Introduction to Machine Learning

Bogdan Ichim

University of Bucharest
Faculty of Mathematics and Computer Science
Department of Computer Science

Bucharest, July 2024

Organization Issues

1. Grading policy
2. Strongly recommended programming languages
3. Datasets
4. Motivation

Grading policy: Projects

- Project (using ML, choice and details to be discussed): 60%
- A project should contain:
  - Technical report (about 10 – 20 pages)
  - Code
  - Dataset

Remark

Please, note that project recycling is strongly discouraged.

Grading policy: Final exam

- 4 exercises in total, each 10%
- some exercises may contain multiple-choice questions
- calculator for the simple mathematical operations may be used
- no smartphones or other similar smart devices are allowed

Teams for the projects

Let n be the number of a project team members. Then the following number of ML algorithms is strongly recommended for the project:

- n if $n \geq 4$
- 4 if $n < 4$

Strongly recommended programming languages

The following programming languages are strongly recommended for the projects:

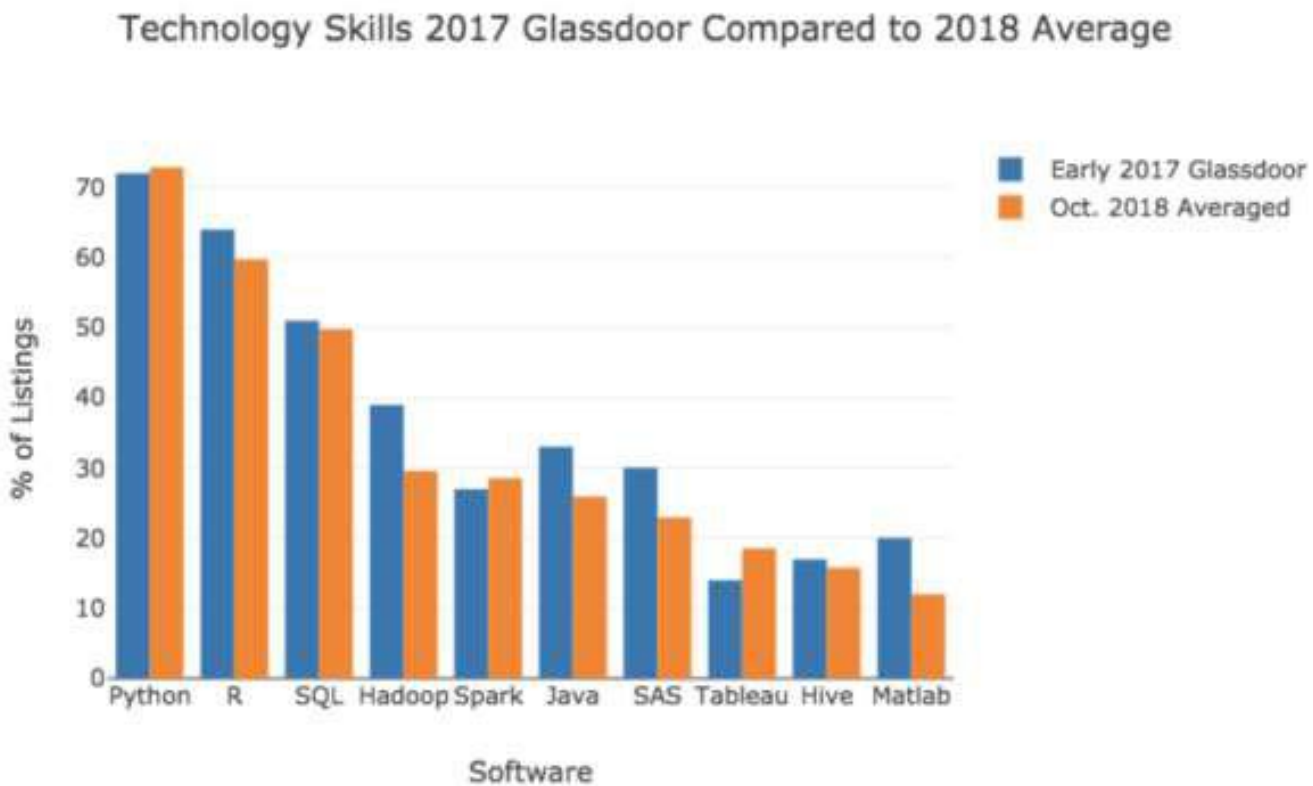
- R
- Python

Remark

Please, note that other programming languages are allowed.

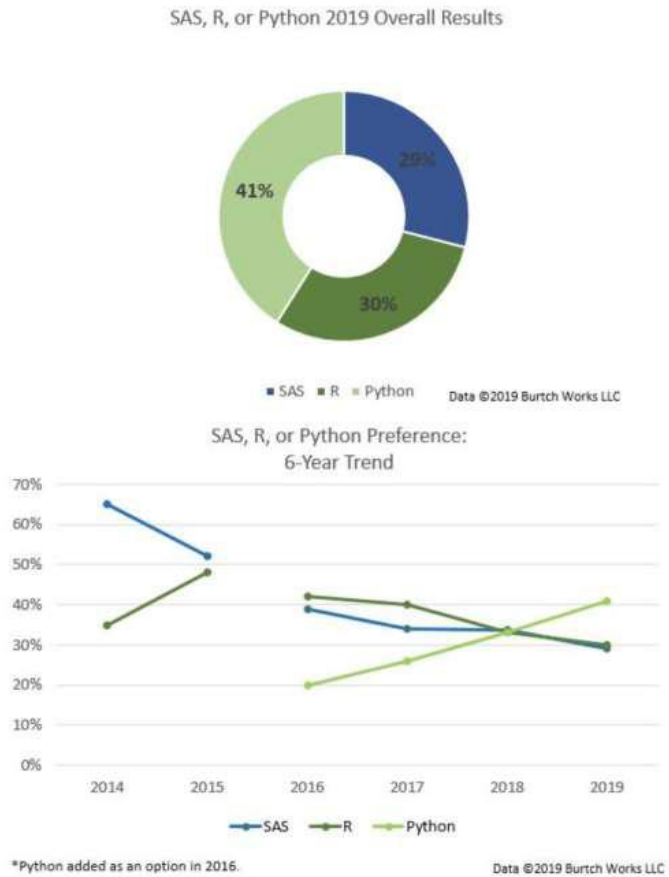
R versus Python

Below is graph of the top tech skills that appear in the Data Scientist job listings.



