CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

Introduction to machine learning

Class #8, April 5 2022

Barbora Hladká hladka@ufal.mff.cuni.cz

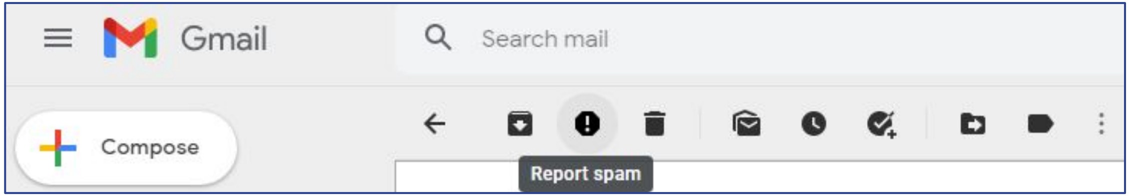
Formal definition of Machine learning (Tom Mitchell, 1997)

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.

I.e., a machine learning problem is defined by a program learning from experience E with respect to a task T while its success is being measured by the performance (measure) P which should improve at performing task T with experience E.

Example :: Spam filtering

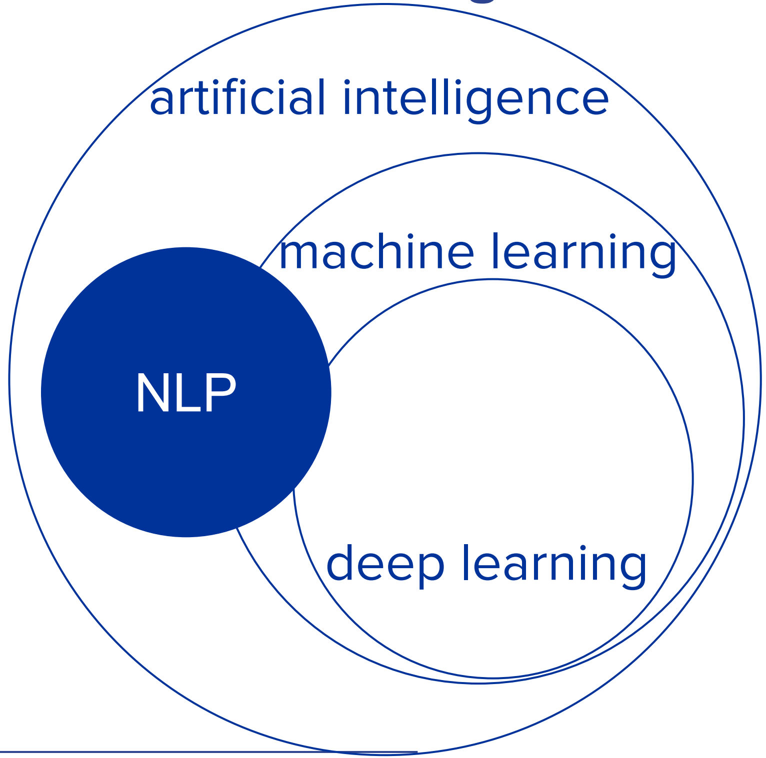
- task T is classifying emails as spam/not spam
- experience E is e.g. in Gmail you classify emails in your inbox
- performance P is measured in the fractionof emails classified correctly



Natural language processing and Machine learning

You’ve already met [UDPipe] and [NameTag] systems of Natural language processing (NLP)

see Lecture #5



Computational linguistics

linguistics

studying
language structure

human
language

computer science

modelling
language use

Computational linguistics

linguistics

studying
language structure

human
language

computer science

modelling
language use

= Natural Language Processing

Natural Language Processing

deals with computer and human interaction in both written and spoken natural language.

uses corpora of texts or speeches, sometimes enriched with linguistic information (annotated corpora).

see Lecture #7

NLP :: Historical excursion

hand-written rules
grammar

machine learning
corpus

deep learning

1950s

rule-based
machine translation
(Wiever, 1949)

1990s

statistical
machine translation
(Brown et al., 1990)

2010s

neural
machine translation
(Goodfellow et al., 2016)

hand-written rules
grammar

machine learning
corpus

deep learning

1950s

1990s

2010s

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F)$$

rule-based
machine translation
(Weaver, 1949)

When we had our first results we submitted them to Coling 1988. Here is a part of the rejection review we received:

