ECS171: Machine Learning

L2 Introduction to Machine Learning: data, models, and hypotheses

Instructor: Prof. Maike Sonnewald
TAs: Pu Sun & Devashree Kataria

MOCO Amsterdam garden

Intended Learning Outcomes

- Part 1 (Tuesday)
    - Know the meaning of key terms (denoted in blue)
    - Describe attributes of data including with statistical concepts
    - Describe and apply the data pre-processing methods and describe their limitations and assumptions
- Part 2 (Thursday)
    - Apply methods from class including PCA, regression, classification, outlier detection (statistical and otherwise)
    - Understand the different approaches to learning for different tasks and what questions are relevant for different methods including supervised, unsupervised, reinforcement and association rule approaches
    - Use pre-processing methods for data and appreciate limitations of various methods.

Rec. reading: B 1.1, R 1-7, R 3.4.6

Practical in-lecture examples on Thursday

Computers, algorithms and progress

Antikythera mechanism
2nd century BC

- First ‘analogue computer’
- Computed stars and astronomical events

Ada Lovelace (1815-1852)

- First computer program: algorithm designed to done by a machine.
- Visionary for capability of computers to go beyond number-crunching

ENIAC (Electronic Numerical Integrator and Computer), c. 1946

- First programmable general-purpose electronic digital computer, built during World War II by the United States.

Images: Britanica

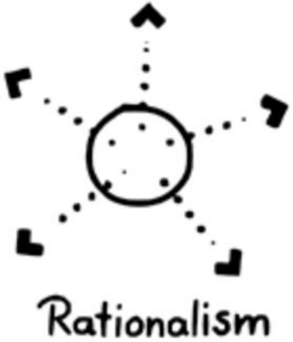
More on computers

Machine Learning and the theory of knowledge:

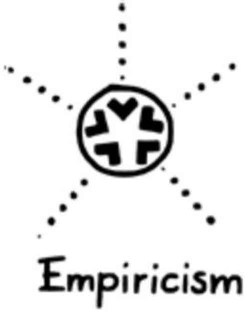
What can we know?

How can we know it?

René Descartes (Fr. 1600s)



David Hume (Scotland. 1700s)



Machine Learning and the theory of knowledge:

What can we know?

How can we know it?

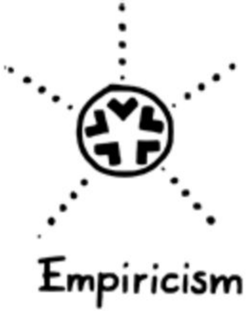
René Descartes (Fr. 1600s)

Rationalism

Knowledge begins with the senses but ends with reason

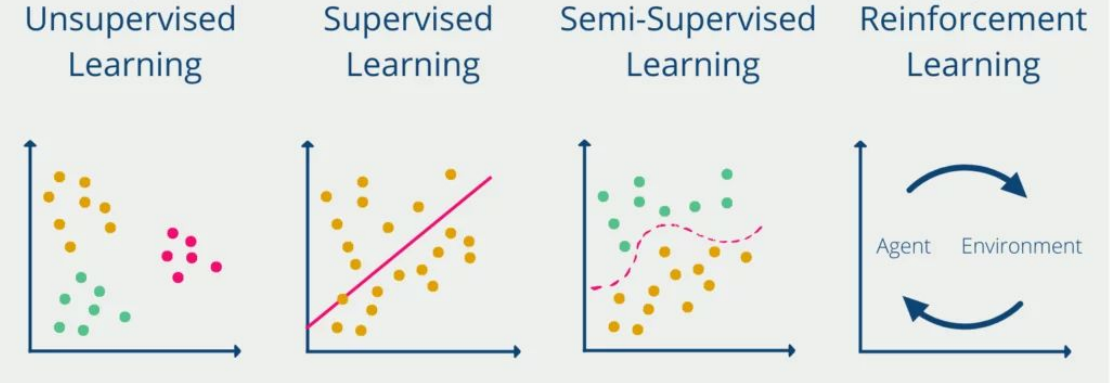
Immanuel Kant (Ger. 1700s)

David Hume (Scotland. 1700s)



Machine Learning (ML) is a subfield of CS and AI

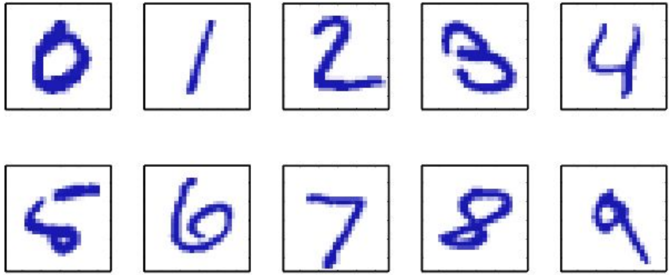
- Make algorithms and statistical models to automatically analyze and draw inferences from **patterns in data.**
- Enables systems to derive meaning from huge volume and heterogeneous, complex data
- The process of building computer algorithms (i.e., models) to identify patterns (i.e., to understand data)
- Provides the mechanisms of learning from data that is learning a function (or algorithm) that maps input variables (X) to output variables (Y).

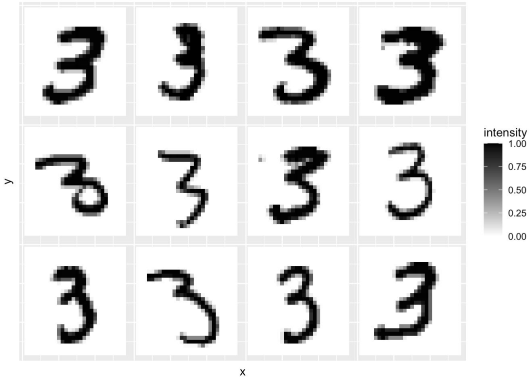


ML does not require pre-defined rules

Example of rule-based non-ML approach

- Goal: Machine for recognize handwritten digits 0-9 (e.g. MNIST dataset)
- Each a 28x28 pixel image
- Non-trivial due to large variation in writing
- One could use handcrafted rule heuristics to distinguish digits…
- Impractical due to numbers of rules and exceptions:
    - This machine gives poor results
- Machine learning techniques learn the ‘rules’





ML extracts ‘rules’ automatically

- Use a *training set* (observations) of N digits $\{x_1, \ldots, x_N\}$ from the pictures to ‘tune’ the parameters of an adaptive ML model
- Classes 0-9 of the N digits in the training set are known in advance, and the classes of each digit is expressed in a *target vector* **t** that stores the correct answer (aka *label*) the ML model should give
- Task: find an ML model for the classes. The ML model can be described as f(x), where a new (or ‘unseen’) digit x is the *input* and the resulting y is the predicted digit.
- The form of function f is determined during a *training phase* or *learning phase* using the training set
- How good the performance of the ML model is (how often it predicts the correct category) is measured by how well new examples of N (with no pre-given category) are predicted. This ability is called *generalisation*

