INTRODUCTION TO MACHINE LEARNING

Recap Days 1-2-3

Pablo Cañas
pablocanas97@gmail.com

NORTHIN SUMMER SCHOOL
Barcelona, July 2023

Course structure

1. **Introduction:** basic concepts, data preparation, linear regression
2. **Fundamental Machine Learning concepts**: Loss functions, optimization, cross-validation, overfitting
3. **Supervised learning**: naive bayes, k-NN, random forests, …
4. **Unsupervised learning**: clustering, dimensionality reduction.
5. **Introduction to Deep Learning**

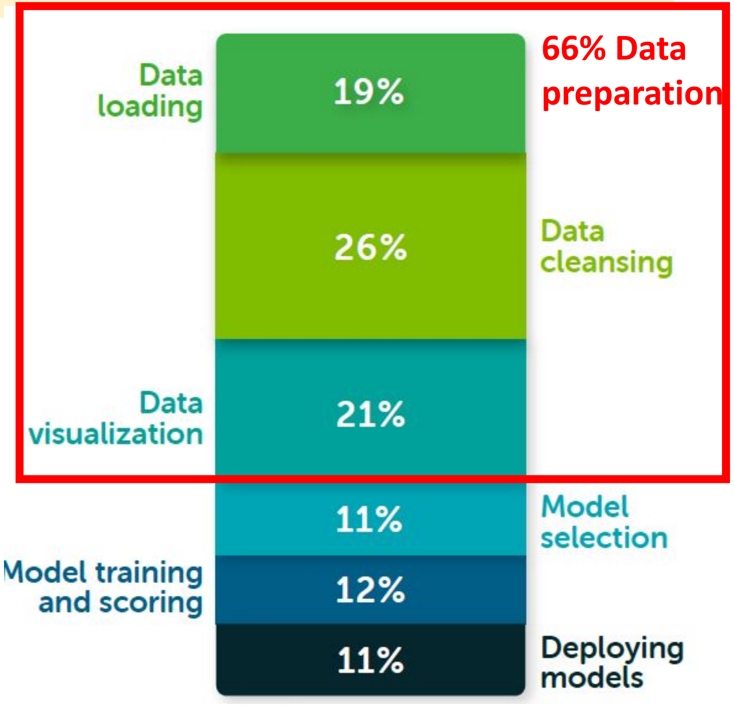
What is Machine learning?

- Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions
- Machine learning algorithms seek to provide knowledge to computers through **data**, observations, and interaction with the world. It is then used to make **accurate predictions** given new observations.
- Machine learning is applied statistics!

Data preparation

- Data is the oil of machine learning
- Prepare and analyse the data before applying machine learning!

