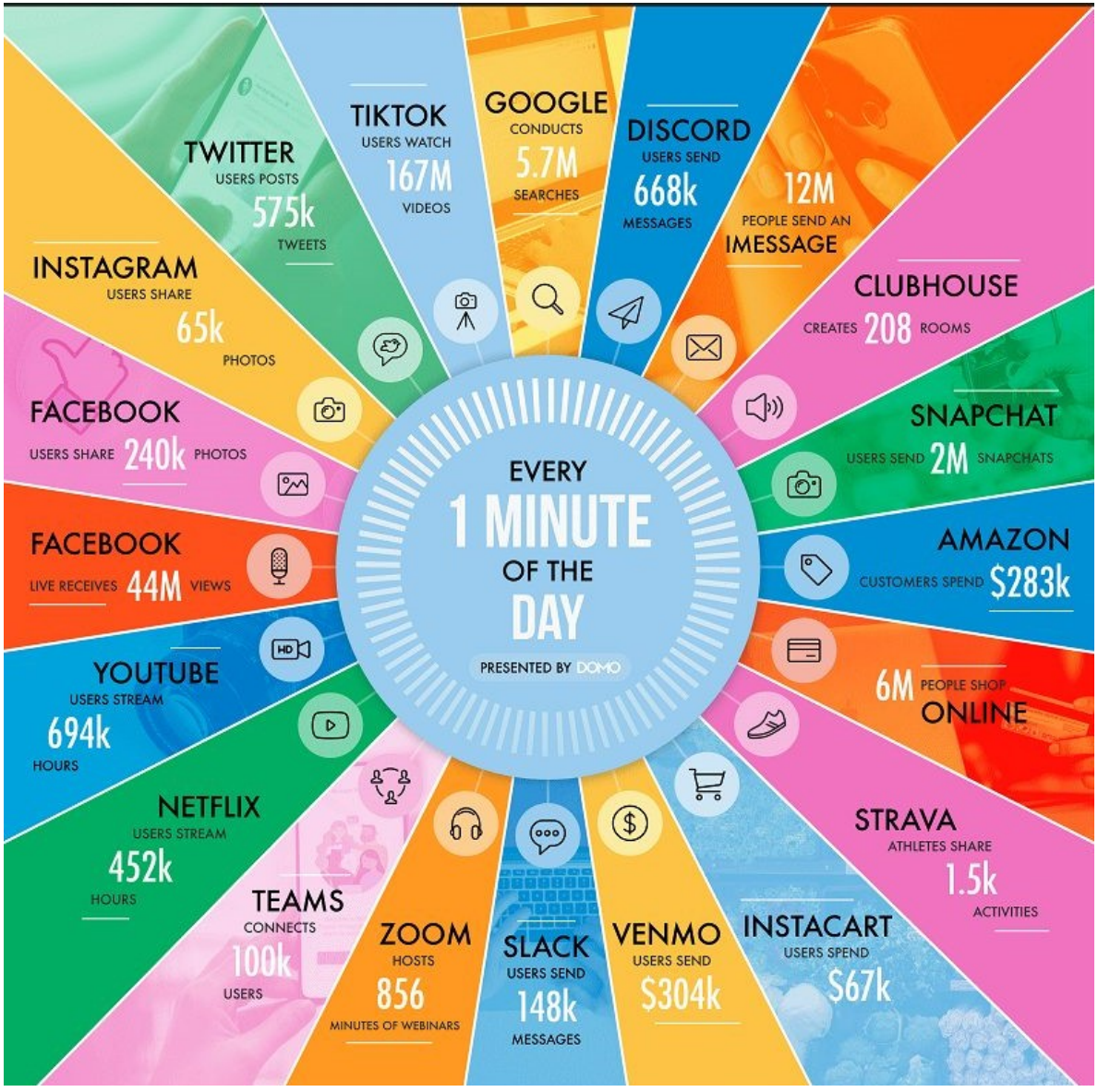
Machine Learning

Charlotte Pelletier

September 12, 2023

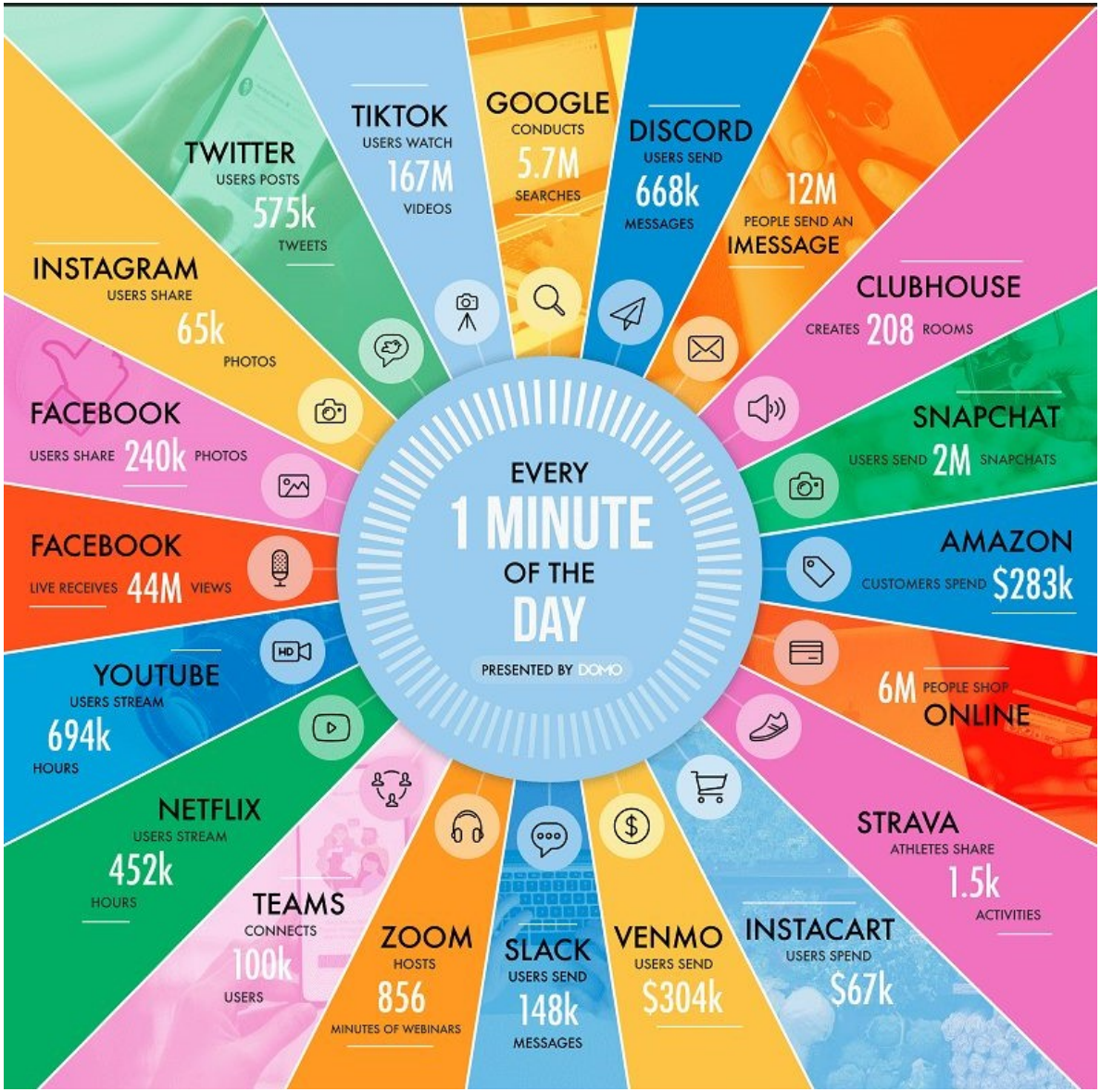
What happens on the Internet every 60 seconds



Source: `https://tekdeeps.com/this-is-what-happens-online-every-60-seconds-in-2021-infographic/`