ML models are used to search for patterns

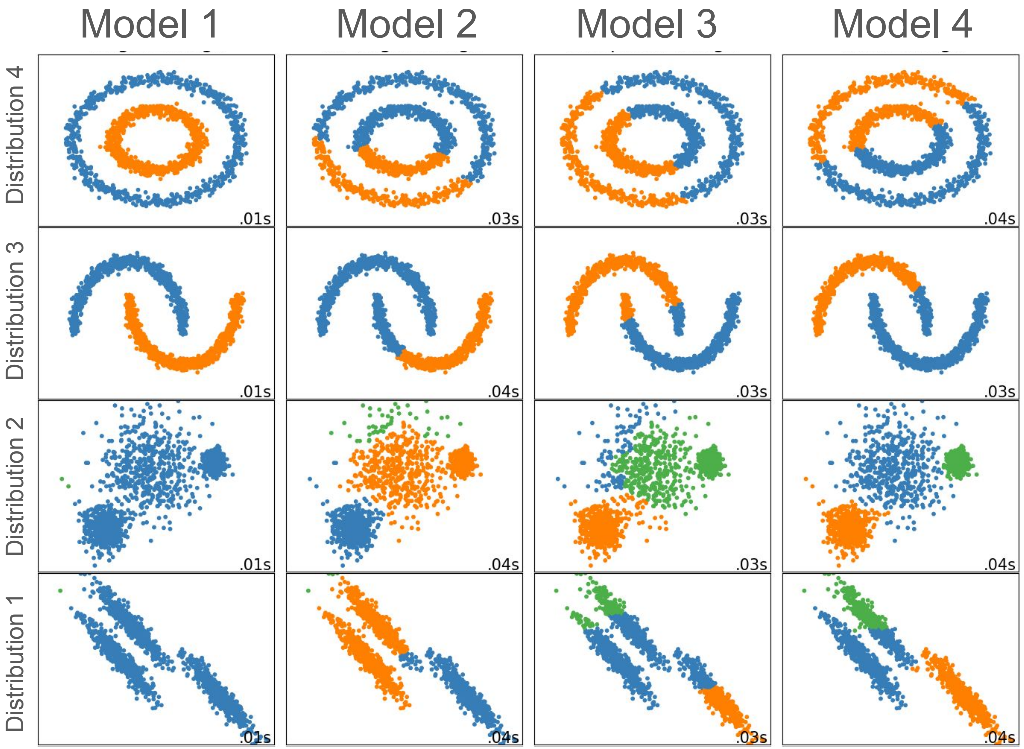
Different models are good at different data distributions

Humans are good at finding patterns in 1, 2 and, 3 dimensions

ML is used in arbitrary dimensions

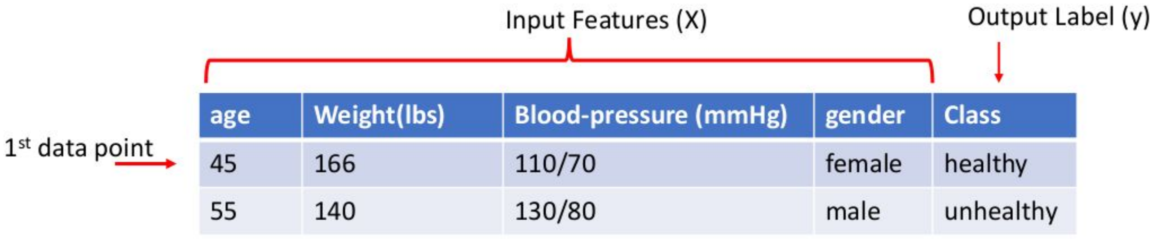
Need to find a 'good' model that is able to find the interesting areas

Determining when a good fit is found is critical



A dataset example

- Data set is a collection of observations or data points that are organized in a table or matrix.
    - Each row represents a data point (i.e., observation), and each column represents an attribute or feature associated with that datapoint
- A dataset can be described in terms of size (i.e., number of rows), attributes, type of attributes, statistical distribution, and correlation among the attributes.
- These terms can be used interchangeably:
    - Observation, example, instance, row, data point
    - Independent variable, predictor, attribute, property, feature
    - Dependent variable, target, class, label



Input Features (X)

Output Label (y)

1st data point

age	Weight(lbs)	Blood-pressure (mmHg)	gender	Class
45	166	110/70	female	healthy
55	140	130/80	male	unhealthy

Data is central to Machine Learning

- Training set is used for training a model
- Test set is used to test the performance of the model
- Observation (or instance), features (or predictors or attributes), target (or class or label).
- Example: aquarium fish
  - Attributes include: name, length, weight, colour, number of finns, tail size…
  - Attributes (or features) can be seen as high-level translations

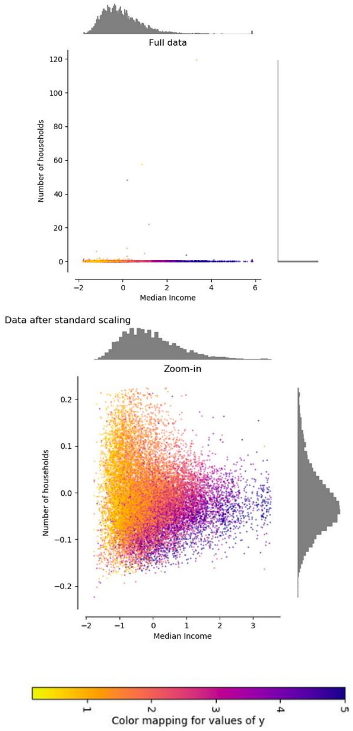
Nano Fish
Swordbill
Guppy
Clown Loach
Betta Fish
Dwarf Gourami
Clownfish
Angelfish
Goldenfish
Cobue Cichlid
Polka Dot Catfish
Black Tetra
VectorStock®
VectorStock.com/17942006

Quality of available data is central to good performance

- An ML model for a task can only succeed if good data is present
- Completeness:
  - The dataset contains all relevant features or observations needed for a given task
  - Incomplete datasets can also have missing labels or sparsity in attribute values
- Size:
  - Sufficient observations must be present
- Validity:
  - Datasets must contain accurate labels (unlabeled data sets exist, labels needed for supervised learning), clean data (outliers and errors), data relevant to the problem

Preprocessing data can improve generalisation

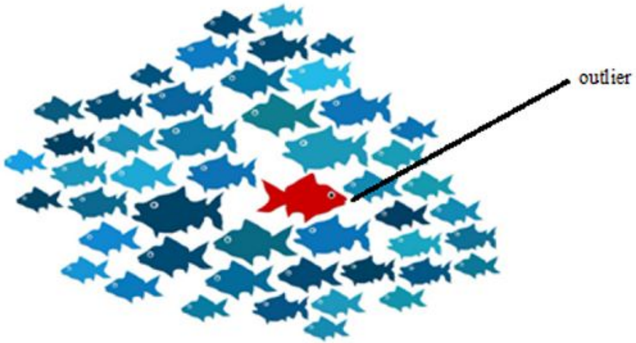
- Preprocessing is taking the variables and transforming them into a new space, where patterns are easier to determine
- Data preprocessing includes:
    - Cleaning (missing data, noisy data)
    - Transformation (scaling, attribute selection, discretization, concept hierarchy generation)
    - Reduction (Aggregation, subset selection, dimensionality reduction)
- The new and unseen data must be preprocessed similarly
- Need to be careful not to discard important information!



scikitlearn

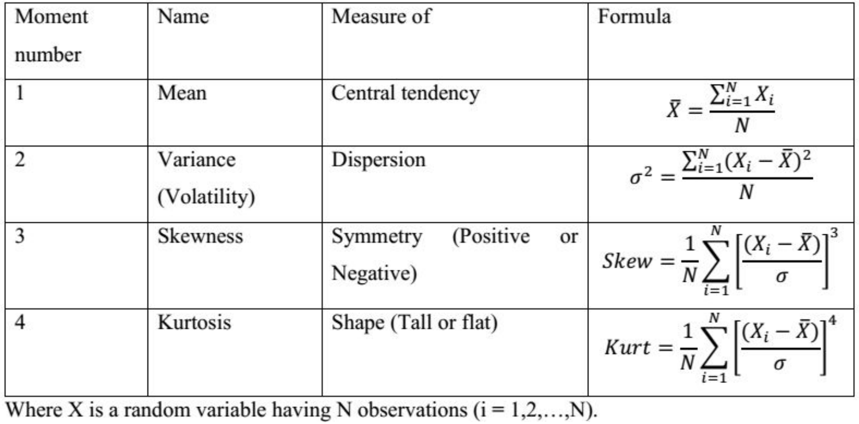
Noisy data: Outlier example

- Outliers in a dataset are those samples that show abnormal distance from the other samples.
- Outliers can affect the overall accuracy of the model trained on the data.
- Methods to detect outliers include:
  - Visualization (such as scatter plot)
  - Data Analysis (statistical approach, box plot)
  - ML algorithms such as One-Class-SVM, Local outlier factor, Isolation Forest