Source: Memegenerator



Source: Anaconda

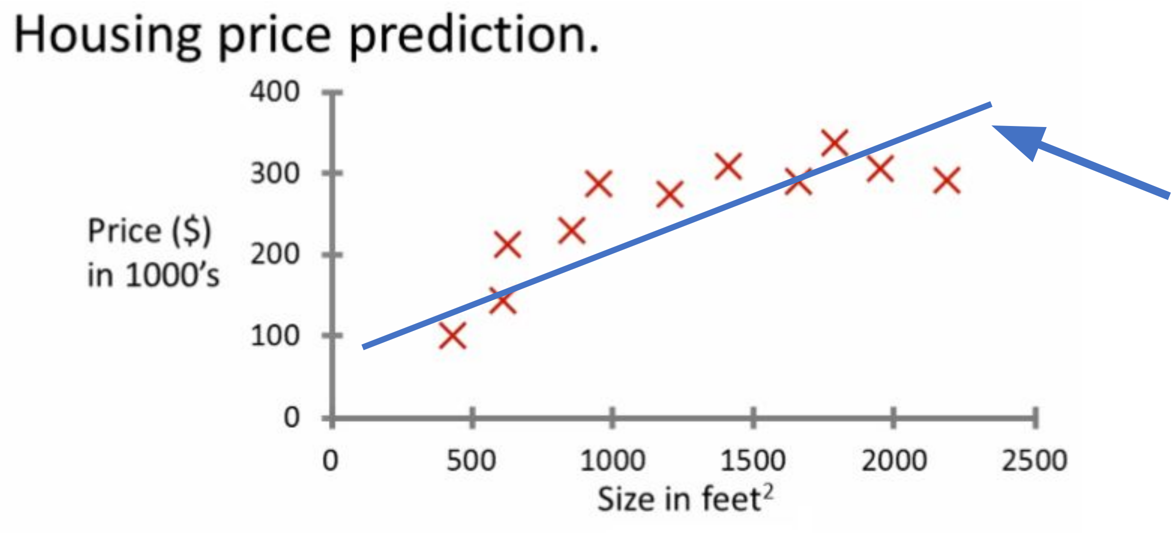
Machine learning algorithms

- **Supervised**: Models are **trained** with **input/output** pairs (X, y) which we relate via a function $y = f(X)$. Model **learns** f to make predictions on new data inputs X.
  - ➢ **Classification:** predictions/outputs y are discrete (class labels)
  - ➢ **Regression:** y is continuous

- **Unsupervised**: Only **inputs** X are given. We compute f such that $y = f(X)$ is a "simpler" representation
  - ➢ **Clustering:** discrete y (groups)
  - ➢ **Dimensionality reduction**: continuous y

Supervised Learning: Regression

- Predict **continuous** valued output. Ex: Housing price prediction.