And in Remote Sensing? > 10 TB of data acquired by Sentinel satellites every year

What happens on the Internet every 60 seconds



Source: `https://tekdeeps.com/this-is-what-happens-online-every-60-seconds-in-2021-infographic/`

And in Remote Sensing? $>$ 10 TB of data acquired by Sentinel satellites every year

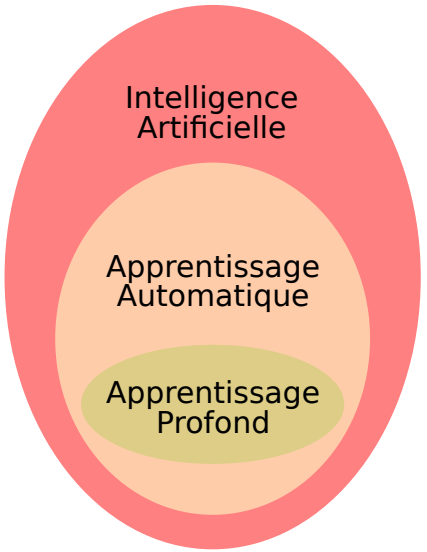
Introduction

This course is motivated by the abundant presence of data in geoscience and remote sensing, made possible by digitisation:

- image, text and time series data
- measurements and new sensor technologies
- social networks, etc.

You can find the content of this class under different namings depending on the scientific field you are working on:

- machine learning (ML)
  - statistics and computer science field
- pattern recognition (PR)
  - signal and image processing field
- artificial intelligence (AI)
  - mainstream



We will dig into the foundations of machine learning and some algorithms to understand how they work.

Introduction

This is an introductory course in machine learning with applications to remote sensing and geoscience data:

- data description and explortation, visualisation;
- discrimination and classification;
- regression and prediction.

It consists of foundationals and practices in, what we are broadly naming, artificial intelligence (AI). It has links with other modules, including:

- deep learning
- data mining and knowledge discovery
- image processing

Syllabus

The course consists of 30 hours of class and practical sessions.

Syllabus:

Introduction
- Overview
- Data analysis
- Validation and hyperparameter selection

I. Regression (Charlotte Pelletier)
- Linear regression
- Logistic regression
- Regularisation
- Evaluation

II. Decision trees and model averaging (Audrey Poterie)
- Decision trees
- Random Forests

III. Support Vector Machines and the kernel trick (Thomas Corpetti)
- SVM
- Kernels

Evaluation:

- a project-based evaluation (by group of 2 or 3) (40 %) with milestones
- a written exam (60 %) at the end of the module on the whole module

In practice

The practical labs will be performed in **Python**, using common libraries such as `Numpy`, `Pandas`, et `Scikit-Learn`.

Objectives: At the end of the module, you will be able to:

- understand and distinguish between the main categories of data problems
- study algorithms for labelling or predicting data
- grasp the complexity of certain problems and the mathematical tools needed to solve them

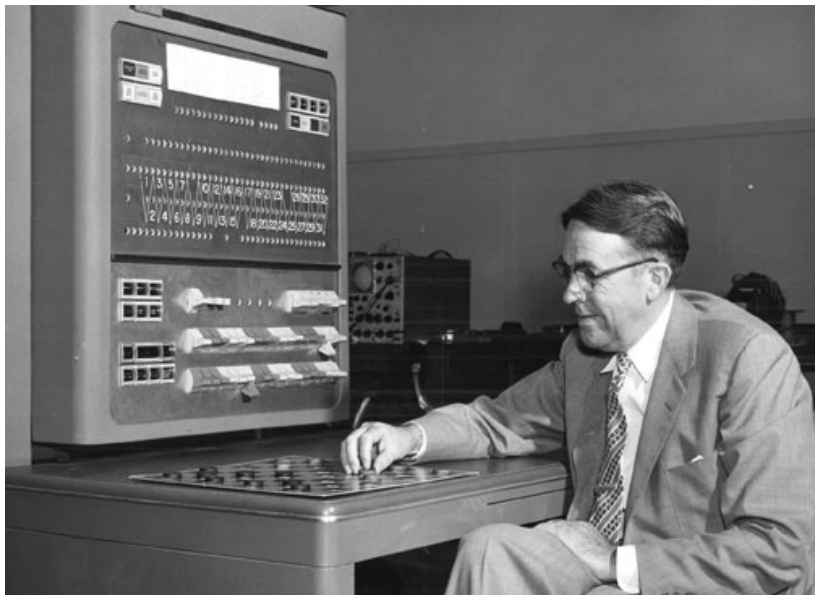
Section

Source : https://xkcd.com/1838/

What machine learning is?

Some definitions provided in the literature

- "Field of study that gives computers the ability to learn without being explicitly programmed" (*Samuel, 1959*).
- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (*Mitchel, 1997*).



What pattern recognition is?

Some definitions in the literature:

- "The assignment of a physical object or event to one of several pre-specified categories" (*Duda et Hart, 1973*).
- "Given some examples of complex signals and the correct decisions for them, make decisions automatically for a stream of future examples" (*Ripley, 1993*).
- The process of giving names w to observations w (*Schuermann, 1993*).
- "The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories" (*Bishop, 2006*)

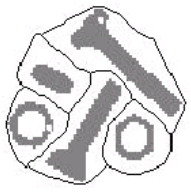
Common objective

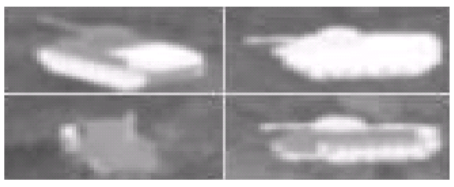
Enabling the machine to automatically process volumes of data (signals, images, etc.) to solve a given problem.

Examples

Vision

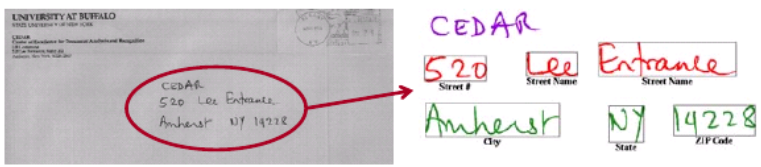
- Inspection of manufactured components
- Military detection