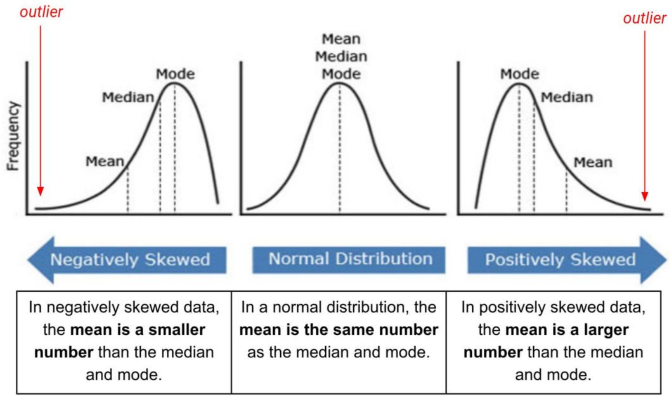
Preprocessing: Statistics to Analyze Distribution of Data

- The *Mean* is the average of the data
- The *Median* is the value in the middle of the dataset (50% of the values smaller/larger or equal)
- The *Standard Deviation* (σ, square root of *variance*) measures the dispersion of data relative to its mean
- *Skewness* is a measure of the symmetry
- *Kurtosis* is the shape (tall, flat etc)
- Example: Assume that each observation has one single attribute x. Count the frequency of occurrence



Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).



In negatively skewed data, the **mean is a smaller number** than the median and mode.	In a normal distribution, the **mean is the same number** as the median and mode.	In positively skewed data, the **mean is a larger number** than the median and mode.

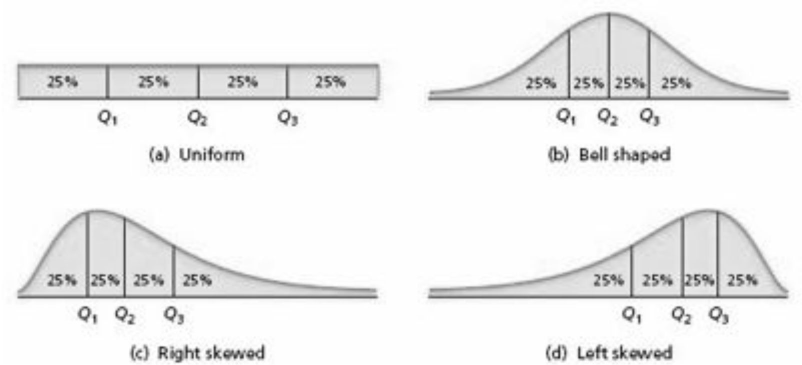
Preprocessing: Impact of Skewed Data in Building ML Models

- Direction of outliers
    - Right skewed data has most of the outliers on the right side of the distribution.
- Model selection depends on understanding the distribution of data.
    - Linear models work on this assumption that distribution of attributes and class variables are similar.
- ML models work better for the portion of data according to the direction of the distribution.
    - In a right-skewed distributed data, the trained model better represents the population distributed on the right-side of the distribution.
- Data that is different from a normal distribution is *non-linear*

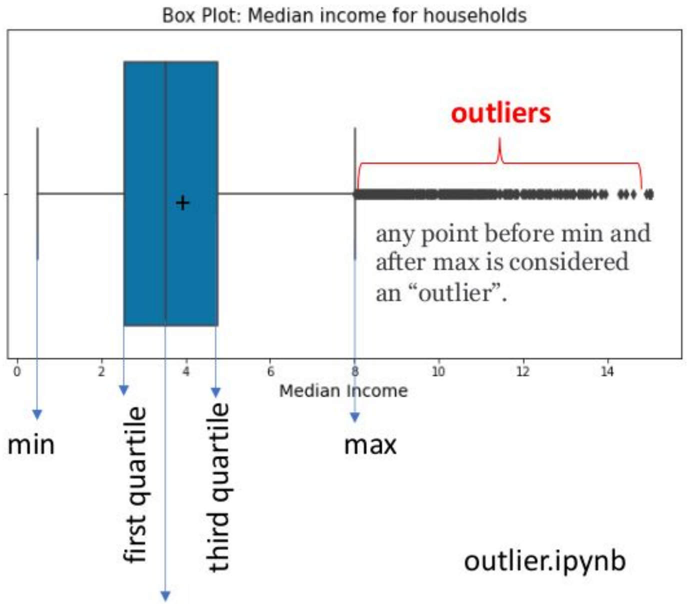
Preprocessing: Box Plot data visualisation

Can tell us about:

- Outliers (any data point before min and after max)
- Symmetrical data
- Skewed data
  - Example: Upper tail is longer than the lower tail (skewed to the right)
- Quartile:



median < mean (+)

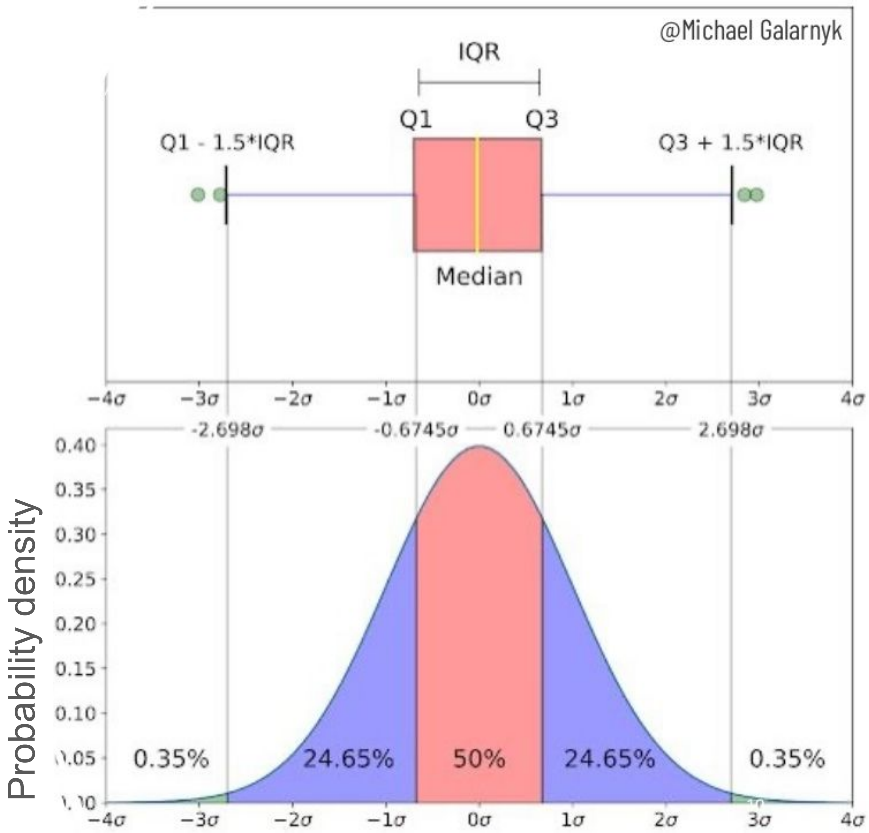


any point before min and after max is considered an "outlier".