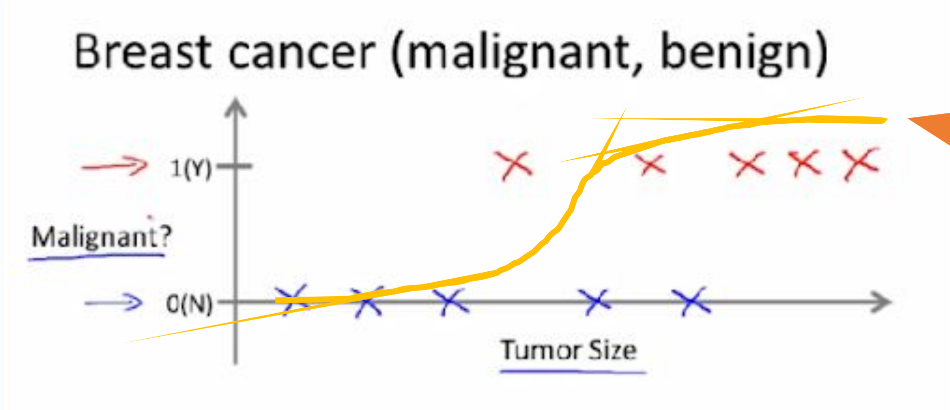


Linear regression model, predict price

Source: Coursera

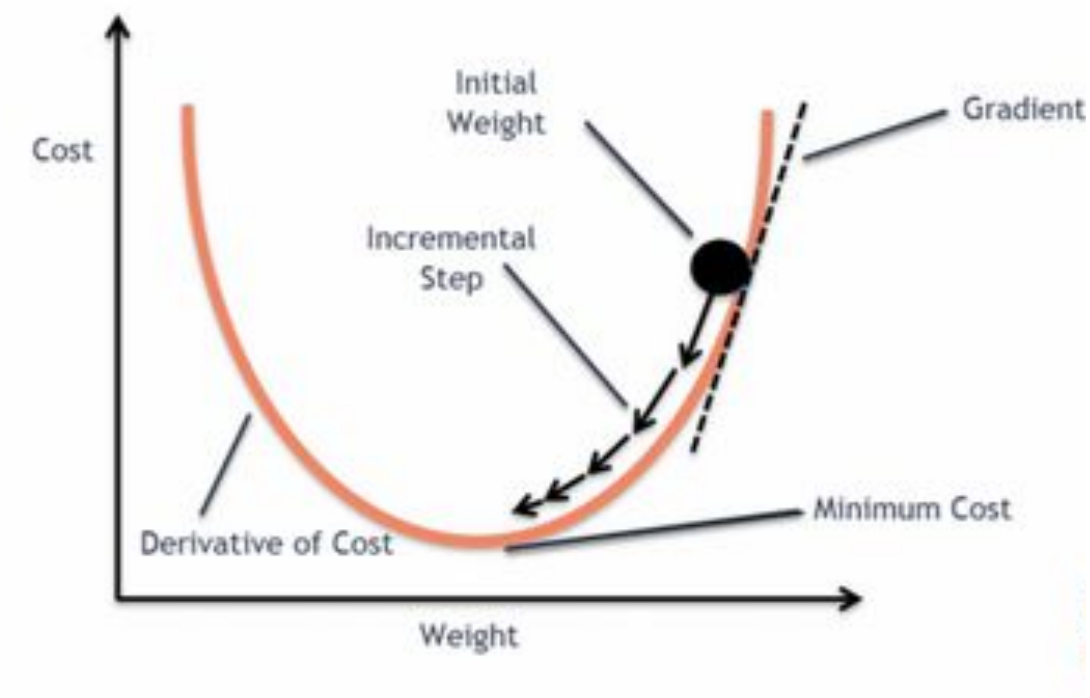
Supervised Learning: Classification

- Predict **discrete** valued output. Ex: Breast cancer diagnose $y \in \{0,1\}$.



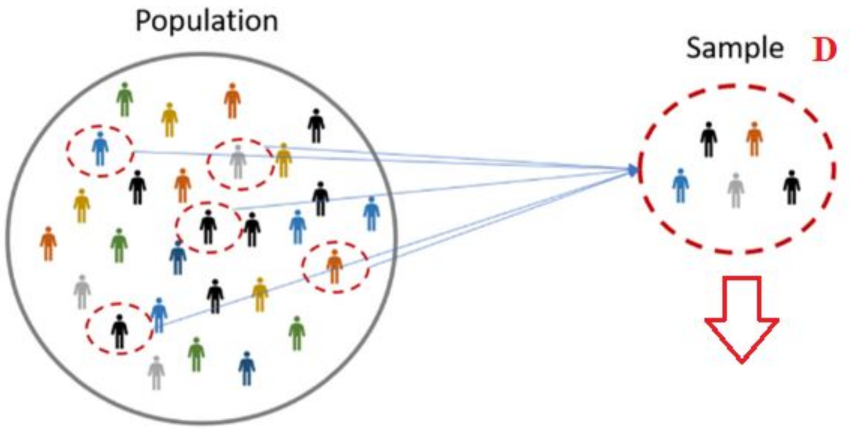
Logistic regression model, predict class probabilities

Gradient descent: find the weights minimizing cost



Model performance: finite samples

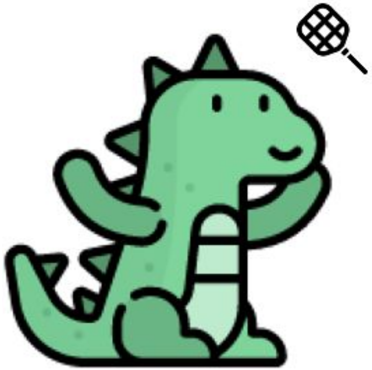
- Most datasets are **samples** taken from an **infinite** population.
- We are interested in modelling the **whole population**, we just have access to a **finite sample**.



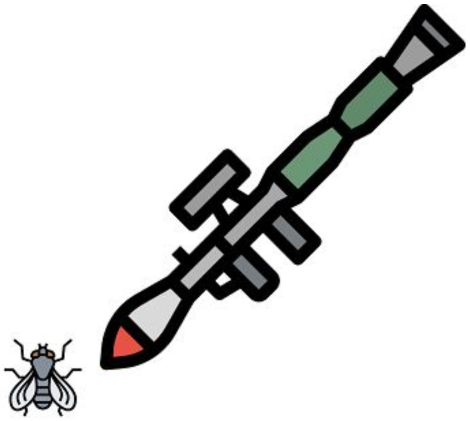
Source: Hotcubator

Underfitting and overfitting

- Simple models might not be able to fit the training data (**underfit**)
- Complex models fit the training data very well (**overfit**), but fail to **generalize** to new examples