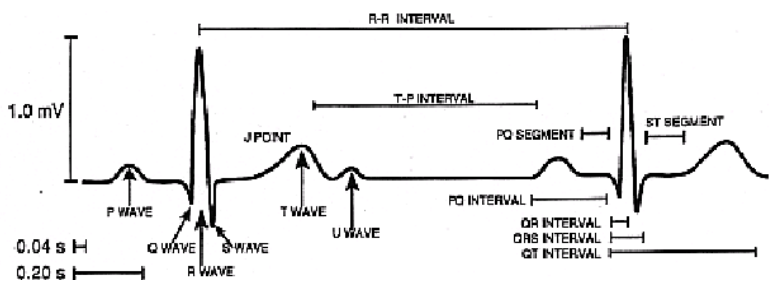
Character recognition

- Mail classification
- Cheque processing



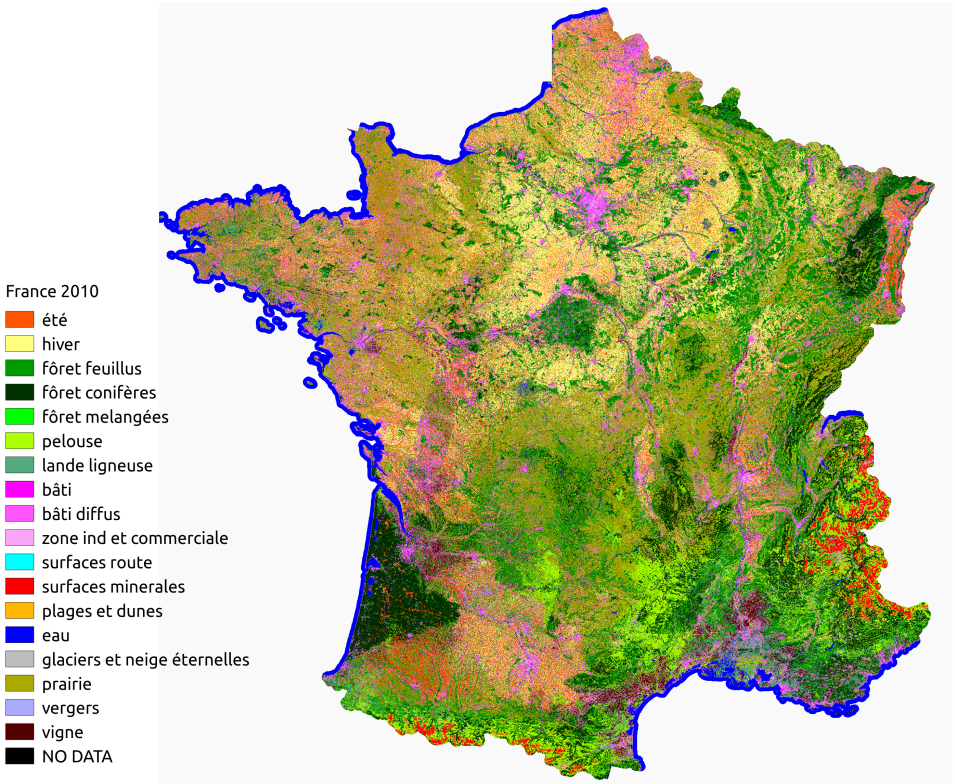
Diagnostic assistance

- Medical imagery: EEG, ECG, etc.
- To assist doctors



Specific examples: crop-type identification

French land cover land use map: produced yearly applying Random Forests to Sentinel-2 image time series[1]



[1] Credit: CESBIO THEIA CES OSO

Types of Problems

Unsupervised Learning: Understanding Data

- **Clustering**: Organizing objects into groups with a certain similarity (taxonomy of animal species).
- **Probability Density Estimation**: Estimating the probability distribution of training data (estimating the distribution of noise).
- **Dimensionality Reduction**: Reducing the dimensionality of data to better interpret/visualize it (recommendation).

Types of Problems

Unsupervised Learning: Understanding Data

- **Clustering**: Organizing objects into groups with a certain similarity (taxonomy of animal species).
- **Probability Density Estimation**: Estimating the probability distribution of training data (estimating the distribution of noise).
- **Dimensionality Reduction**: Reducing the dimensionality of data to better interpret/visualize it (recommendation).

Supervised Learning: Learning to Predict

- **Classification**: Assigning a class to an observation (character recognition, weather forecasting for rain).
- **Regression**: Predicting a real value based on an observation (temperature forecasting in weather).

Types of Problems

Unsupervised Learning: Understanding Data

- **Clustering**: Organizing objects into groups with a certain similarity (taxonomy of animal species).
- **Probability Density Estimation**: Estimating the probability distribution of training data (estimating the distribution of noise).
- **Dimensionality Reduction**: Reducing the dimensionality of data to better interpret/visualize it (recommendation).

Supervised Learning: Learning to Predict

- **Classification**: Assigning a class to an observation (character recognition, weather forecasting for rain).
- **Regression**: Predicting a real value based on an observation (temperature forecasting in weather).

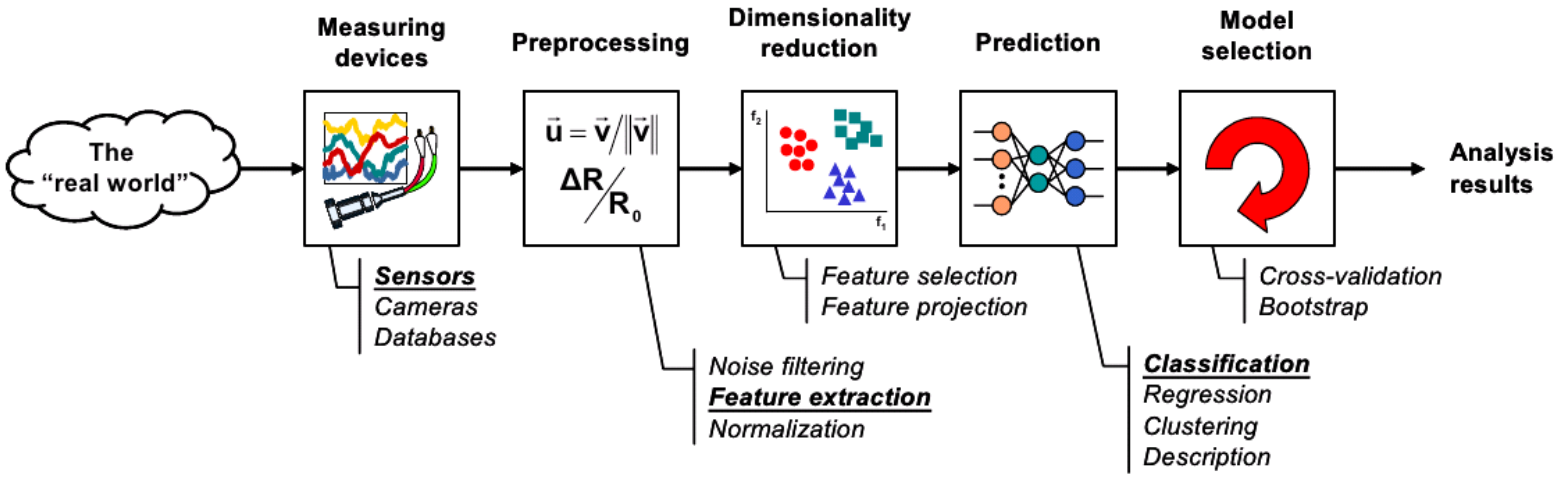
Reinforcement Learning: Learning Through Play

- Learning to maximize a reward (autonomous driving, games, control systems).