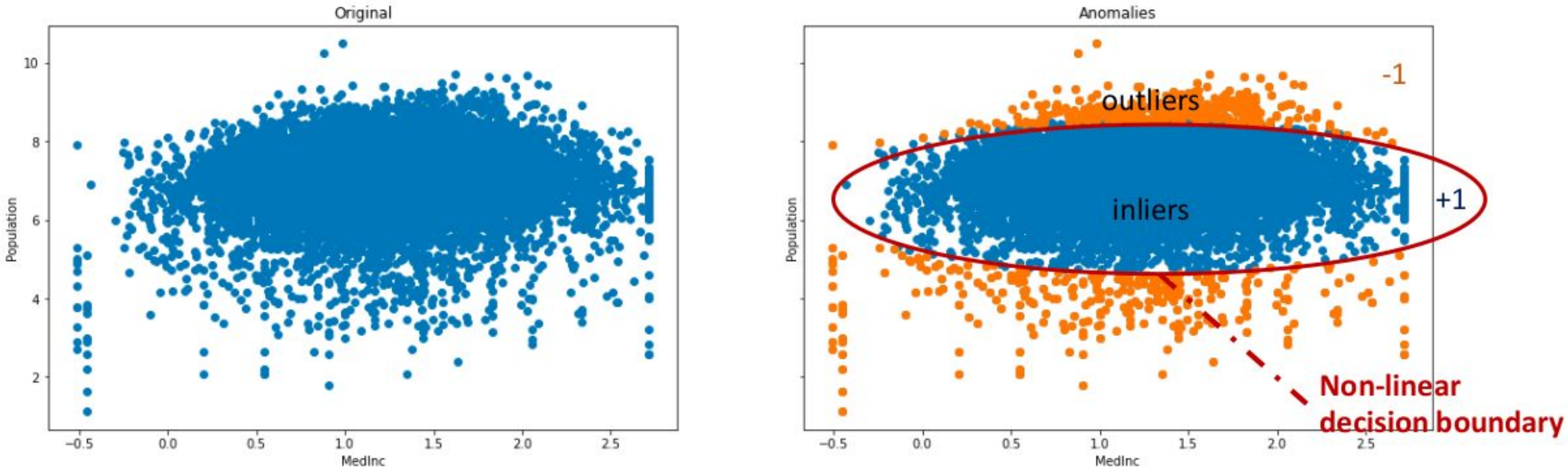
outlier.ipynb

Preprocessing: Box plot of a ‘normal’ or ‘bell-shaped’ distribution

- In this example, outliers are 0.7% of the data.
- In a perfectly normal distribution, mean is equal to median.
- See “outlier.ipynb”



Preprocessing: Anomaly Detection through Outliers



- Outlier detection is used for anomaly detection.
- A machine learning technique used for outlier detection is One-class-SVM.
- A trained One-class-SVM model returns +1 or -1 to indicate whether the data is an "inlier" or "outlier" respectively.
- One-class-svm can be trained with unlabeled data (unsupervised machine learning)
- See “one-class-svm-outlier.ipynb”

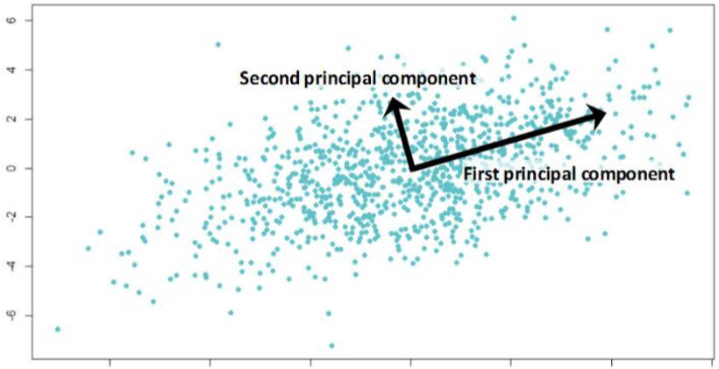
Preprocessing: High dimensional data

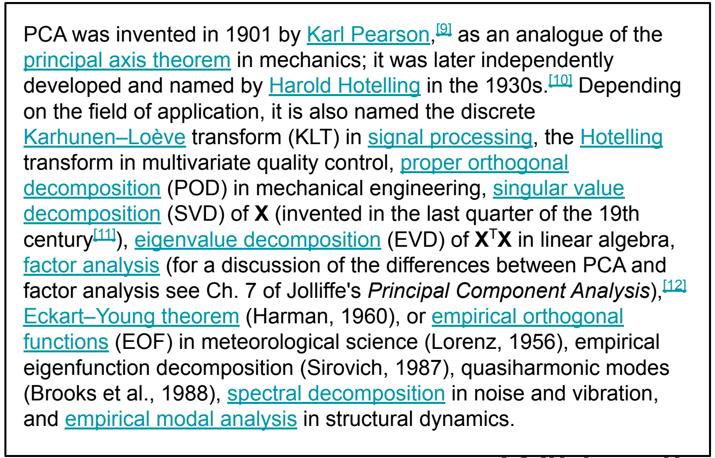
- High Dimensional means that the number of dimensions are staggeringly high —so high that calculations become extremely difficult.
- With high dimensional data, the number of features can exceed the number of observations.
  - Example: in medical datasets or gene datasets where the number of attributes tens of thousands.
- In case of dealing with high dimensional data, Feature Selection and Dimensionality Reduction techniques can reduce the number of dimensions and the complexity of computation and perform optimization.

PCA example for preprocessing

- Principal Component Analysis (aka see below)
- Basic principle: Some axes of variability are more important than others
- Use examples:
  - Dimensionality reduction/data compression
  - Data visualization and Exploratory Data Analysis
  - Create uncorrelated features/variables that can be an input to a prediction model
  - Uncovering latent variables/themes/concepts
  - Noise reduction in the dataset
- See “pca_example.ipynb”

linkedin.com

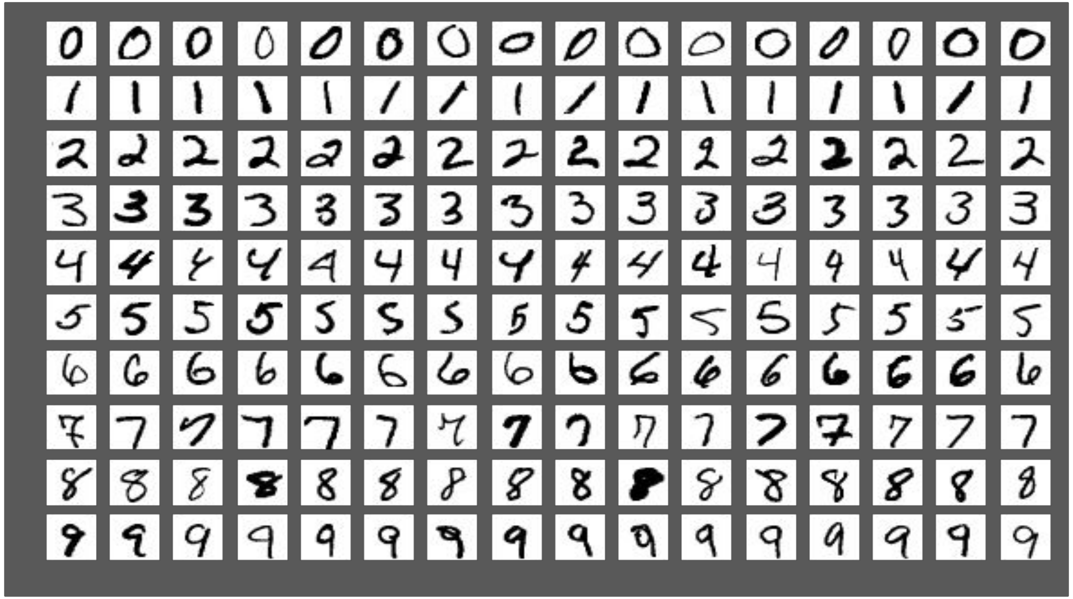



PCA was invented in 1901 by Karl Pearson,[9] as an analogue of the principal axis theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s.[10] Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, singular value decomposition (SVD) of **X** (invented in the last quarter of the 19th century[11]), eigenvalue decomposition (EVD) of $\mathbf{X}^T\mathbf{X}$ in linear algebra, factor analysis (for a discussion of the differences between PCA and factor analysis see Ch. 7 of Jolliffe's *Principal Component Analysis*),[12] Eckart–Young theorem (Harman, 1960), or empirical orthogonal functions (EOF) in meteorological science (Lorenz, 1956), empirical eigenfunction decomposition (Sirovich, 1987), quasiharmonic modes (Brooks et al., 1988), spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics.