Underfitting



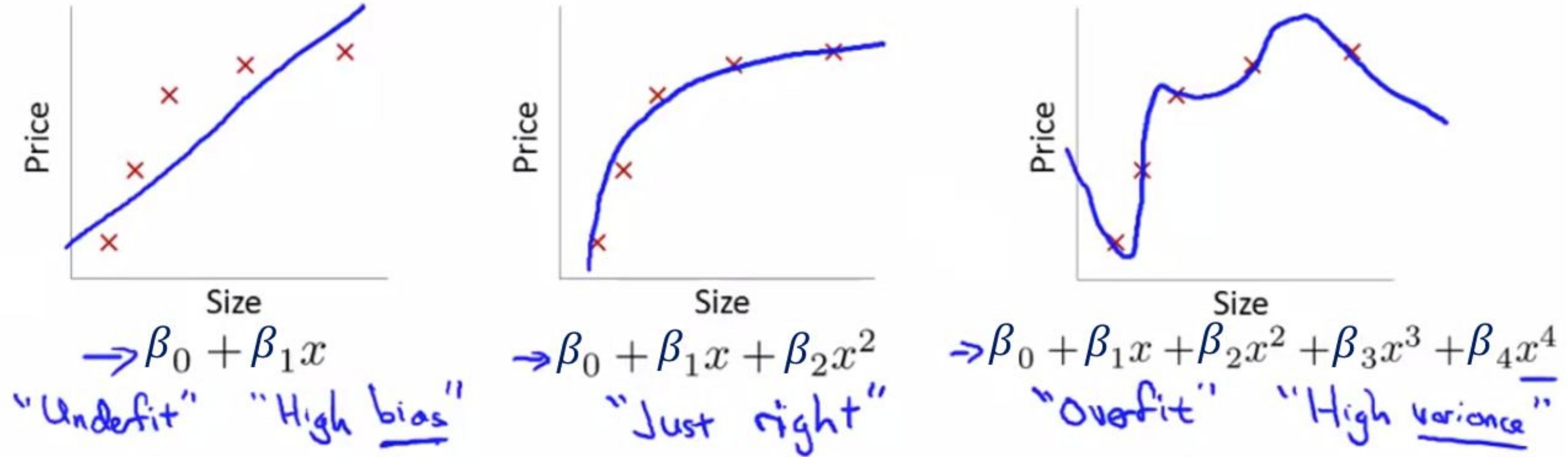
Overfitting

Image Credits: https://livebook.manning.com/

Underfitting and overfitting

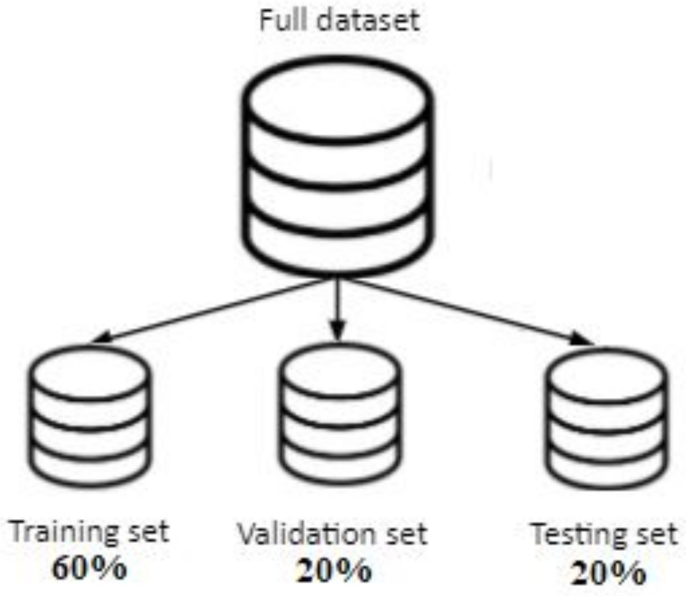
- Simple models might not be able to fit the training data (**underfit**)
- Complex models fit the training data very well (**overfit**), but fail to **generalize** to new examples