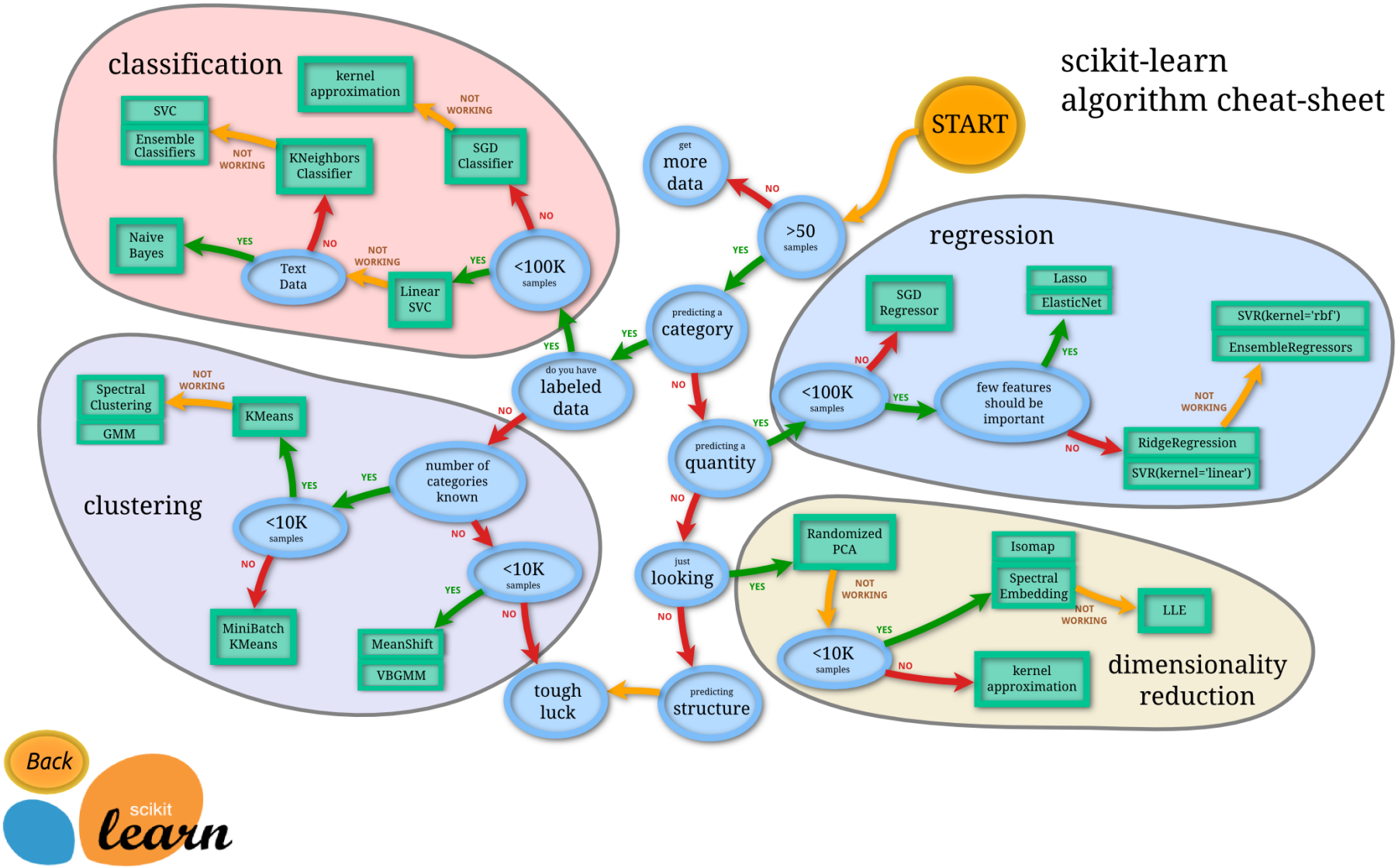
Components of a Machine Learning System

In practice, a typical system consists of

- an acquisition system (sensor, database, manual or automatic labeling)
- a set of data preprocessing (cleaning, formatting, conversion, normalization)
- a variable extraction system (manual or automatic extraction, selection, dimensionality reduction)
- an algorithm (clustering, classification, regression)



Finding Your Algorithm



Training Data

Unsupervised Learning

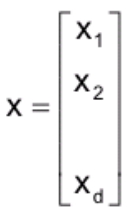
- $\mathbf{x} \in \mathbb{R}^d$ is an observation with d real variables.
- The training set is defined by observations $\{\mathbf{x}_i\}_{i=1}^m$ where m is the number of training examples (data points).
- Examples are often represented in the form of a matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ defined as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top$ containing training examples as rows and variables as columns.
- d and m define the dimensionality of the learning problem.

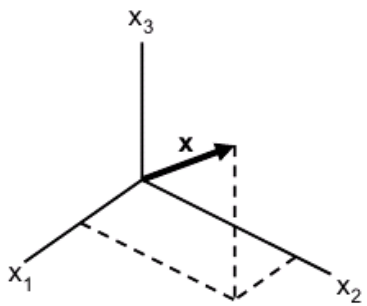
Supervised Learning

- Each observation $\mathbf{x}_i$ is associated with a value to predict $y_i \in \mathcal{Y}$.
- Just like observations, the values to predict (labels) can be concatenated into a vector $\mathbf{y} \in \mathcal{Y}^n$
- The space of values to predict $\mathcal{Y}$ is:
  - $\mathcal{Y} = \{-1, 1\}$ for binary classification or $\mathcal{Y} = \{1, \ldots, m\}$ for multi-class classification (m classes).
  - $\mathcal{Y} = \mathbb{R}$ for regression.

Variables and Shapes

- A variable is a distinctive trait or characteristic of an object. It can be symbolic (e.g., a color) or numeric (e.g., size).
- **Definition**
  - A combination of variables is represented using a vector $\mathbf{x}$ of dimension d.
  - The d-dimensional space is called the feature space ($\mathbb{R}^d$).
  - Objects are represented as points in the feature space. This representation is called a scatterplot.





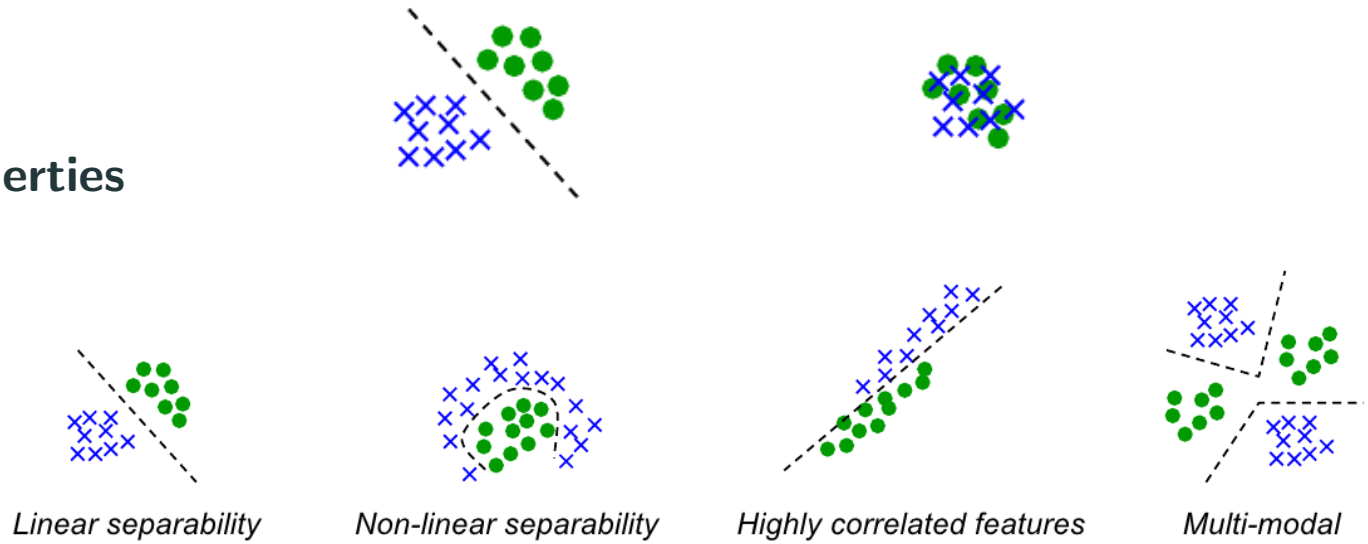
- A shape is a set of variables from a given observation. In discrimination problems, a shape consists of a vector of variables and a label.

What Makes a "Good" Variable?

The quality of a variable depends on the learning problem.

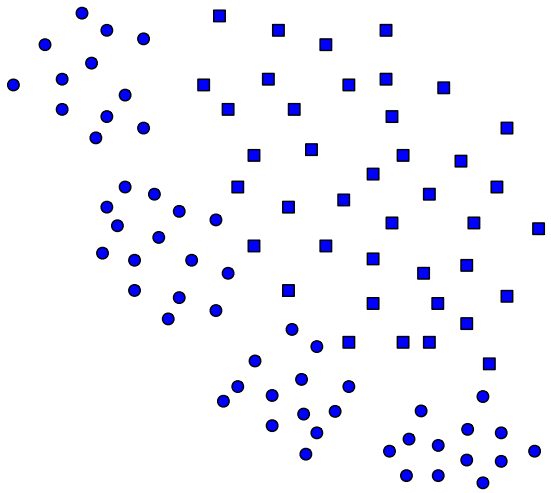
- **Discrimination**: Examples from the same class should have similar variables, while examples from different classes should have different variables.
- **Regression**: The variable should help in better predicting the value (it should be correlated with the values to predict).

Other Properties



Section

Data Description/Exploration



Consider a training set $\{\mathbf{x}_i\}_{i=1}^m$ composed of examples of dimension d.