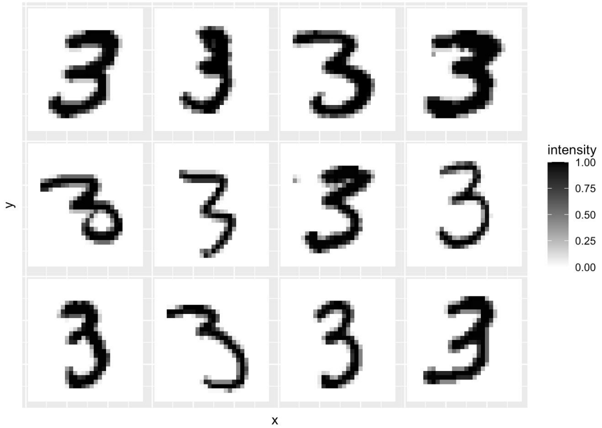
Wikipedia

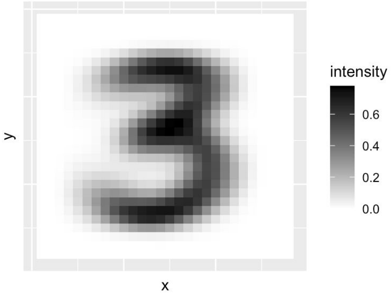
PCA on MNIST dataset

- The MNIST dataset has 784 dimensions for each example - 784 pixels
- Do we need all 784 dimensions?

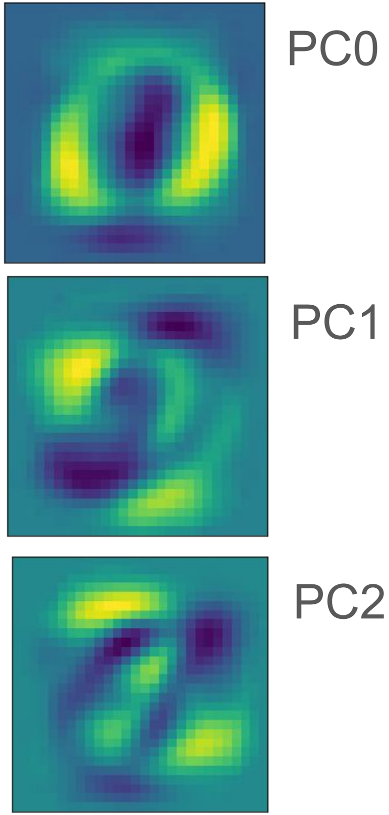


Principal Components don't look intuitive





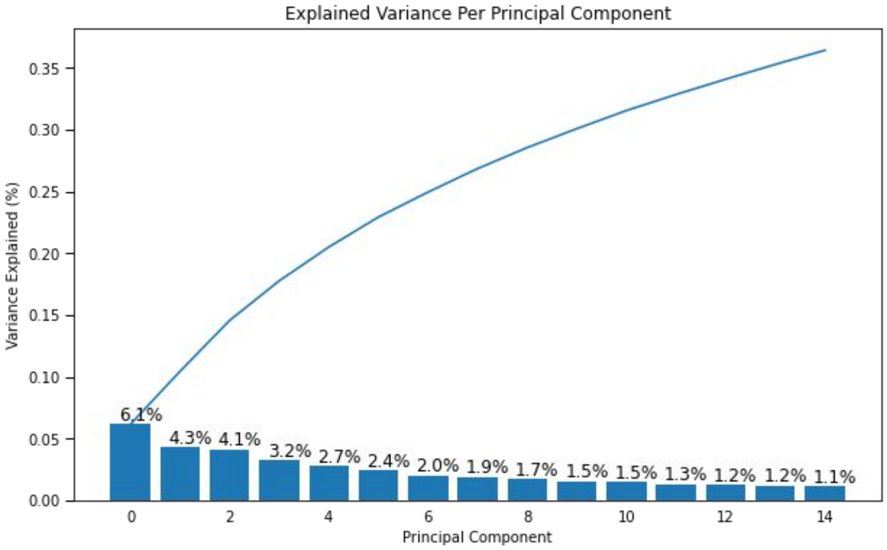
Average 3

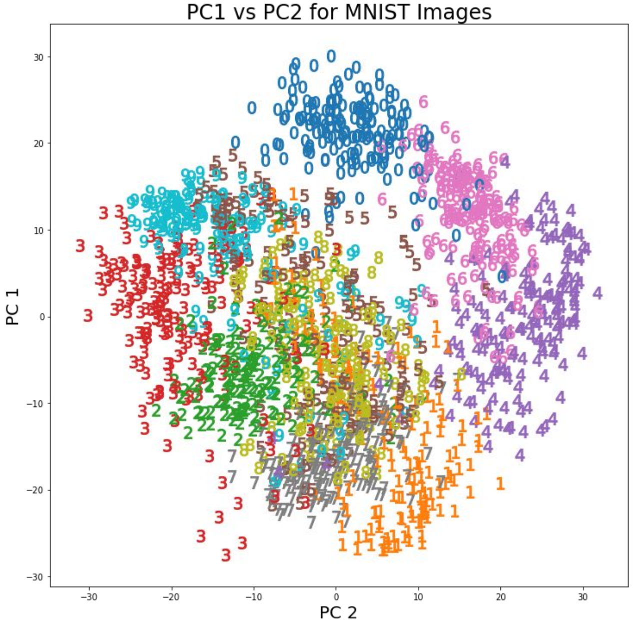


Whole dataset

Does the dataset have a normal distribution?

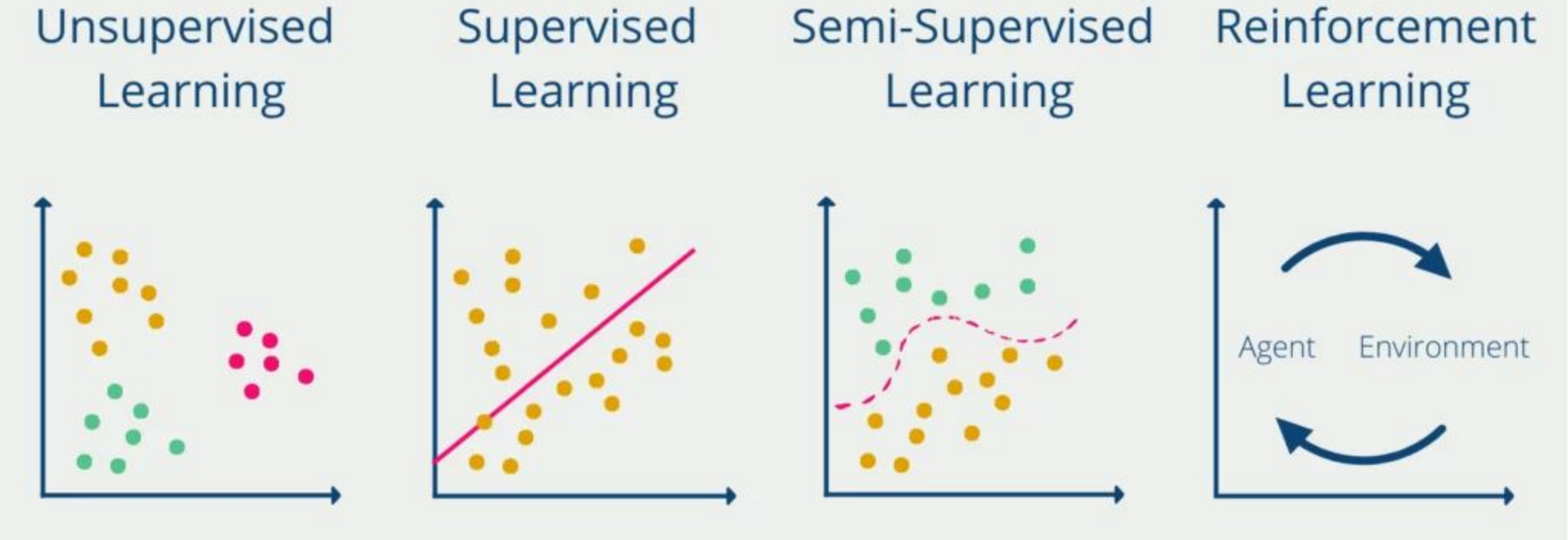
Adding more Principle components allows us to capture more of the variance