Source: Coursera



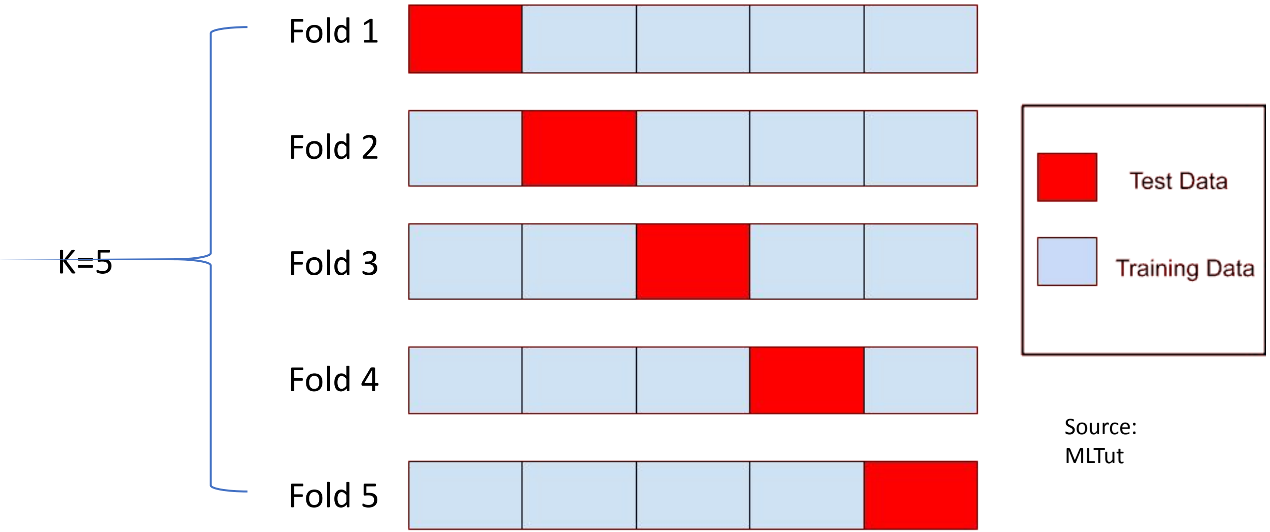
Model selection: Train/Validation/Test sets

1. **Fit** model **parameters** on **training** set
2. **Choose model**/hyperparameter configuration with **lower validation error**
3. **Evaluate** model performance with **testing** set



Model evaluation: k-fold cross-validation

- More **efficient** way to compute validation error if we have **little data**
- Average error over the k red portions → **Validation error**



Source:
MLTut

Performance metrics for classification

- In binary classification (Yes/no, 0/1), we use the **confusion matrix** which has 4 values:
  - **True positives:** positive examples classified as positive
  - **True negatives:** negative examples classified as negative
  - **False positives:** negative examples classified as positive
  - **False negatives:** positive examples classified as negative

		Class	
		A	B
Classified	A	TP	FP
	B	FN	TN

Credit: Robert West, EPFL

Accuracy: overview

- Represents the % of correctly predicted cases

$$A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$

- Good metric when
  - Classes are not skewed
  - Errors (FP, FN) have the same importance

Source: Avinash Pandey

Accuracy: skewed example

$$A = \frac{85}{100} = 0.85$$

Classifier 1

		Class: Fraud	Class: ¬Fraud
Classified	Fraud	5	10
	¬Fraud	5	80

Credit: Robert West, EPFL

$$A = \frac{90}{100} = 0.90$$

Always ¬Fraud

		Class: Fraud	Class: ¬Fraud
Classified	Fraud	0	0
	¬Fraud	10	90

Credit: Robert West, EPFL

Accuracy: errors with different importance

$$A = \frac{75}{100} = 0.75$$

Classifier 1

		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

Credit: RobertWest, EPFL

$$A = \frac{80}{100} = 0.80$$

Classifier 2

		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

Credit: RobertWest, EPFL

Precision, recall, F1-score

- **Precision**: What fraction of positive predictions are actually positive?