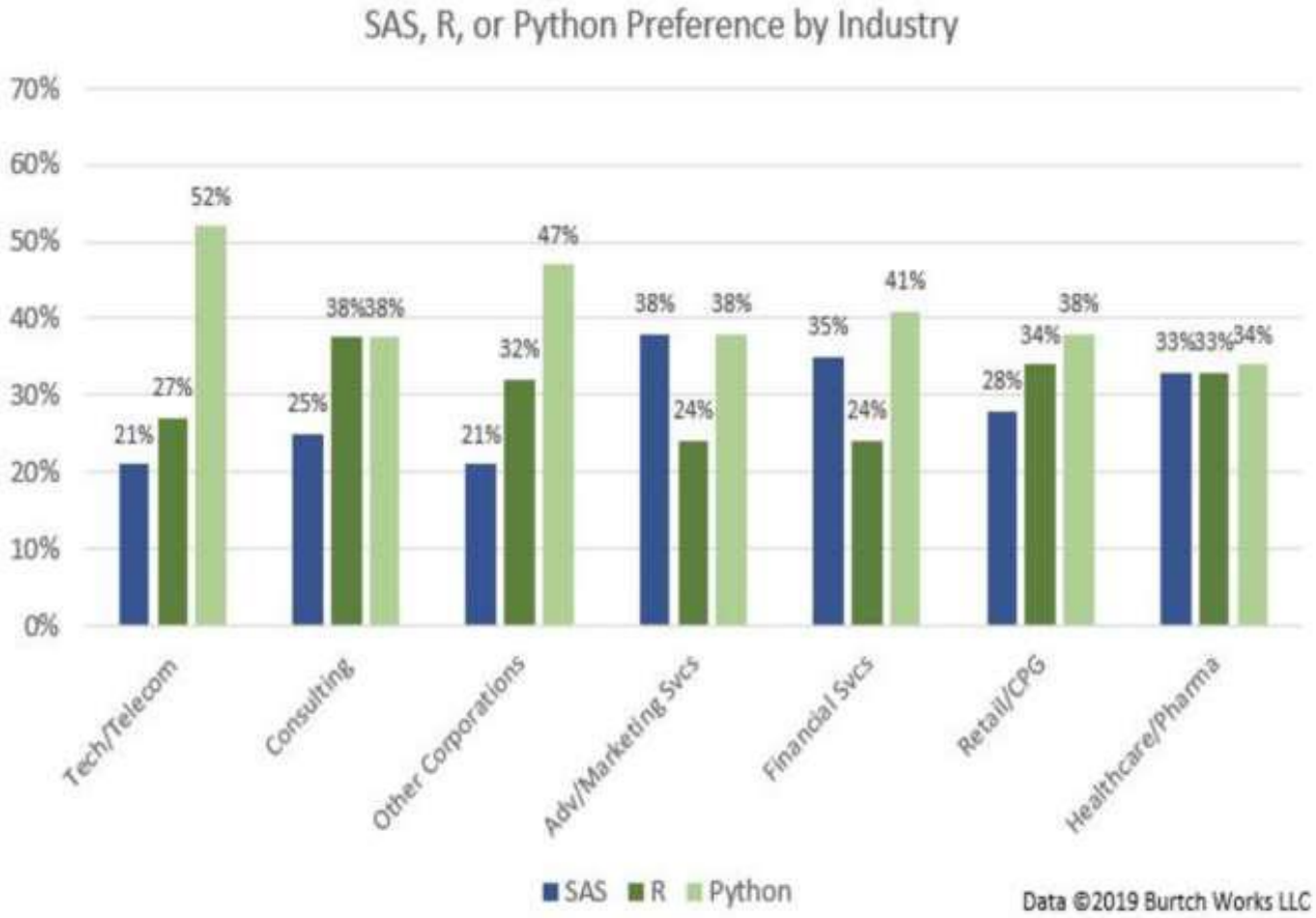
R versus Python

Python preference shows an ascending trend.



R versus Python

The preference depends on the industry.



R versus Python

Variance in Python

```
import numpy as np
vec = [1, 2, 3, 4, 5, 6, 7]
np.var(vec)
```

```
4.0
```

Variance in R

```
library(stats)
vec <- c(1, 2, 3, 4, 5, 6, 7)
stats::var(vec)
```

```
[1] 4.666667
```

R utilizes Bessel's correction when calculating variance, which changes the formula from returning **population variance** to **sample variance**

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N-1}$$

✅ Using Bessel's correction gives an unbiased estimator as demonstrated in this example *(sd of 2 implies a variance of 4)*

```
map_dbl(1:100000, ~ {
  x <- rnorm(n = 5, mean = 0, sd = 2)
  sum((x - mean(x))^2) / length(x)
}) |> mean()
```

```
[1] 3.190721
```

```
map_dbl(1:100000, ~ {
  x <- rnorm(n = 5, mean = 0, sd = 2)
  sum((x - mean(x))^2) / (length(x) - 1)
}) |> mean()
```

```
[1] 3.993807
```

Datasets

You can get some inspiration (especially concerning the projects) by playing with the datasets listed here:

- Wiki list of datasets for ML research
- Kaggle Datasets
- OpenML
- PMLB (supervised, accessible through Python API)
- R Datasets (accessible in R)
- IMDb Datasets

Motivation

Below is list of the top fastest growing jobs in the next decade according to www.visualcapitalist.com.