For PCA to work perfectly data must have a normal distribution

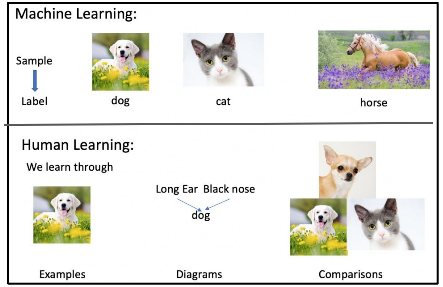
Machine learning categories



Examples of categories and questions

Supervised learning

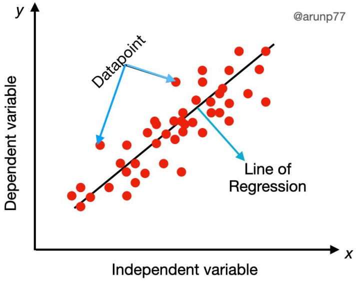
Classification predicts a label



@CMU ML Blog

Example question: Can you show me a picture of a dog?

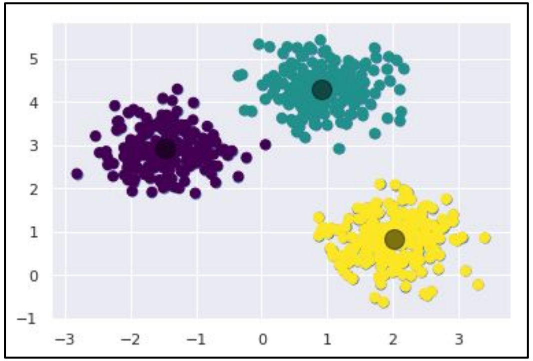
Regression predicts a value



Example question: What is the sales value for next month?

Unsupervised learning

Clustering can help us find rules



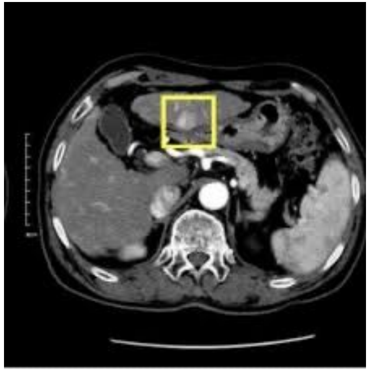
Example question: Can one recommend products based on consumer purchase history?

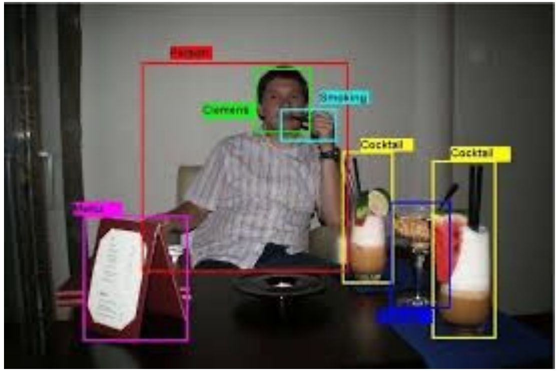
Supervised Machine Learning

- Needs labeled training set
- It has the knowledge about the correct output
- It is called "supervised" learning because the algorithm is guided by the correct output labels during training, which supervise the learning process.
- Example uses:
    - Prediction task in regression problems where the goal is to predict a “value”.
    - Classification task in classification problems where the goal is to predict a “label” or “category” or “class”.

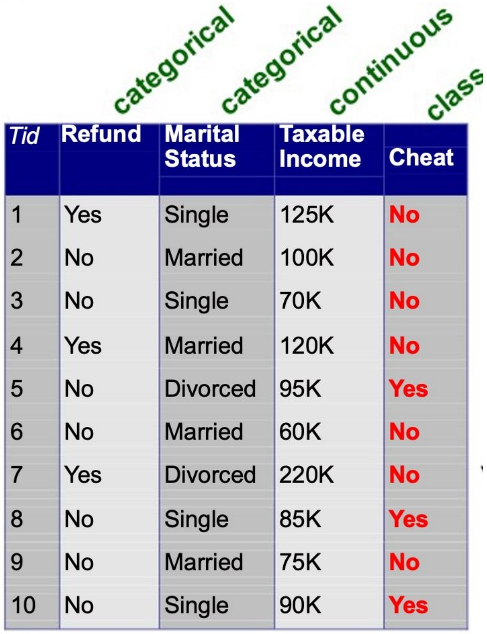
Supervised Machine Learning

- Needs a labelled dataset
- Predicts a label or category
- Classification types include:
  - Binary (only two labels)
  - Multi-class (more than two labels)
- Example Applications:
  - Object Recognition
  - Face Detection
  - Face Recognition
  - Scene Recognition
  - Malware Detection
  - Cancer Detection



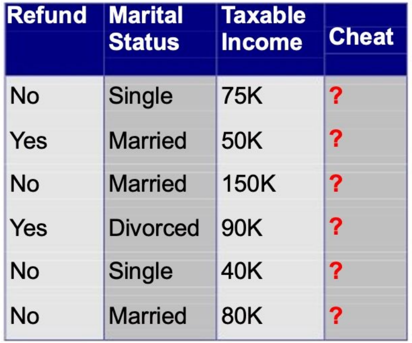


Example of dataset labels for supervised classification


categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

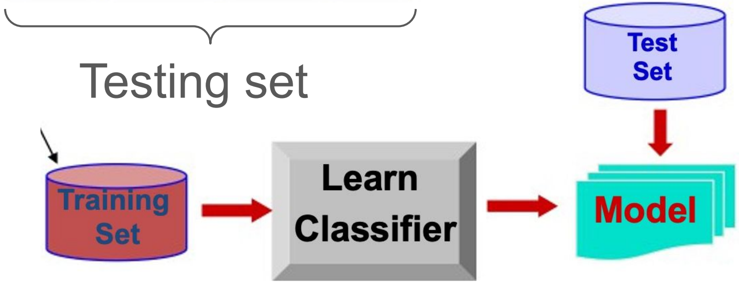
Training set



Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Testing set

Model to predict the “Cheat” label



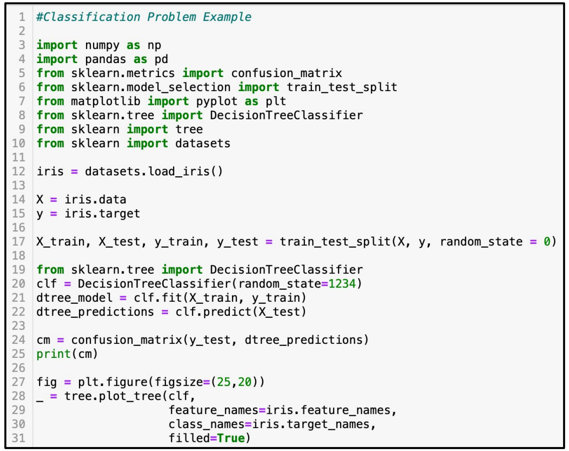
Multi-class Classification Example

The ‘Iris’ dataset

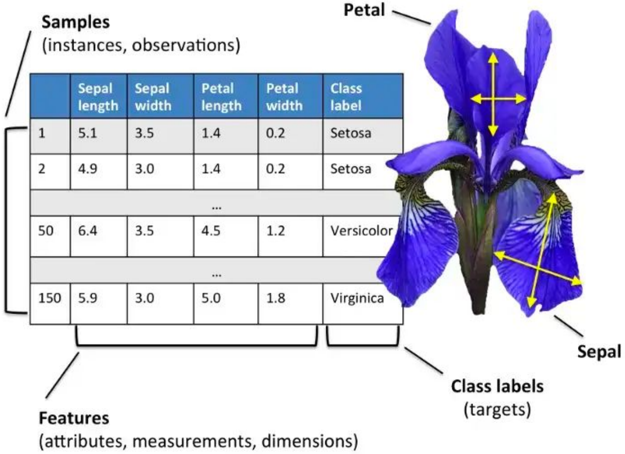
Iris Versicolor

Iris Setosa

Iris Virginica



```
#Classification Problem Example

import numpy as np
import pandas as pd
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
from sklearn import datasets

iris = datasets.load_iris()

X = iris.data
y = iris.target

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0)

from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=1234)
dtree_model = clf.fit(X_train, y_train)
dtree_predictions = clf.predict(X_test)

cm = confusion_matrix(y_test, dtree_predictions)
print(cm)

fig = plt.figure(figsize=(25,20))
_ = tree.plot_tree(clf,
                   feature_names=iris.feature_names,
                   class_names=iris.target_names,
                   filled=True)
```