> The validity of a statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950 (cf. Hutchins, MT – Past, Present, Future, Ellis Horwood, 1986, p. 30ff and references therein). The crude force of computers is not science. The paper is simply beyond the scope of COLING.
>
> Anonymous Coling review, 1 March 1988

Frederick Jelinek, ACL Lifetime Achievement Award, 2009

Example of an NLP task :: Part-of-speech class labeling

Assigning a part-of-speech (POS) to each word in a text (= tagging)

Time|noun flies|verb like|preposition an|article arrow|noun

Words can have more than one POS

Time|adjective flies|noun like|verb an|article arrow|noun

Teach machines/computers POS labeling

1. Think of POS classes, morphological categories and their values

POS	categories
noun (N)	gender (Masculine, Feminine, …)
adjective (A)	number (Singular, Plural)
pronoun (P)	case (e.g. for Czech, nominative 1, genitive 2, …)
verb (V)	person (1st,2nd,3rd)
number (C)	…
adverb (D)	
preposition (R)	
conjunction (J)	
particle (T)	
interjection (I)	

Teach computers POS labeling

2. Design a tagset

● ● ●

pos number tense

pos = {noun N, adjective A, verb V, preposition R, …}

number = {singular S, plural P, …}

tense = {present P, …}

Teach computers POS labeling

3. Create an exercise book
   i.e. manually annotate as many texts as possible

Nadal	NS-
plays	VSP
with	R--
control	AS-
aggression	NS-
…	…

Teach computers POS labeling

4. Convert the exercise book to frequency tables

Teach computers POS labeling

word	tag	frequency in the exercise book
Nadal	NS-	10 (Nadal was labeled with NS- 10 times)
plays	AS-	15
plays	VSP	15
with	R--	40
control	NS-	5
control	VPP	13
…	…	…

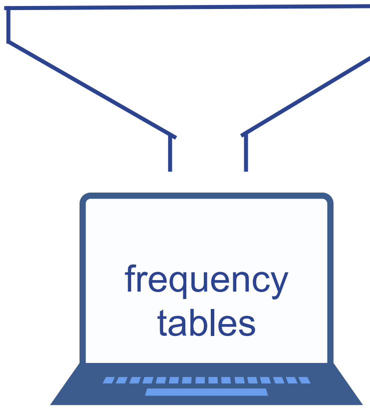
Teach computers POS labeling

tag_{i-1}	tag_i	frequency in the exercise book
NS-	VSP	10 (NS- was followed by VSP 10 times)
NS-	NP-	1
VSP	R--	15
R--	NS-	8

- tag bigram tag_{i-1} tag_i = a sequence of two adjacent tags in annotated texts
  - tag_{i-1} followed by tag_i
- in general *n*-grams = a sequence of *n* adjacent tags/words/letters/…

CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW

5. Nadal plays tennis



Teach computers POS labeling

Nadal	NS-	10
plays	NS-	15
plays	VSP	15
with	R--	40
control	NS-	5
control	VPP	13
…	…	…
NS-	VSP	10
NS-	NP-	1
VSP	R--	15
R--	NS-	8
…	…	…

Nadal	NS-
plays	VSP
tennis	R--

tennis and preposition??

That's statistical machine learning

This is machine learning! Supervised machine learning.

= a teacher supervising the entire learning process. We, teachers, create examples and tell machines how to learn from them.

Supervised machine learning in details

real world objects

words in text

passengers on Titanic

What computers extract from examples?

- Humans use their reason, intuition, and their real world knowledge.
- Computers need to extract a set of useful context clues that are then used for a given task.

- Formally, the context clues are called attributes or features and should be exactly and explicitly defined.
- Then each object (e.g. a word) is characterized by a list of features, which is called feature vector.

Features

= are properties of objects that we can observe or measure.

Feature values can be of several types

- numerical
  - either discrete or continuous
- categorical
  - any list of discrete values, non-numerical
- binary (0/1, Yes/No)
  - can be viewed as a kind of categorical