Occupation	Percent employment change, 2020–2030P	Numeric employment change, 2020-2030P	Median annual wage, 2020
Wind turbine service technicians	68.2%	4,700	$56,230
Nurse practitioners	52.2%	114,900	$111,680
Solar photovoltaic installers	52.1%	6,100	$46,470
Statisticians	35.4%	14,900	$92,270
Physical therapist assistants	35.4%	33,200	$59,770
Information security analysts	33.3%	47,100	$103,590
Home health and personal care aides	32.6%	1,129,900	$27,080
Medical and health services managers	32.5%	139,600	$104,280
Data scientists and mathematical science occupations, all other	31.4%	19,800	$98,230
Physician assistants	31.0%	40,100	$115,390

Schedule Monday 08.07.2024

Lecture 1 . . . . . July 08, 09:00-11:00, 214
Lecture 2 . . . . . July 08, 11:00-13:00, 214

Break

Study session 1 . . . . . July 08, 16:00-18:00, 218
Laboratory 1 . . . . . July 08, 18:00-20:00, 218

Schedule Tuesday 09.07.2024

Lecture 3 ..... July 09, 09:00-11:00, 214
Lecture 4 ..... July 09, 11:00-13:00, 214

Break

Study session 2 ..... July 09, 16:00-18:00, 218
Exam ..... July 09, 18:00-19:00, 010
Laboratory 2 ..... July 09, 19:00-20:00, 218

Schedule Wednesday 10.07.2024

Lecture 5 . . . . . July 10, 09:00-11:00, 214
Lecture 6 . . . . . July 10, 11:00-12:00, 214
Scientific presentation . July 10, 12:00-13:00, IMAR 309

Break

Study session 3 . . . . . July 11, 16:00-18:00, 218
Laboratory 3 . . . . . July 11, 18:00-20:00, 218

Schedule Thursday 11.07.2024

Lecture 7 . . . . . July 12, 09:00-11:00, 214
Lecture 8 . . . . . July 12, 11:00-13:00, 214

Break

Study session 4 . . . . . July 12, 16:00-18:00, 218
Laboratory 4 . . . . . July 12, 18:00-20:00, 218

Schedule Friday 12.07.2024

Lecture 9 . . . . . July 08, 09:00-11:00, 214
Lecture 10 . . . . . July 08, 11:00-13:00, 214

Break

Study session 5 . . . . . July 08, 16:00-18:00, 218
Laboratory 5 . . . . . July 08, 18:00-20:00, 218

Break

Structure of the Lecture 1

- What is Data Science?
- What is Machine Learning?
  - Related fields
  - Definition
  - Data E (Experience), associated types of learning
  - Tasks T and applications
  - Performance P measures
- Example

What is Science?

Donald E. Knuth:

> *"Science is what we understand well enough to explain to a computer. Art is everything else we do."*

What is Data?

Clive Humby, UK Mathematician and architect of Tesco's Clubcard, 2006 (widely credited as the first to coin the phrase):

> *"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."*

Gary Wolf, Quantified Self Co-Founder, 2012 :

> *"People are saying, 'Big Data is the new oil.'"*

Virginia Rometty, IBM CEO, 2013:

> *"I want you to think about data as the next natural resource."*

What is Data?

Abhishek Mehta, CEO Tresata, 2013:

> *"Just like oil was a natural resource powering the last industrial revolution, data is going to be the natural resource for this industrial revolution. Data is the core asset, and the core lubricant, for not just the entire economic models built around every single industry vertical but also the socioeconomic models."*

Kevin Plank, founder and CEO of Under Armour, 2016:

> *"Data is the new oil. The companies that will win are using math."*

Qi Lu, the chief of Microsoft's Applications and Services, 2016:

> *"Data is the new oil."*

What is Data Science?

Data Science is focusing on business applications of quantitative fields like Machine Learning, Artificial Intelligence, Statistics, Mathematics, etc.

Down to earth, one can see Data Science as a set of methods and techniques for extracting useful information from high-dimensional sets of data. In other words, it is an interdisciplinary field that extracts value from data.

Nowadays Data Science relies heavily on Machine Learning.

What is Data Science?

The list bellow includes some of the most useful topics associated with Data Science.

- Machine Learning
- Artificial Intelligence
- Optimization and Statistics
- Combinatorics and Discrete Mathematics
- Decision and Voting Theory

Some examples of Data Science use cases are:

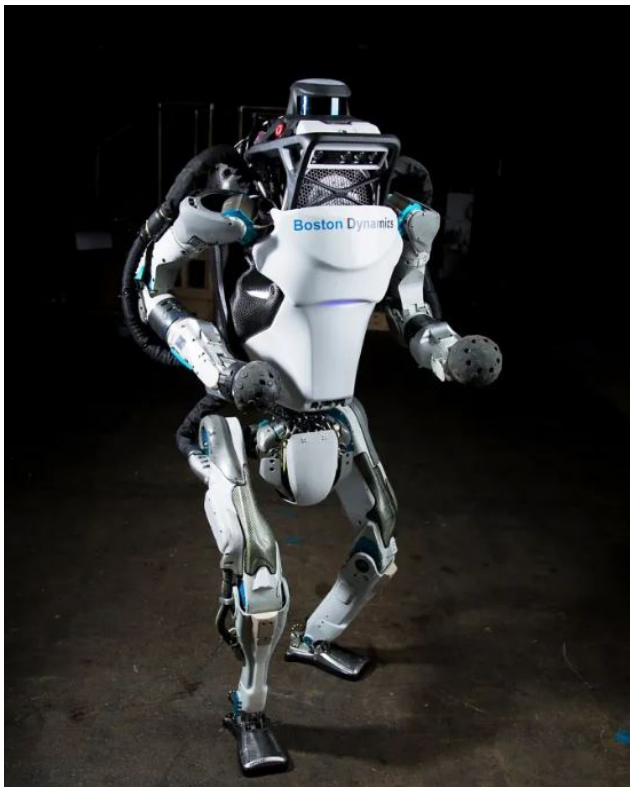
- predict revenue for the next quarter
- a model that detects fraudulent credit card transactions
- find customers who will most likely churn next month