Samples
(instances, observations)

Petal

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Sepal

Class labels
(targets)

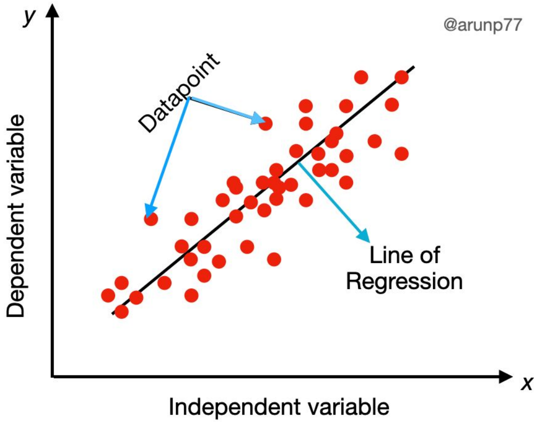
Features
(attributes, measurements, dimensions)

- Practical : ClassificationProblemExample-Python.ipynb

Supervised regression

Practical: regression_example.ipynb

- Needs a labelled dataset
- Predicts a value
- Example Applications:
    - Engine Performance
    - Prediction
    - Business Revenue
    - Real-Estate Market Prediction
    - Stock Market Prediction
    - Weather Data Analysis
    - Heart rate prediction
- Right: The line is the predicted linear regression model. The red dots are plotted according to the test data (X) and actual labels. E.g. x for sales-price and y for business-revenue.

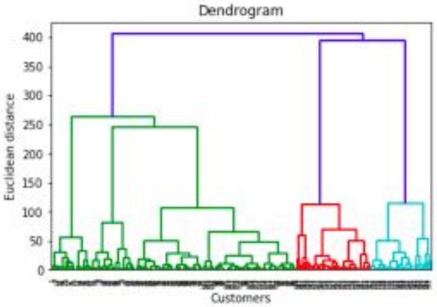


- The slope of the line indicates the correlation between the input and the output.
- How is the input correlated with the the output? Linear or non-linear correlation?

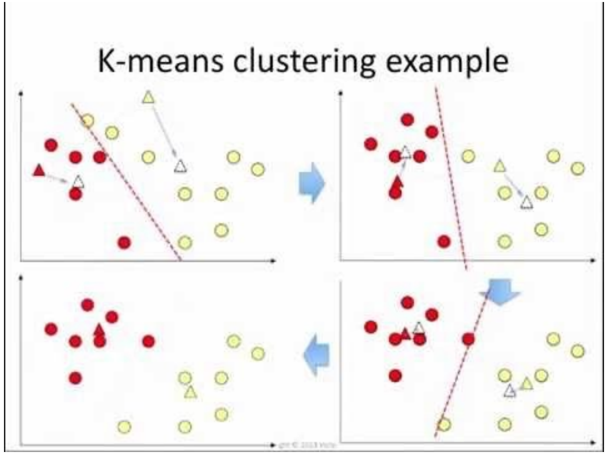
Important Notes for supervised learning

- Some ML models can be used for both classification and regression problems
- Example:
    - There are two types of decision trees: 1) regression trees, 2) classification trees.
- Some Regression models can be used to solve classification problems
    - Example:
        - If ”business-revenue” is greater than a threshold, label with “High-Revenue”
- Logistic regression is a classification model and predicts a label or category

Unsupervised Machine Learning



- Works with unlabeled data sets
- Does not need the knowledge about the correct output
- Called "unsupervised" learning because the algorithm is not guided by output labels during training. Task is to learn to identify meaningful patterns or relationships “unsupervised”
- Applications include:
    - Data exploration and association rule discovery . Example: Clustering (discover similarities and differences)
    - Anomaly Detection
- Example of popular clustering algorithms:
    - K-means Clustering
    - Hierarchical Clustering such as single-linkage
    - Self-Organizing Map


K-means clustering example

Association Rule Discovery

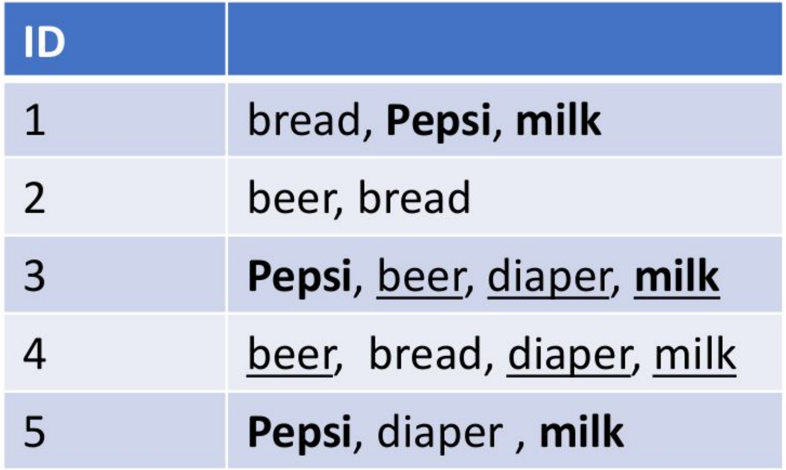
- Given a dataset:
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items
- Example algorithms: Apriori
  - Measures: Confidence, Support, Lift

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

$$Lift(\{X\} \rightarrow \ \{Y\}) = \frac{Confidence(\{X\} \rightarrow \{Y\})}{Support(Y)}$$

- A high lift value indicates a stronger association between the antecedent and consequent
- High support and confidence values indicate the frequency and reliability of the rule.



ID	
1	bread, **Pepsi**, **milk**
2	beer, bread
3	**Pepsi**, beer, diaper, **milk**
4	beer, bread, diaper, milk
5	**Pepsi**, diaper , **milk**

Rules of the form of:
{Antecedent} -> {Consequent}

Example Rule:
{Diaper, milk} -> {beer}
Confidence = 2/3 = 0.67
Support = 2/5=0.4
Lift = 0.67 / 0.6 = 1.11

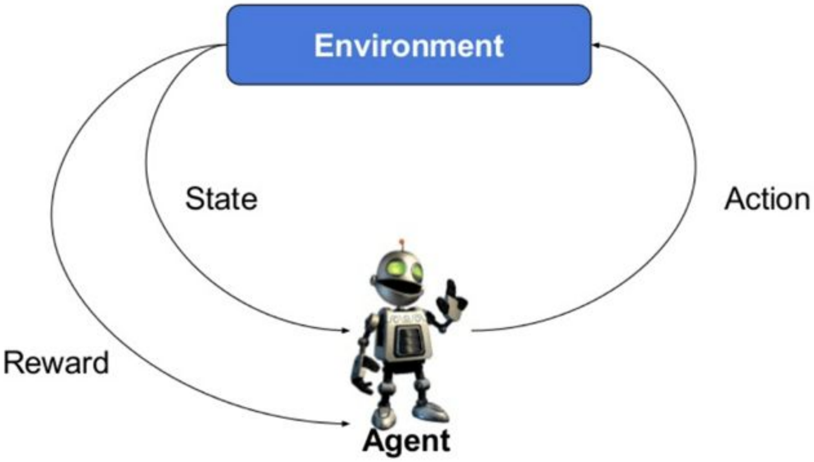
Reinforcement Learning (RL)