$$P = \frac{TP}{TP+FP}$$

- **Recall:** What fraction of positive examples did I recognize as such?

$$R = \frac{TP}{TP+FN}$$

- **F1-score:** Harmonic mean of precision and recall

$$\mathrm{F1} = 2\frac{P \cdot R}{P+R}$$

Naive Bayes

- Naive Bayes: **classification** algorithm based on Bayes rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Dataset (X, y) samples, where
  - y: class variable (play golf yes/no)
  - X: features or parameters
- **Given** a new sample $\boldsymbol{X}$ we want to **predict** class $\boldsymbol{y}$

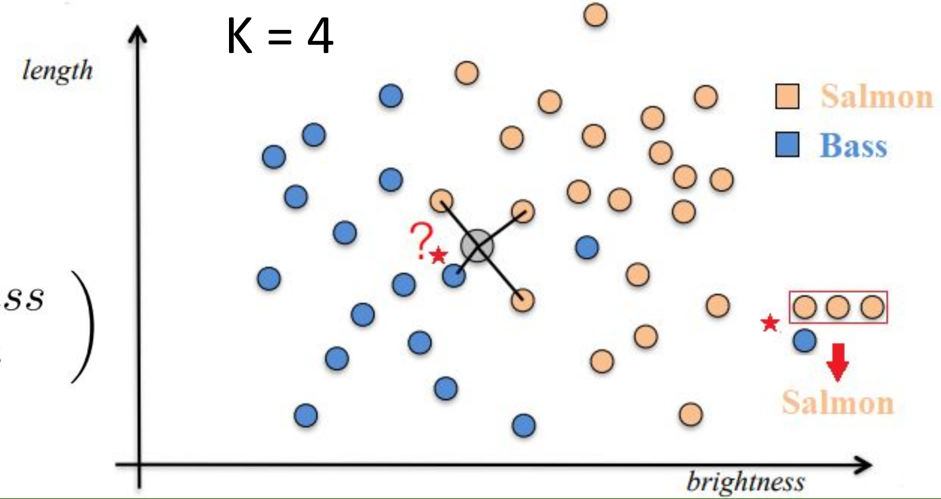
	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes

K-nearest neighbours

1. Given a new x, find its **k nearest neighbours** according to some distance measure
2. **Classify** the point according to the **majority of labels** of its nearest neighbours

Is it a bass or a salmon?

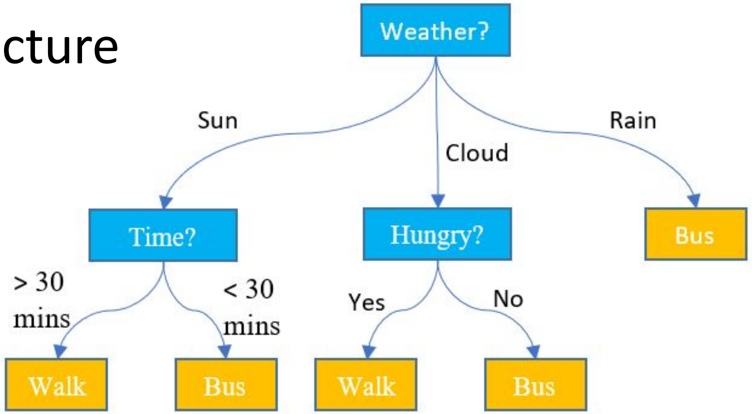
$$\xrightarrow{\text{Some algorithms}} \mathbf{x} = \begin{pmatrix} brightness \\ length \end{pmatrix}$$



Credit: Bob West, EPFL

Decision trees

- Supervised algorithm that can be used for **classification** or **regression** tasks.
- The model follows a flow-chart tree structure
  - Nodes are tests on a single attribute
  - Branches are attribute values
  - Leaves are class labels (classification) or output values (regression)



Source: SQLshack

WE WANT TO HEAR FROM YOU

Send us your feedback through [this form]