Machine Learning, Artificial Intelligence and Data Science

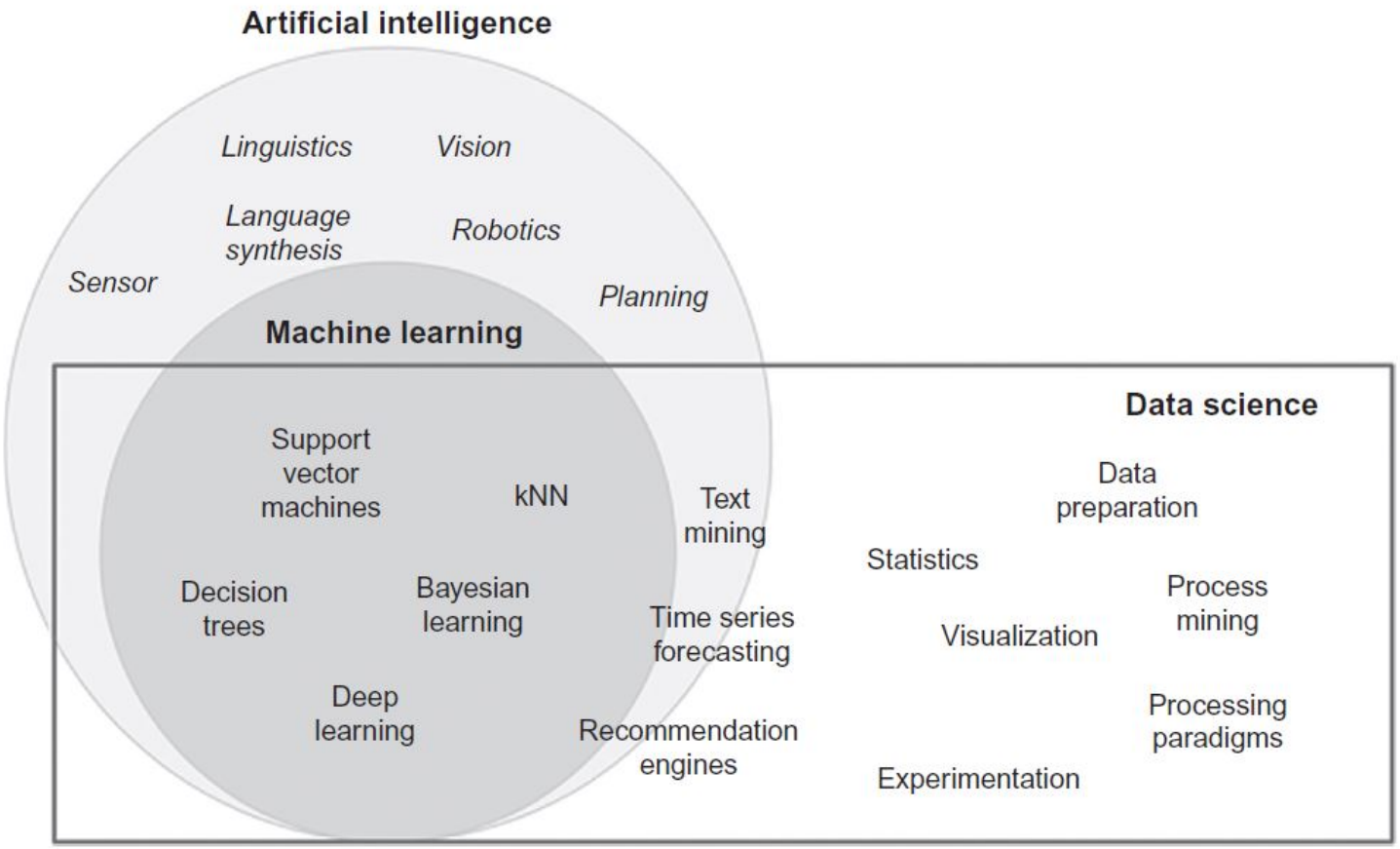
Machine learning, Artificial intelligence and Data Science are all related to each other. Unsurprisingly, they are often used interchanged and conflated with each other.

Machine Learning, Artificial Intelligence and Data Science

Artificial Intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions. Some examples: facial recognition, automated driving, sorting mail based on postal code.

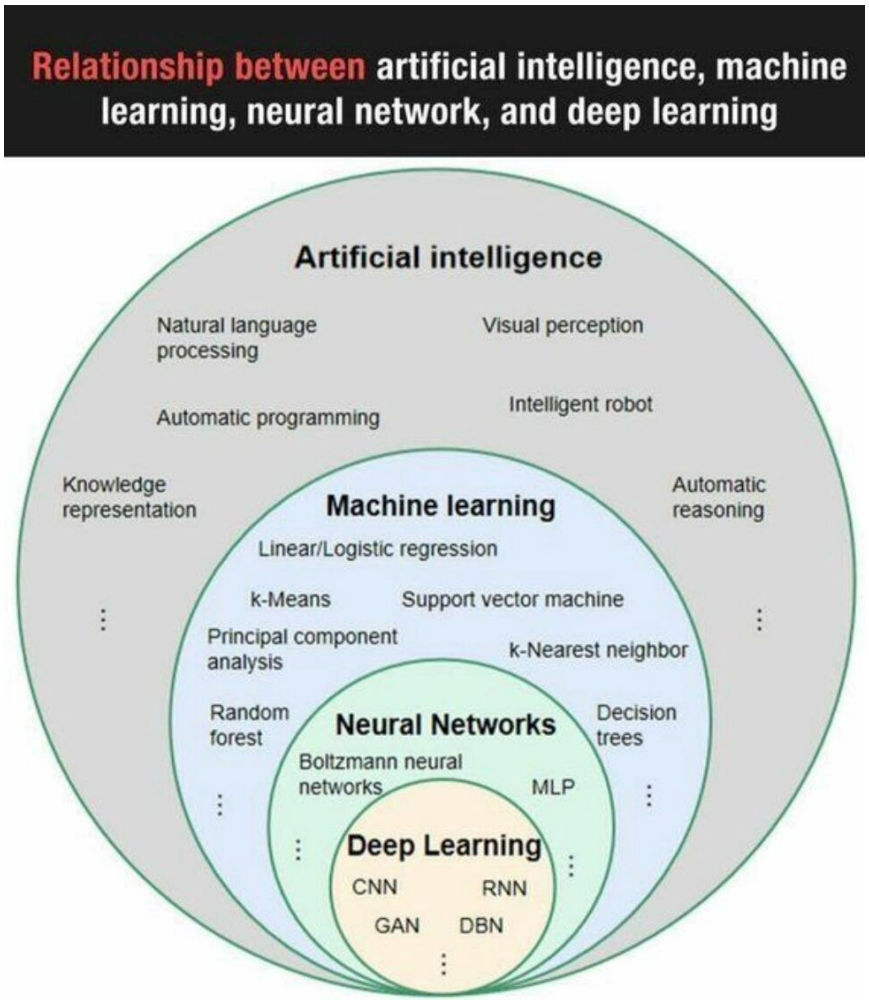


Machine Learning, Artificial Intelligence and Data Science



Taken from [KD]

Machine Learning and Deep Learning



What is Machine Learning?