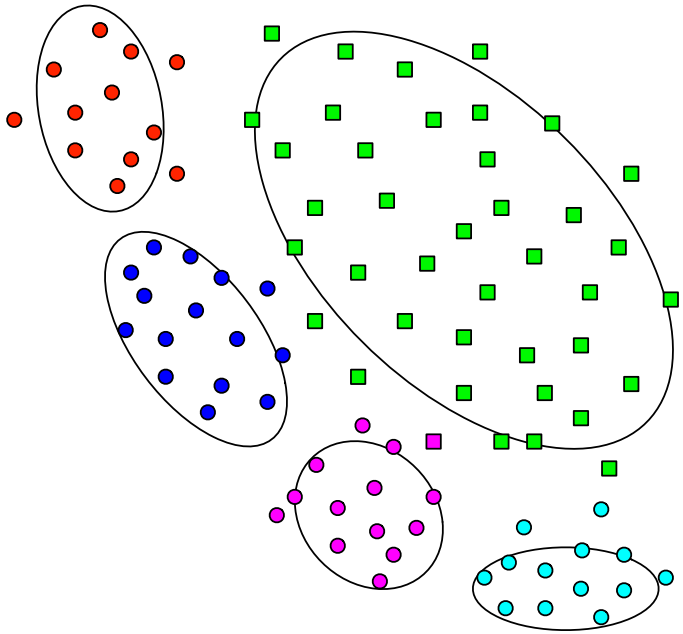
Objectives

- **Clustering** $\{\mathbf{x}_i\}_{i=1}^m \Rightarrow \{\hat{y}_i\}_{i=1}^m$ where $\hat{y}$ represents membership in a group.
- **Probability Density Estimation** $\{\mathbf{x}_i\}_{i=1}^m \Rightarrow p(\mathbf{x})$.
- **Dimensionality Reduction** $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^m \Rightarrow \{\tilde{\mathbf{x}}_i \in \mathbb{R}^{d'}\}_{i=1}^m$ with $d' \ll d$.

Clustering

Objective

- Organize the training examples into groups.
- $\{\mathbf{x}_i\}_{i=1}^n \Rightarrow \{\hat{y}_i\}_{i=1}^n$ where $\hat{y} \in \mathcal{Y}$ represents a group (cluster) $\{1, \ldots, m\}$
- Parameters:
  - m number of groups
  - Similarity measure (characterizing similarities between observations)



Methods

- k-means clustering.
- Gaussian mixture models.
- Hierarchical clustering.

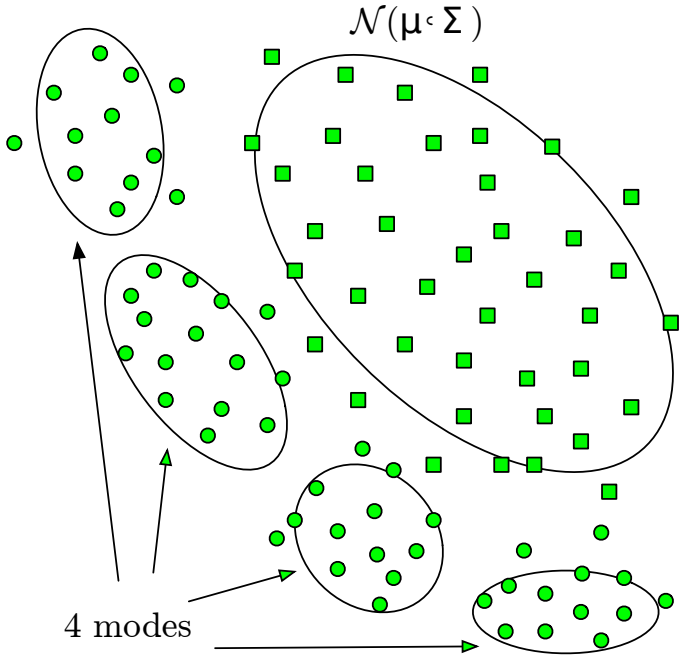
Examples

- Taxonomy of animals.
- Gene clustering.
- Social networks.

Probability Density Estimation

Objective

- Estimate the probability distribution of the data.
- $\{\mathbf{x}_i\}_{i=1}^m \Rightarrow p(\mathbf{x})$ where $p(\mathbf{x})$ is a probability density ($\int p(\mathbf{x})d\mathbf{x} = 1$)
- Model can be generative.
- Parameters:
  - Type of distribution (Gaussian, ...)
  - Distribution parameters (μ, Σ)



Methods

- parzen windows
- histogram
- Gaussian Mixture Models

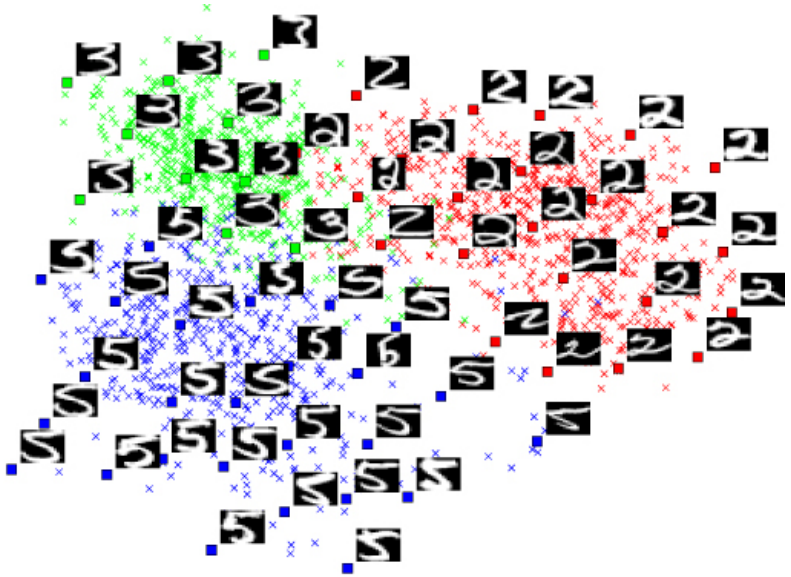
Examples

- noise estimation
- data generation (faces, ...)
- novelty detection

Dimensionality Reduction

Objective

- Project the data into a low-dimensional space.
- $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n \Rightarrow \{\tilde{\mathbf{x}}_i \in \mathbb{R}^{d'}\}_{i=1}^m$ with $d' \ll d$ (often $d' = 2$).
- Parameters:
  - Type of projection.
  - Similarity measure.



Methods

- Variable selection.
- Principal Component Analysis (PCA).
- Non-linear reduction.

Examples

- Data preprocessing.
- Vector visualization.
- Data interpretation.
- Recommendation systems.

Section

Prediction

Consider a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ composed of n observations $\mathbf{x}_i \in \mathbb{R}^d$ of dimension d and target values $y_i \in \mathcal{Y}$.

Objective

- We aim to learn from the training data a prediction function $f(\cdot) : \mathbb{R}^d \to \mathcal{Y}$.
- Types of predictions:
  - **Classification**
    $f(\cdot)$ predicts a class / category (discrete output), either in binary classification $\mathcal{Y} = \{-1, 1\}$ or multiclass $\mathcal{Y} = \{1, \ldots, m\}$.
  - **Regression**
    $f(\cdot)$ predicts a real value ($\mathcal{Y} = \mathbb{R}$).