- A class of ML algorithm that learns to achieve a goal by trial-and-error interactions with an uncertain environment.
- The goal is to figure out a policy which selects an action in each state to optimize a reward function.
- It is about exploration vs. exploitation i.e., exploiting the best action so far to gain maximum reward.
- Drawback: may lead to suboptimal solution.
- Two popular RL algorithms:

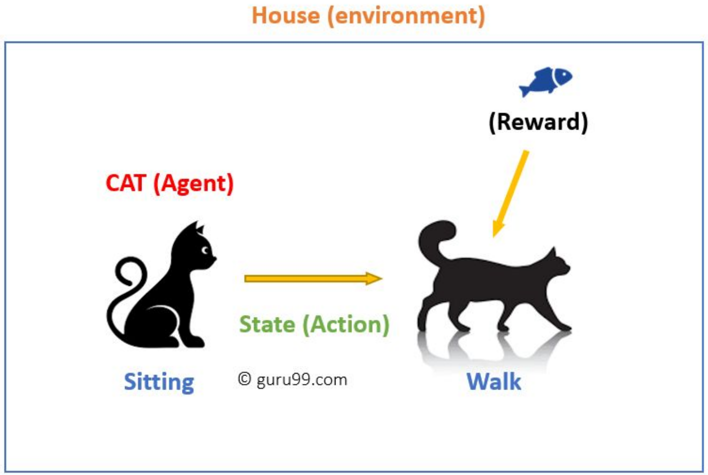
  1. Markov Decision Process

  2. Q learning

Reinforcement Learning (RL)



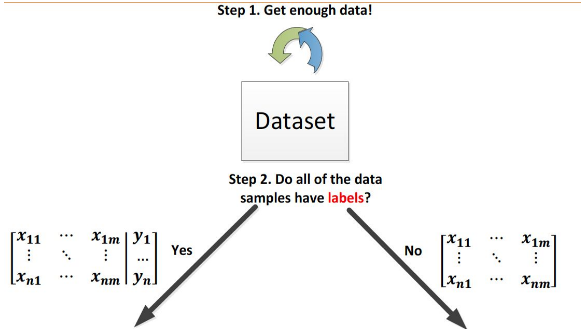
- Getting either a negative or positive experience
- Based on reward,revise agent behavior
- When an agent does something right (walk, sit)



© guru99.com

- There is no supervisor unlike supervised learning, only a reward feedback is used to train the algorithm
- Performs Sequential decision making
- Time has a crucial role in RL problems
- Agent's actions determines the subsequent data it receives,e.g., moving a robot to the left or right.

Machine Learning Workflow (Overview)



Step 1. Get enough data!

Dataset

Step 2. Do all of the data samples have labels?

Yes

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \cdots \\ x_{n1} & \cdots & x_{nm} & y_n \end{bmatrix}$$

No

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

Machine Learning Workflow (Overview)

@Ilias Tagkopoulos

Step 1. Get enough data!

Dataset

Step 2. Do all of the data samples have labels?

$$\left[\begin{array}{ccc|c} x_{11} & \cdots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \cdots \\ x_{n1} & \cdots & x_{nm} & y_n \end{array}\right]$$

Yes

No

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

Supervised Learning

Reinforcement Learning

Unsupervised Learning

Step 3: The task is to predict a continuous variable, assign a new sample to a class, or perform an optimal action?

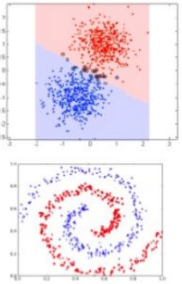
Step 3: The task is to cluster data together, find latent factors or complete missing data?

Assign to a class

Predict a continuous variable

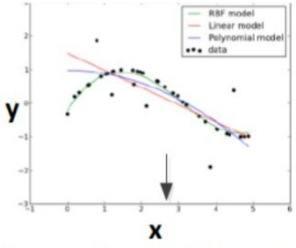
Perform an optimal action

CLASSIFICATION



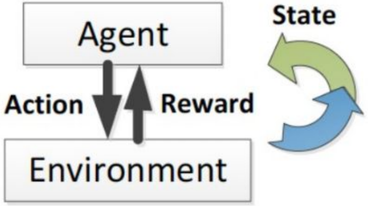
- **Bayesian Classification (Naïve Bayes)**
- **Linear Discriminant Analysis**
- **Artificial Neural Networks**
- **Decision Trees**
- **Support Vector Machines**

REGRESSION



Linear, polynomial, logistic, ...

REINFORCEMENT LEARNING (*)



Agent

Action

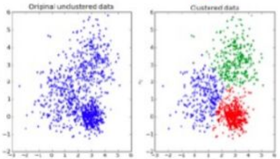
Reward

Environment

State

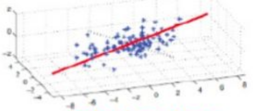
Markov Decision Process (MDP), POMDP, Q-learning,

Clustering



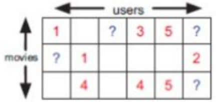
- **K-means**
- **Hierarchical clustering**
- **SOM**

Dimensionality Reduction



- **PCA**
- **ICA**

Missing Data



users

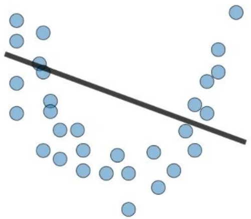
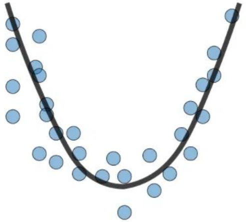
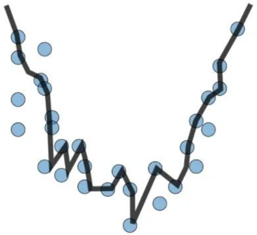
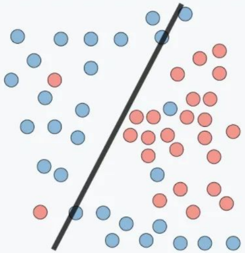
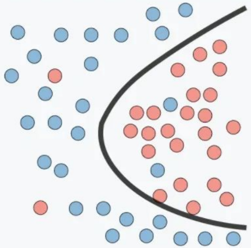
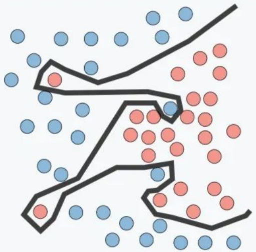
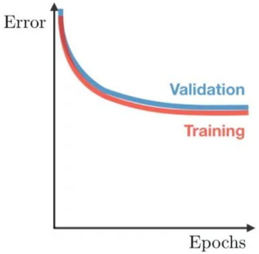
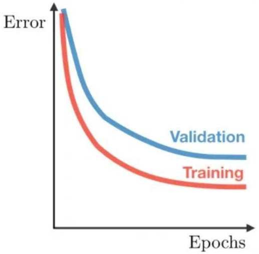
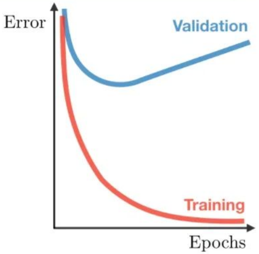
movies

1		?	3	5	?
?	1				2
	4		4	5	?

- **Collaborative filtering**
- **Market Basket analysis**

Assessing model fit is key to success

- Indicate the fit of a trained model and the causes of poor performance in machine learning.
- Goal of training a model: **Obtaining a generalized model to perform well on unseen data.**
- **Overfitting:**
    - **A function (ML model) is too closely aligned with a limited set of data**
- **Underfitting:**
    - **A function (ML model) is not aligned with a limited dataset**

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data