Machine Learning (ML) is not a single approach but rather a diverse array of techniques.

ML is a blend of probability theory and statistics, linear algebra, optimization, and control theory, all worth studying in their own.

ML tools embrace classification, regression, clustering techniques, density estimation, feature (or representation) learning, matrix factorization, Bayesian networks, Markov random fields, and many others.

Related Fields and Terminology

- Artificial Intelligence
- Probability Theory and Statistical Inference
- Computational Statistics (high-dimensional statistics)
- Combinatorics
- Discrete Geometry
- Optimization
- Functional Analysis
- Data Mining
- Decision and Voting Theory

Remark

Terminology differs across different fields!!!

Definition of Machine Learning (ML)

According to [M], a machine learning algorithm is an algorithm that is able to learn from data:

> *"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 ."*

Three ingredients:

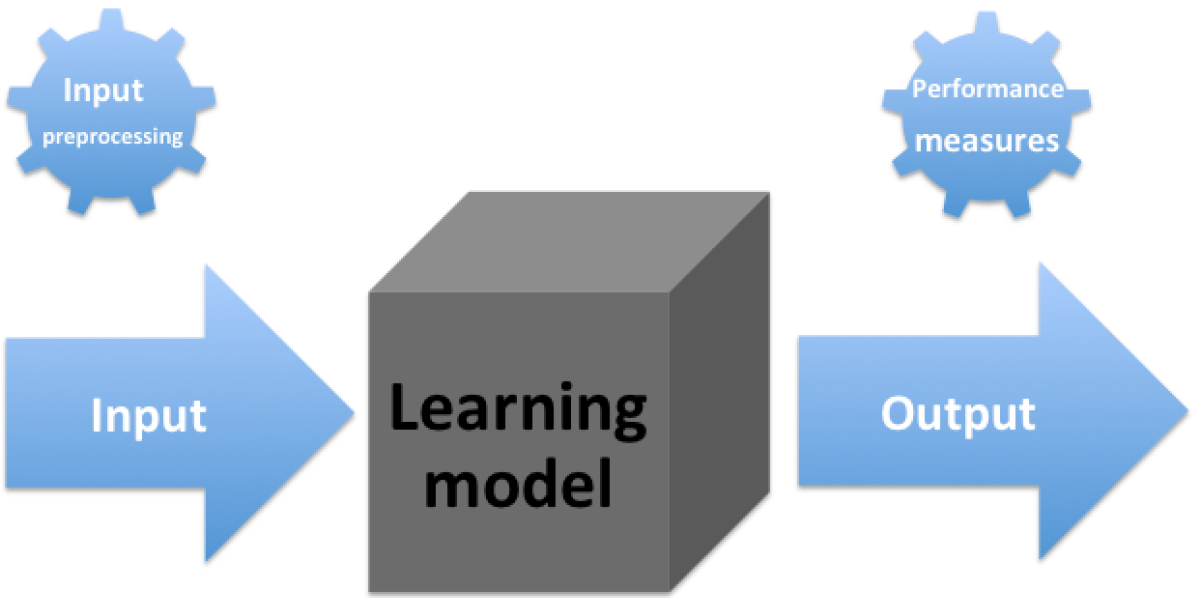
- **Experience E**
- **Task T**
- **Performance P**

Goal of Machine Learning

Goal of ML/SL: learning from data.

One wants to: execute task T based on experience E with optimal performance P.

Machine Learning Model



Main Idea

In the realm of ML these three ingredients interact in the following way:

- Select a ML algorithm (model) to solve the task T.
- The data in E is used to **train** (**estimate**) the algorithm (model) by maximizing the performance P on the training set E.
- By definition, in a ML algorithm, P should increase (error should decrease) with E.

1. Data E (Experience)
   - $\mathcal{X}$ - input space (measurement space, feature space, signal domain)
   - $\mathcal{Y}$ - output space (label space, response space, signal range)
2. Task T
   - to determine a function $f : \mathcal{X} \to \mathcal{Y}$.
3. Performance P
   - reward or utility function (its negative is a loss function)

ML is the solution of choice when dealing with tasks that are too complex to be carried out by completely solving a problem.

(1) We approach to machine learning as an input/output problem.
   - Input $x \in \mathcal{X}$: contains available information for the solution of the problem, the so called **predictors** i.e.:
     - historical data
     - explanatory factors
     - features of the individuals
     - qualitative features
   - Output $y \in \mathcal{Y}$: contains the solution of the problem, for instance:
     - explained (dependent) variables
     - forecasted data
     - qualitative response or classification results

(2) We distinguish between discrete-time and continuous-time setups and between deterministic and stochastic situations since they lead to very different levels of mathematical complexity.