Linear Function

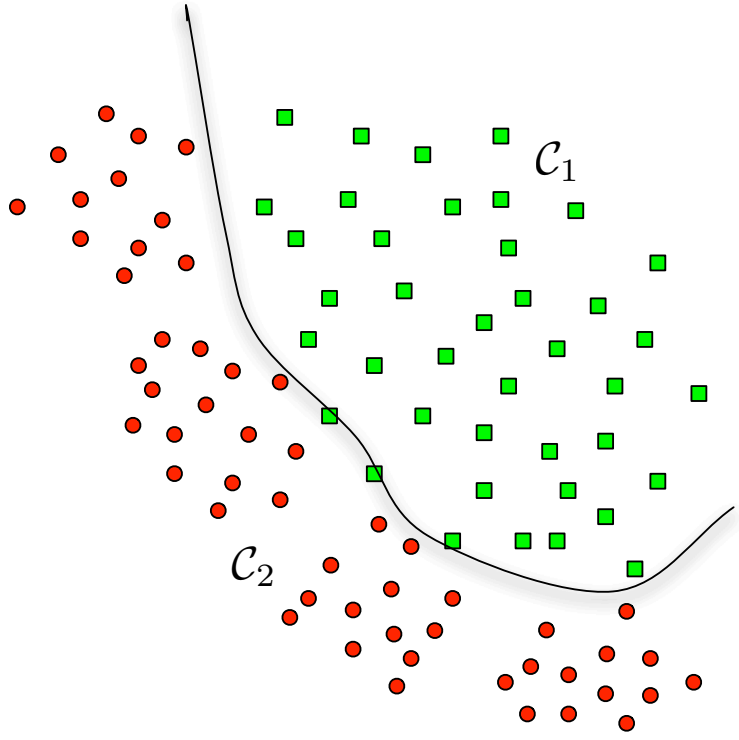
$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \mathbf{w}^\top \mathbf{x} + b$$

parameterized by $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$

Binary Classification

Objective

- Learn a function that predicts either class -1 or 1.
- $\{\mathbf{x}_i, y_i\}_{i=1}^m \Rightarrow f(\mathbf{x})$.
- Prediction: sign of $f(\cdot)$
- $f(\mathbf{x}) = 0$: decision boundary.
- Parameters:
  - Types of functions
  - Performance measurement



Methods

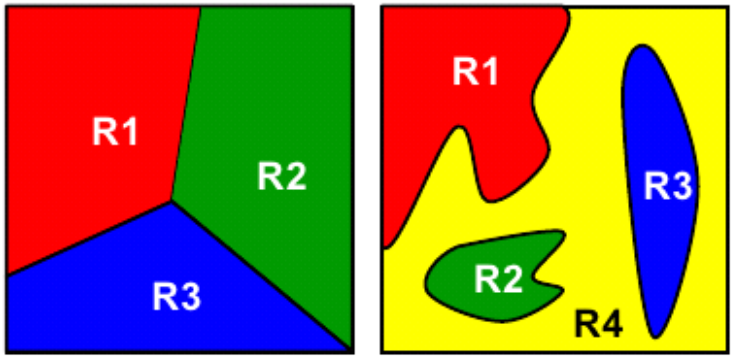
- Bayesian Methods
- Linear Discriminant Classifier
- Support Vector Machine (SVM)
- Decision Trees

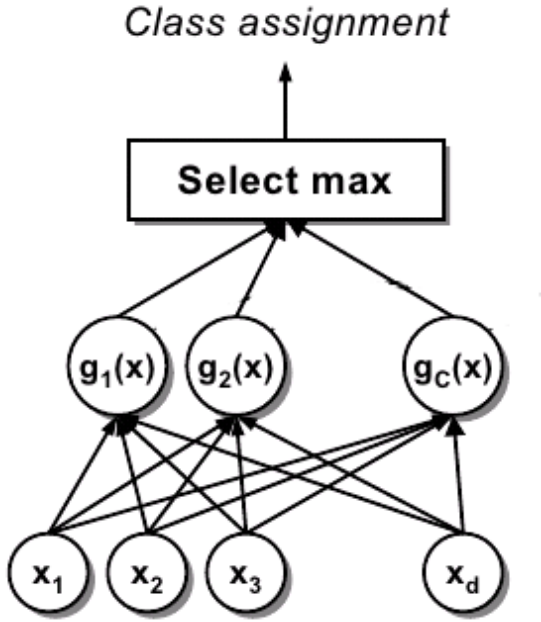
Examples

- Character Recognition.
- Diagnostic Assistance.
- Parts Inspection.
- Weather (Rain) Prediction.

Multiclass Classification

- The role of a classifier is to partition the space of variables into multiple regions to which classes are assigned.
  - The boundaries are called decision boundaries.
  - Classifying a vector of variables $\mathbf{x}$ involves determining which region it belongs to and assigning it the label of that region.
- The classifier can be represented by a set of discriminant functions: the classifier assigns $\mathbf{x}$ to class j if $g_j(x) > g_i(x)$ for all $i \neq j$.

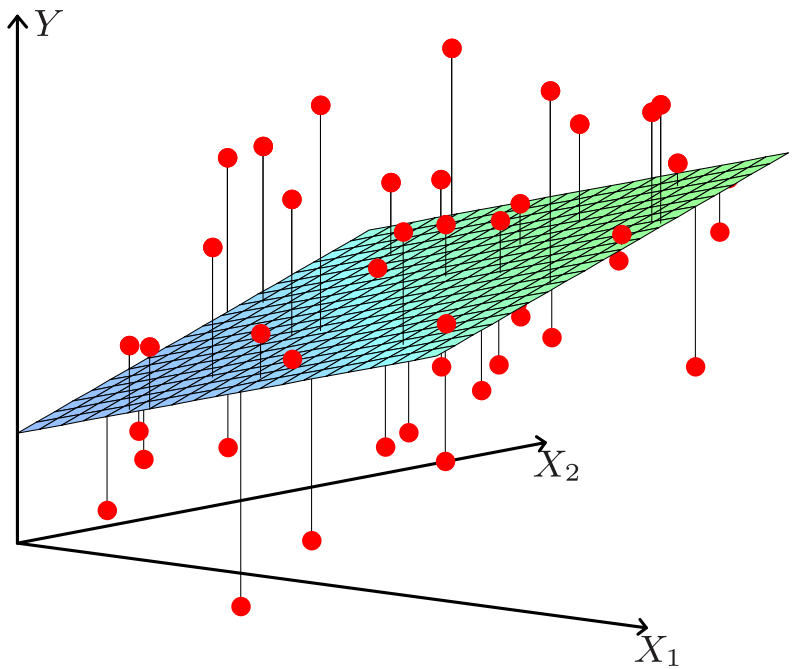




Regression

Objective

- Learn a function that predicts a real value.
- $\{\mathbf{x}_i, y_i\}_{i=1}^{n} \Rightarrow f(\mathbf{x})$.
- Parameters:
  - Type of function.
  - Performance measurement.
  - Prediction error.



Methods

- Least Squares.
- Ridge Regression.
- Kernel Regression.

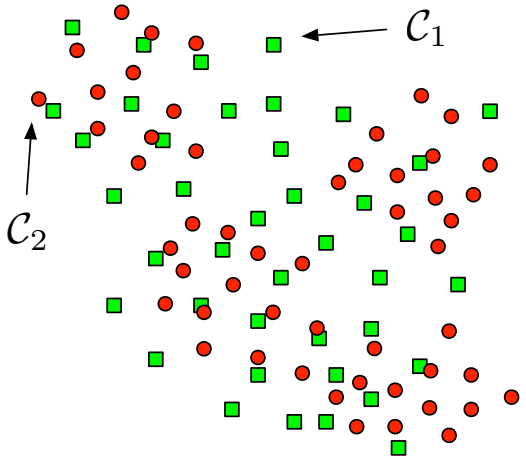
Examples

- Motion Prediction.
- Cholesterol Level Prediction.
- Weather (Temperature) Prediction.

