracy using subset
2. Update
   a) Compute "say" for algorithm *ith* proportional to its accuracy
   b) Update weights of data instances
      - Increase weight if incorrect
      - Decrease weight if correct

Test stage

Execution: apply all the algorithms in the ensemble for each data instance.

Get overall "say": accumulate the "say" or importance of classifiers for the predictions

Final prediction: prediction with the highest accumulated importance or "say"

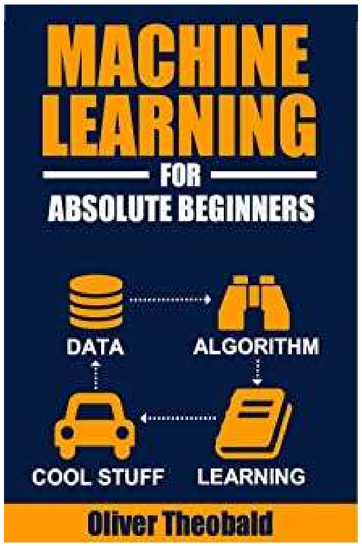
[1] Y. Freund & R. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*. 55: 119–139, 1997

CONCLUSIONS

- Machine Learning requires to prepare raw data in order to remove noise, errors or impose requirements of algorithms
  - Data cleaning & annotation, Feature selection & transforms, …
- Must understand the type of ML algorithms needed to solve a particular problem (it may be a mix of different types)
- Most of the training strategies oriented to avoid overfitting (i.e. high train accuracy but low test accuracy)
- Finding best algorithms/ensembles may involve running multiple times with different settings (hyperparameter tuning)

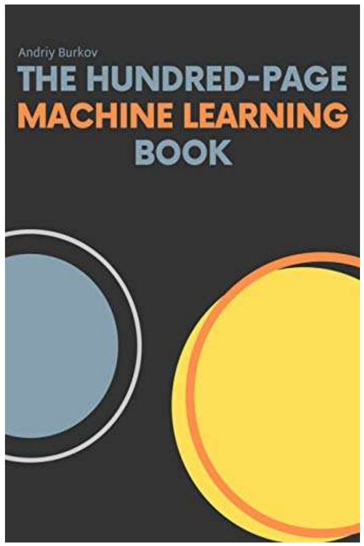
WANT TO LEARN MORE?

Beginner



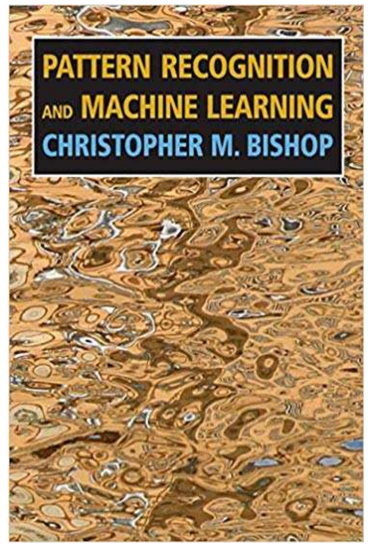
2nd Ed 2021
https://amzn.to/2TUhHXW

Intermediate



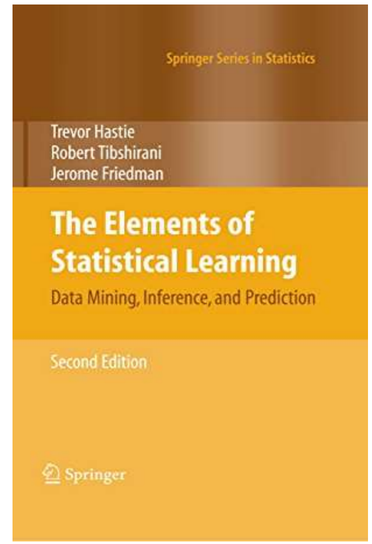
2019,
https://bit.ly/2TRmtW4

Intermediate



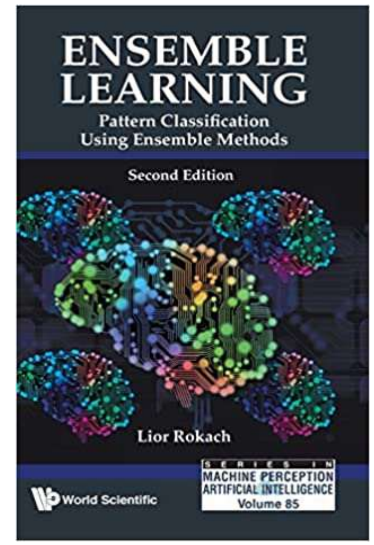
2006
https://bit.ly/2V7wz5W

Expert



2009
https://bit.ly/3jhvPTw

Expert



2nd Ed 2019
https://amzn.to/3io2R57

- As for practical work, please do check tutorials and up-to-date examples available for popular ML & DL frameworks (TensorFlow, PyTorch, scikit-learn, Spark ML, Torch, Keras,…)

6th Summer School on INtelligent signal processing for FrontIEr Research and Industry

30th August 2021, University Autónoma de Madrid

Introduction to Machine Learning and Deep Learning (Part I)

Juan Carlos San Miguel

Associate Professor at UAM & Researcher at VPULab &
Director of Master in Deep Learning for Audio and Video Signal Processing
Juancarlos.sanmiguel@uam.es

ANY QUESTIONS?