Data E (Experience)

- $\mathcal{X}$ - input space (measurement space, feature space, signal domain)
- $\mathcal{Y}$ - output space (label space, response space, signal range)

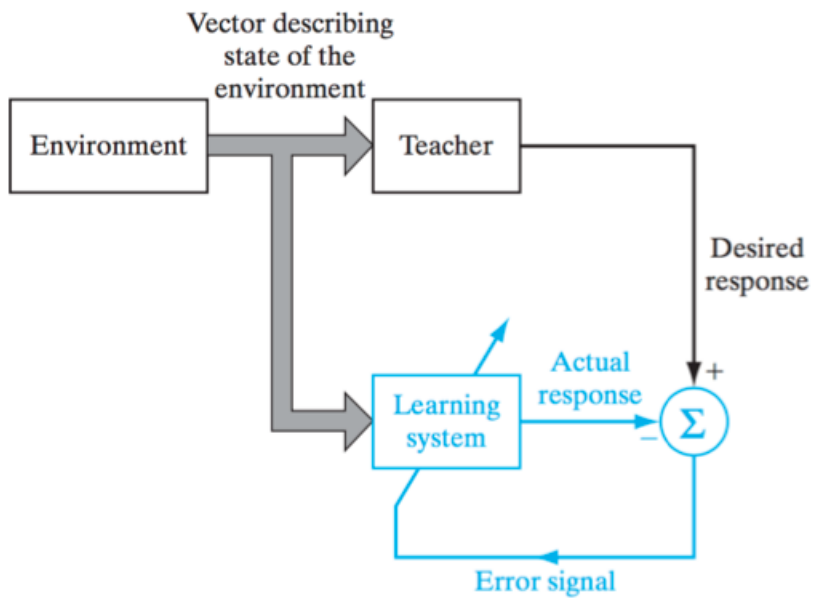
The nature of the data available determines the kind of learning algorithms that can be implemented.

The main groups are:

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**

Supervised learning

- **Supervised learning**: the dataset contains features, but each example is also associated with a label or target.

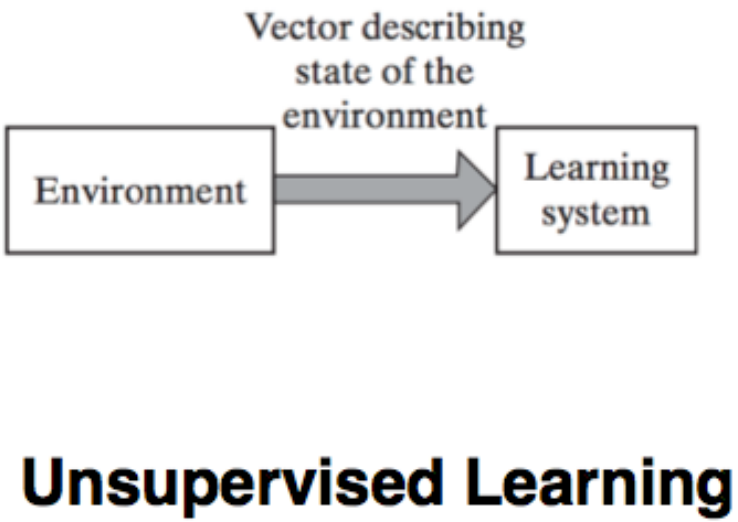


Taken from [H]

Supervised Learning

Unsupervised learning

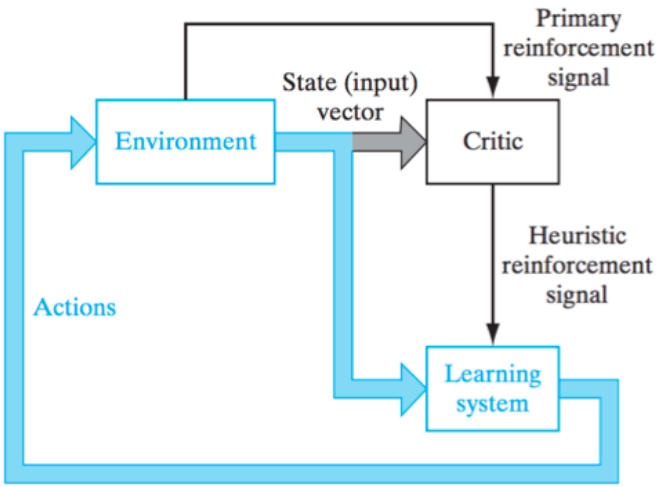
- **Unsupervised learning**: the dataset contains features, but no label or target is given. The structural properties are learnt and data is grouped based on some measure of similarity.



Taken from [H]

Reinforcement learning

- **Reinforcement learning**: algorithms that do not use a fixed dataset and interact with the environment, so there is a feedback loop between the learning system and its experiences.



Taken from [H]

Reinforcement Learning

- Example from [M, Chapter 13]: A learning robot. The robot, has a set of sensors to observe the state of its environment, and a set of actions it can perform to alter this state.

Task T: Classification

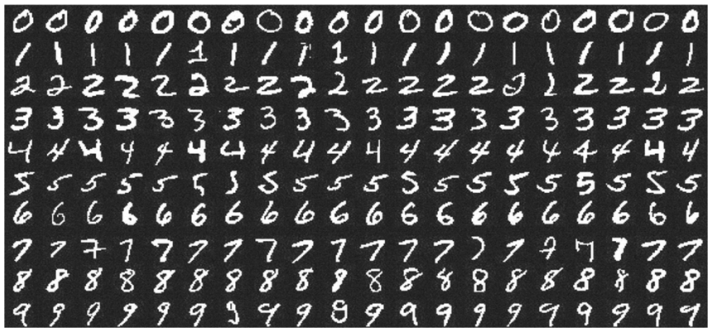
- **Classification**: assign input to one of the k categories, one is interested in producing for example $f : \mathcal{X} \longrightarrow \{1, \ldots, k\}$, with $\mathcal{X}$ containing features (think of $\mathbb{R}^n$). Function f can also be a probability distribution over classes.