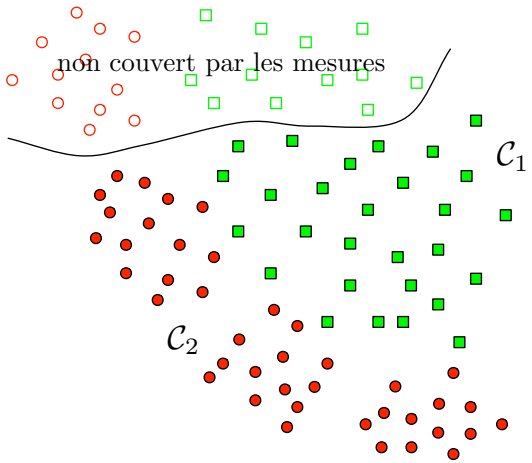
Section

Real Data (1)

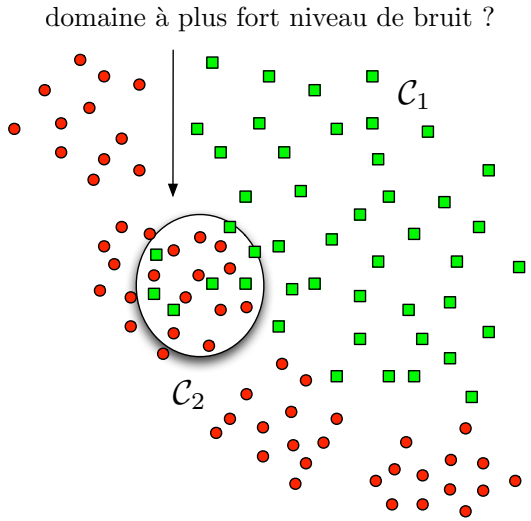
- Inadequate



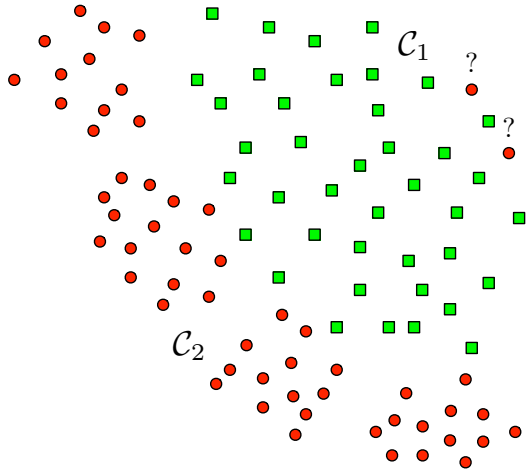
- Non-representative



- Contaminated by Noise



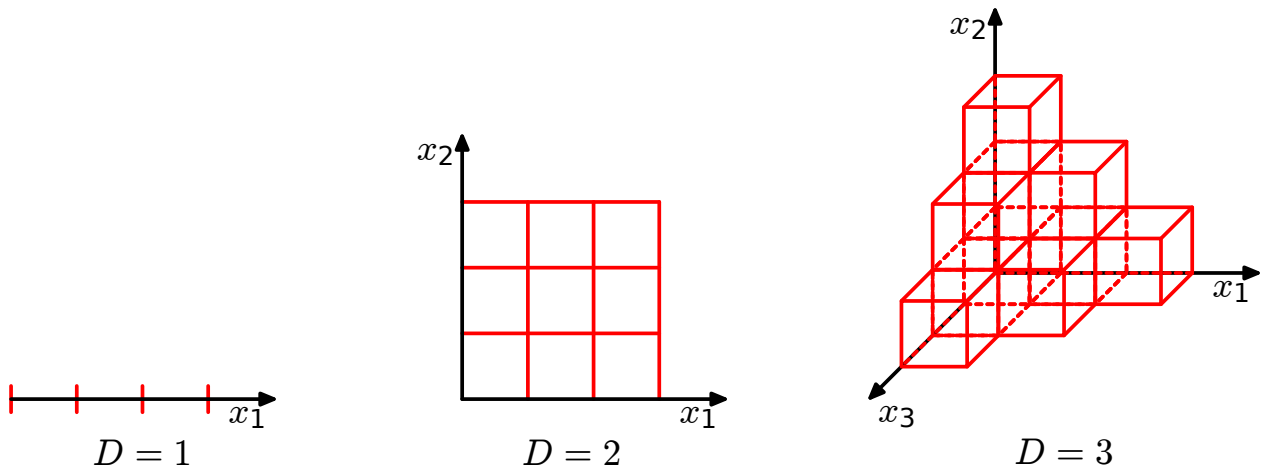
- Outliers



Real Data (2)

Dataset Size
We always have a finite number n of training points.

Curse of Dimensionality

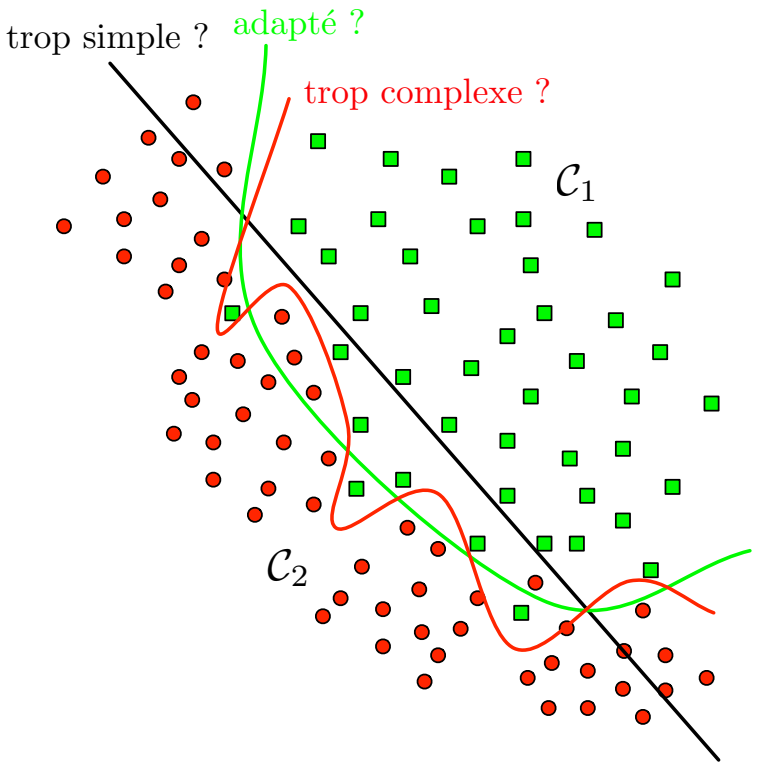


The curse of dimensionality expresses the fact that the number of data must grow exponentially with dimension to maintain an equivalent density.

How to Choose?

Model	Learning	Prediction
Too Simple	- -	- -
Appropriate	+	+
Too Complex	++	- -

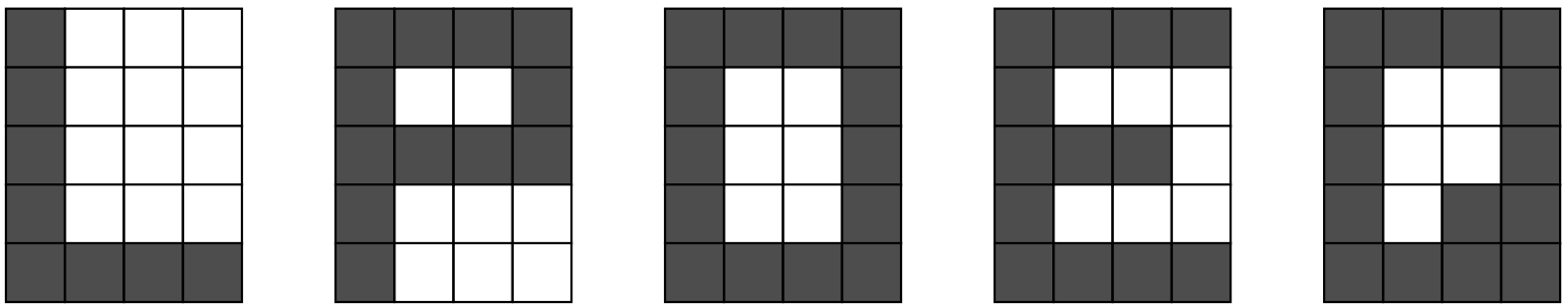
- A model that is too complex leads to overfitting.
- We aim to learn to predict!



Validation

- Splitting data into training/validation sets.
- Maximizing performance on validation data.
- Validation requires a good performance measure.

An Example of Shape Recognition Task



- Develop an algorithm to discriminate uppercase letters L, P, O, E, Q.
  - Determine a set of variables.
  - Propose a classification method based on a binary tree.