Examples:

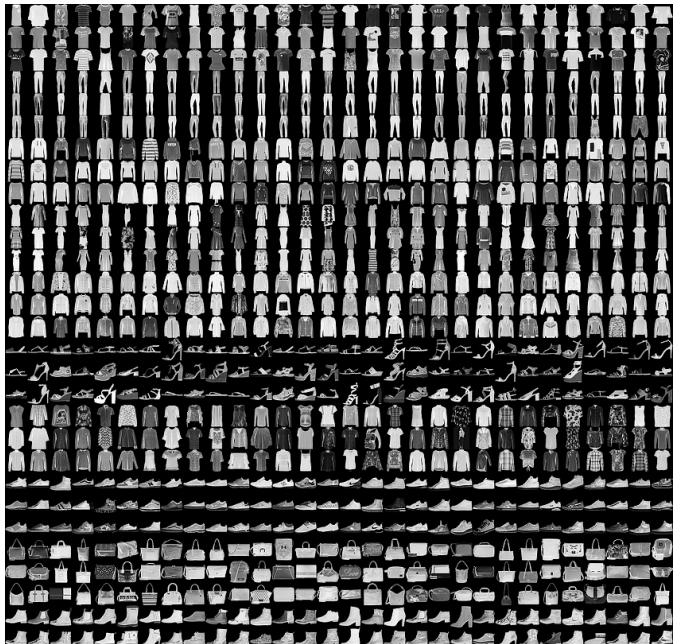
- Document classification
- Speech recognition (pronounced digits, male/female)
- Pattern recognition
- Biological (medical) classification
- Credit scoring
- Object recognition
- Handwriting recognition
- Recommender systems

Classification: Standard Examples and Datasets

- Classification of handwritten digits (MNIST)



- Classification of clothing (Fashion-MNIST)



Task T: Clustering

- **Clustering**: grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters), that is learn $f : \mathcal{X} \longrightarrow \{1, \ldots, k\}$ where k need not to be specified.
  Examples:
  - Distribution-based clustering
  - Density-based clustering
  - Connectivity-based clustering (hierarchical clustering)

  Applications
  - Medical imaging
  - Business and marketing (market research, recommender systems)
  - World wide web (social network analysis)
  - Computational chemistry

Task T: Regression

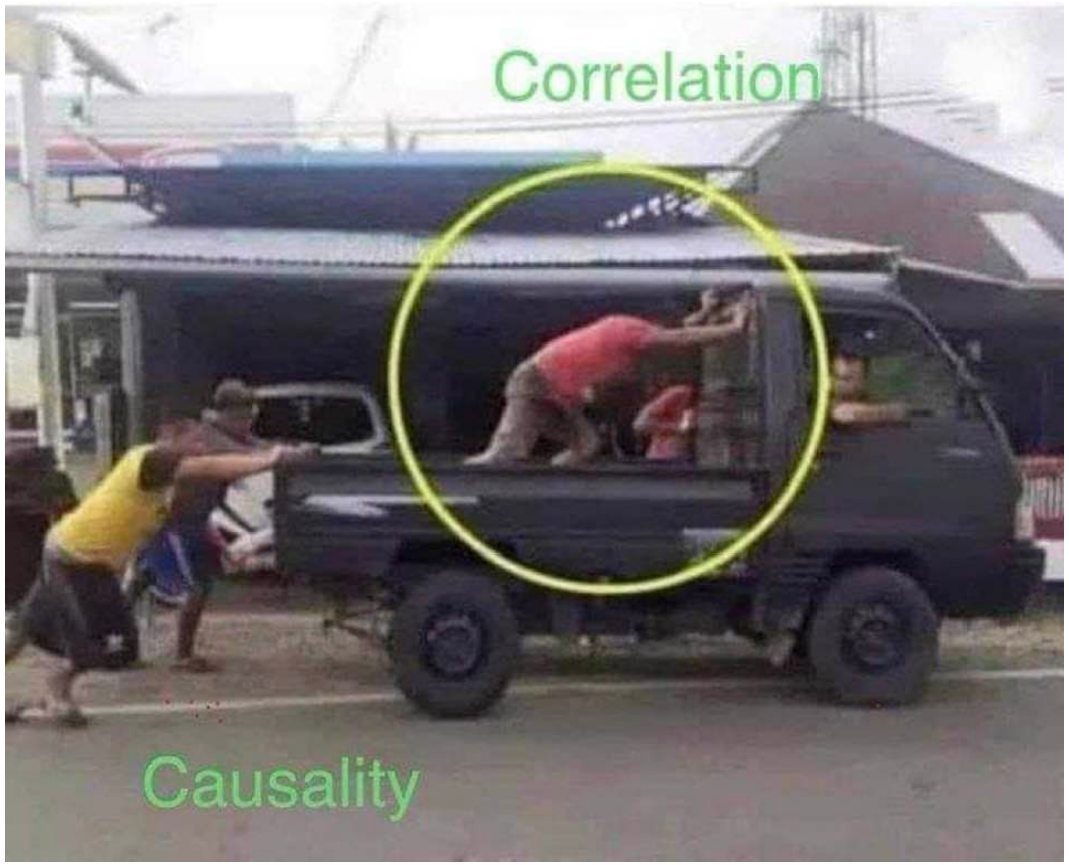
- **Regression**: learn a mapping from the input (covariates) space to the output space $f : \mathcal{X} \to \mathcal{Y}$ (learn an estimator) (think of multivariate case $f : \mathbb{R}^n \to \mathbb{R}$). Examples:
  - Prediction of the expected claim amount that an insured person will make (used to set insurance premiums).
  - Prediction of future prices of securities. Used for algorithmic trading.
  - Prediction of number of passengers in a given flight.
  - Prediction of energy consumption.
  - Logistics and infrastructure management applications.

Regression: correlation vs causality

Has the cat damaged the roof?

Regression: correlation vs causality

While causation and correlation can exist at the same time, **correlation does not imply causation!**



Taken from John List

Task T: Transcription

- **Transcription**:
  unstructured representation of some kind of data $\longrightarrow$ transcribe into discrete, textual form.

  Examples:
  - Optical character recognition: a photograph containing an image of text $\longrightarrow$ this text in the form of a sequence of characters (e.g., in ASCII or Unicode format). Google Street View uses deep learning to process address numbers in this way.
  - Speech recognition: an audio waveform $\longrightarrow$ a sequence of characters or word ID codes describing the words that were spoken in the audio recording. Deep learning is a crucial component of modern speech recognition systems used at major companies including Microsoft, IBM, and Google.

Task T: Machine translation

- **Machine translation**: the input of a sequence of symbols in some language $\longrightarrow$ a sequence of symbols in another language. This is commonly applied to natural languages, such as translating from English to French. Have a look at this article in the NY Times Magazine.

Task T: Anomaly detection

- **Anomaly detection**: a set of events or objects $\longrightarrow$ some of them marked as unusual or atypical. Example: credit card fraud detection. By modeling your purchasing habits, a credit card company can detect misuse of your cards. If a thief steals your credit card or credit card information, the thief's purchases will often come from a different probability distribution over purchase types than your own. The credit card company can prevent fraud by placing a hold on an account as soon as that card has been used for an uncharacteristic purchase.

  Recent example: Detecting ICS attacks using recurrent neural networks, see here.

Task T: Synthesis and sampling

- **Synthesis and sampling**: the ML algorithm is asked to generate new examples that are similar to those in the training data. Synthesis and sampling via machine learning can be useful for media applications where it can be expensive or boring for an artist to generate large volumes of content by hand. Examples:
  - Video games: automatically generate textures for large objects or landscapes, rather than requiring an artist to manually label each pixel.
  - Speech synthesis: a written sentence $\longrightarrow$ an audio waveform containing a spoken version of that sentence. This is a kind of structured output task, but with the added qualification that there is no single correct output for each input, and we explicitly desire a large amount of variation in the output, in order for the output to seem more natural and realistic.

One related example to play with:

- Hand writing generation by Google Brain. Here

Performance

The performance evaluation P is needed during:

- Training: maximization of P determines the algorithm hyperparameters.
- Testing: the differences in P obtained during training and testing allow us to assess if we are in an under or overfitting situation.

The choice of P modifies hence how the ML algorithm is going to perform. The pertinence of a given P depends on the task.

Performance of classification

- For classification (or clustering): we often measure the **accuracy** of the model. Accuracy is the proportion of examples for which the model produces the correct output. We can also obtain equivalent information by measuring the **error rate**, the proportion of examples for which the model produces an incorrect output. The error rate is often referred to as the **expected 0-1 loss**. The 0-1 loss on a particular example is 0 if it is correctly classified and 1 if it is not. However, there are other possibilities for which it is important to understand the notion of **confusion matrix** and its derivative concepts.

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive** (Type I error)
	Predicted condition negative	**False negative** (Type II error)	**True negative**

condition positive (P)
the number of real positive cases in the data
condition negatives (N)
the number of real negative cases in the data

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Wikipedia

Performance of regression

- For regression: usually the mean square error (MSE) or the residual sum of squares (RSS) are used. If our dataset (training or testing) contains N samples (epochs) of the form $(y_i, \mathbf{x}_i)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is the regression map then the associated (training or testing) MSE is:

$$\mathrm{MSE}_f := \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 ,$$

  while the associated (training or testing) RSS is:

$$\mathrm{RSS}_f := \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 .$$

What would you choose as performance for the following tests / systems?

- COVID 19 test
- Pregnancy test
- EEG (electroencephalogram) signal processing system to detect awareness in comatose patients

Comment: It is often difficult to choose a performance measure that corresponds well to the desired behavior of the system. Examples:

- Transcription task: should we measure the accuracy of the system at transcribing entire sequences, or should we use a more fine-grained performance measure that gives partial credit for getting some elements of the sequence correct?
- Regression task: should we penalize the system more if it frequently makes medium-sized mistakes or if it rarely makes very large mistakes? These kinds of design choices depend on the application.

Comercial and societally relevant applications

- Customer profiling
- Ad placement optimization
- Financial time series prediction
- Control and monitoring of large technological systems (production plants, energy grids, internet)
- Computer games
- Brain-computer interfaces
- Automatic health diagnostic systems
- Surveillance (communication scanning, face recognition, traffic monitoring)
- Military (autonomous missiles and drones, satellite data interpretation, battlefield robotics)
- Speech and language technology

Comercial and societally relevant applications

- **Time series prediction**
  - financial and macroeconomics time series forecasting
  - local weather development (important for short-term power yield prediction in windmill farms)
  - predicting the consequences of action (robot action planning)
- **System control**
  - steering (or auto-piloting) engines and vehicles
  - controlling chemical production plants
- **Fault monitoring**
  - monitor power grids or power plants
  - monitor any technological device
  - driver sleep detection
- **Temporal pattern generation**
  - generating motions of robots and game characters

Example: predicting house prices

Mullainathan, S., and Jann S. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2): 87-106 Harvard, (pdf)

- Consider 10,000 randomly selected owner-occupied units from the 2011 metropolitan sample of the American Housing Survey
- Include 150 variables that contain information about the unit and its location, such as the number of rooms, the base area, and the census region within the United States
- evaluate how well each approach predicts (log) unit value on a separate hold-out set of 41,808 units from the same sample

Example: predicting house prices

Performance of Different Algorithms in Predicting House Values

Method	*Prediction performance (R^2)*		*Relative improvement over ordinary least squares by quintile of house value*				
	Training sample	*Hold-out sample*	1st	2nd	3rd	4th	5th
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]	–	–	–	–	–
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]	–11.5%	10.8%	6.4%	–14.6%	–31.8%
LASSO	46.0%	43.3% [41.5%, 45.2%]	1.3%	11.9%	13.1%	10.1%	–1.9%
Random forest	85.1%	45.5% [43.6%, 47.5%]	3.5%	23.6%	27.0%	17.8%	–0.5%
Ensemble	80.4%	45.9% [44.0%, 47.9%]	4.5%	16.0%	17.9%	14.2%	7.6%

Note: The dependent variable is the log-dollar house value of owner-occupied units in the 2011 American Housing Survey from 150 covariates including unit characteristics and quality measures. All algorithms are fitted on the same, randomly drawn training sample of 10,000 units and evaluated on the 41,808 remaining held-out units. The numbers in brackets in the hold-out sample column are 95 percent bootstrap confidence intervals for hold-out prediction performance, and represent measurement variation for a fixed prediction function. For this illustration, we do not use sampling weights. Details are provided in the online Appendix at http://e-jep.org.

References


S. Haykin. *Neural Networks and Learning Machines.* Pearson, Addison Wesley, 2009.


Herbert Jaeger. *Machine learning.* Course Notes, Jacobs University Bremen.


V. Kotu and B. Deshpande. *Data Science - Concepts and Practice.* Morgan Kaufmann, Elsevier, 2019.


T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.