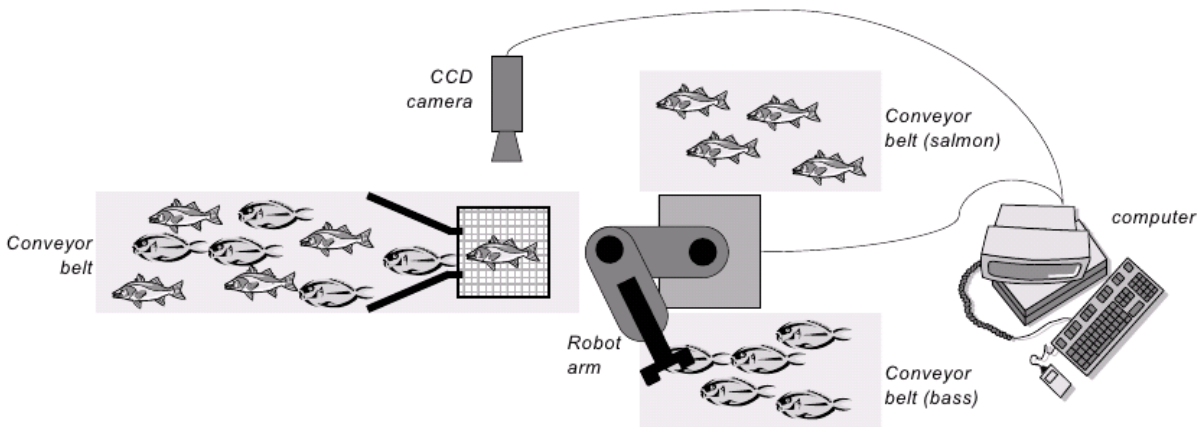
Design Cycle (1)

- Data Collection
  - Tedious and time-consuming but essential.
  - How many examples are sufficient?
- Variable Selection
  - Critical.
  - Can be constructed manually based on prior knowledge or automatically.
- Classifier Selection
  - Which model to choose?
  - How to adjust its parameters?
- Training
  - Train the model to perform well on training data.
- Evaluation
  - Is my model good?
  - Dilemma: Overfitting vs. Generalization.

Design Cycle (2)

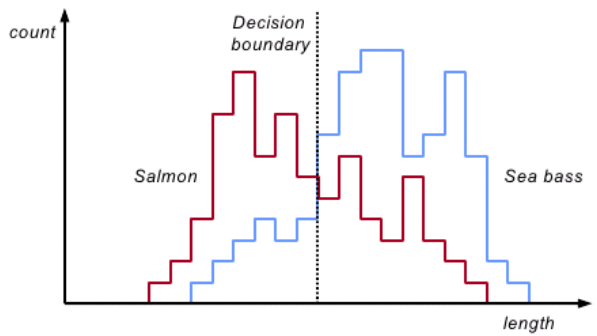
Scenario

- A fishery is looking to develop a computer vision system for automatic sorting of fish based on their types (salmon or sea bass).
- The system consists of:
  - A conveyor belt for transporting fish.
  - 2 conveyor belts for transporting the two species of fish.
  - A robotic arm for sorting.
  - A vision system.
  - A computer for image analysis and controlling the robotic arm based on decisions.



Design Cycle (3)

- Sensor
  - The vision system captures an image of a fish arriving on the sorting system.
- Image Processing
  - Adjustment of grayscale levels.
  - Segmentation to separate the fish from the background of the image.
- Feature Extraction
  - It is assumed that, on average, sea bass are longer than salmon.
  - From the segmented image, the length of the fish is estimated.
- Discrimination
  - Collect specimens of fish from both classes.
  - Plot length histograms for both classes.
  - Choose a length threshold to minimize the classification error.
  - Achieve a disappointing score of 40
  - What's next?

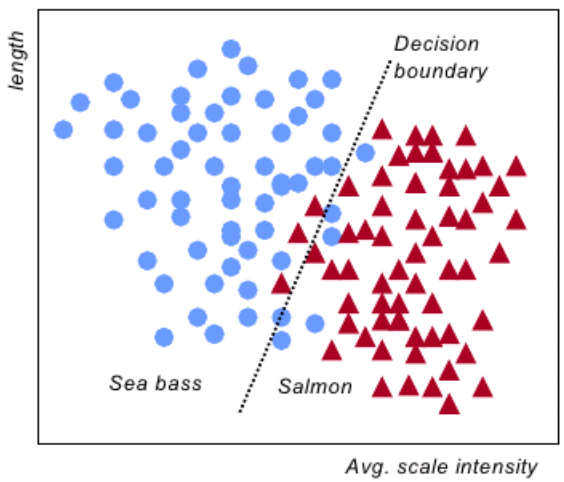


Design Cycle (4)

System Improvement

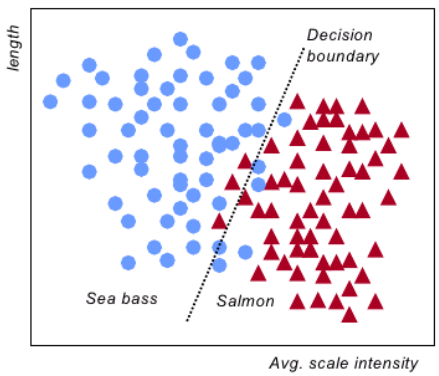
- Aiming for a recognition rate of 95
  - Width, area, eye position relative to the mouth, etc.
  - Variables that do not carry discriminative information.
- Finally, we found a "good" variable: the average grayscale level of the scales.

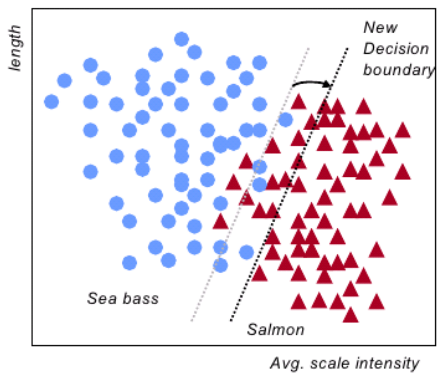
- We combine "length" and "grayscale level" to improve the separability of the classes.
- We calculate a linear decision function to separate the two classes and achieve a recognition rate of 95.7



Cost and Recognition Rate

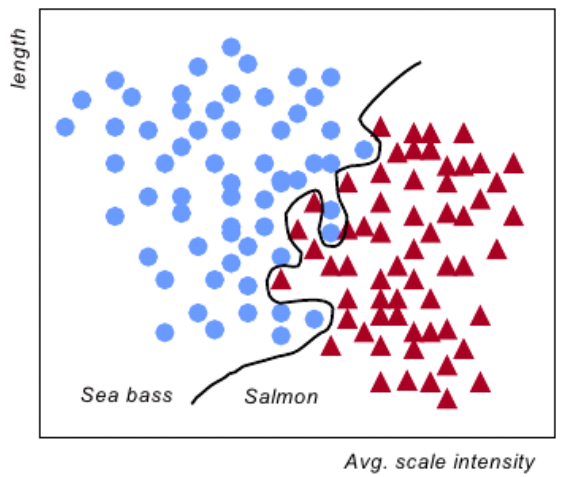
- Our classifier was constructed to minimize the error in discrimination.
- Is this the best choice for our fishery?
  - The cost of classifying a salmon as a sea bass is that the end customer experiences a "good" salmon taste when they buy sea bass.
  - The cost of classifying a sea bass as a salmon is the customer's dissatisfaction with buying sea bass at the salmon price.
  - The costs of misclassification can be different.
- Intuitively, we would like to take this cost into account when constructing our decision boundary.





Generalization

- Our system meets the specifications with a recognition rate of 95.7%.
- By further improving the system through the use of a method allowing a non-linear decision function, we achieve a rate of 99.9975% with the following decision function:



- Satisfied, we deploy our system in the processing plant. However, a few weeks later, the plant manager contacted us to report that in practice, the system correctly recognizes only 75% of the fish.
- Where did we go wrong?

Section

Not Just a Matter of Terminology

Statistical Models	Machine Learning
data points	samples
variables	features
parameters	weights
estimation/fitting	learning
regression/classification	supervised learning
clustering/density estimation	unsupervised learning
response	label
performance	generalization

To delve deeper: `https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3`