Features :: Titanic dataset

numerical discrete

categorical

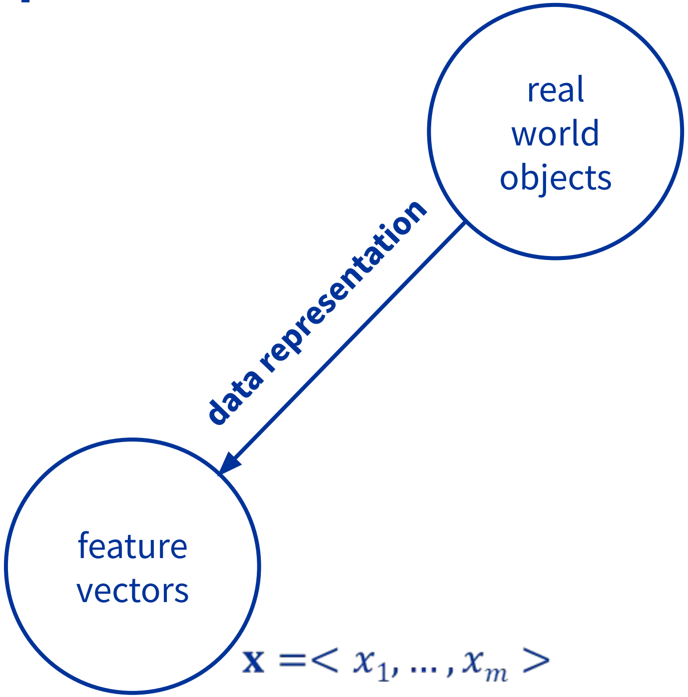
binary

numerical continuous

10 feature vectors

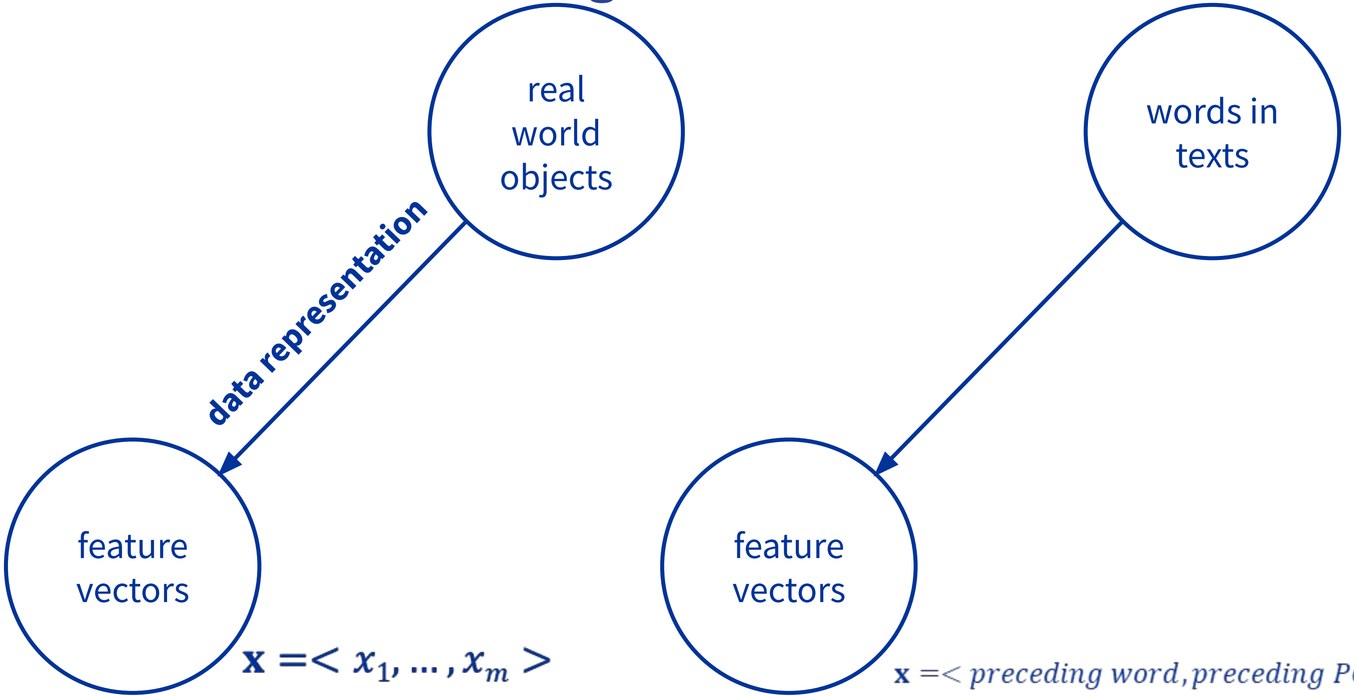
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C

Supervised machine learning



$\mathbf{x} =< x_1, \dots, x_m >$

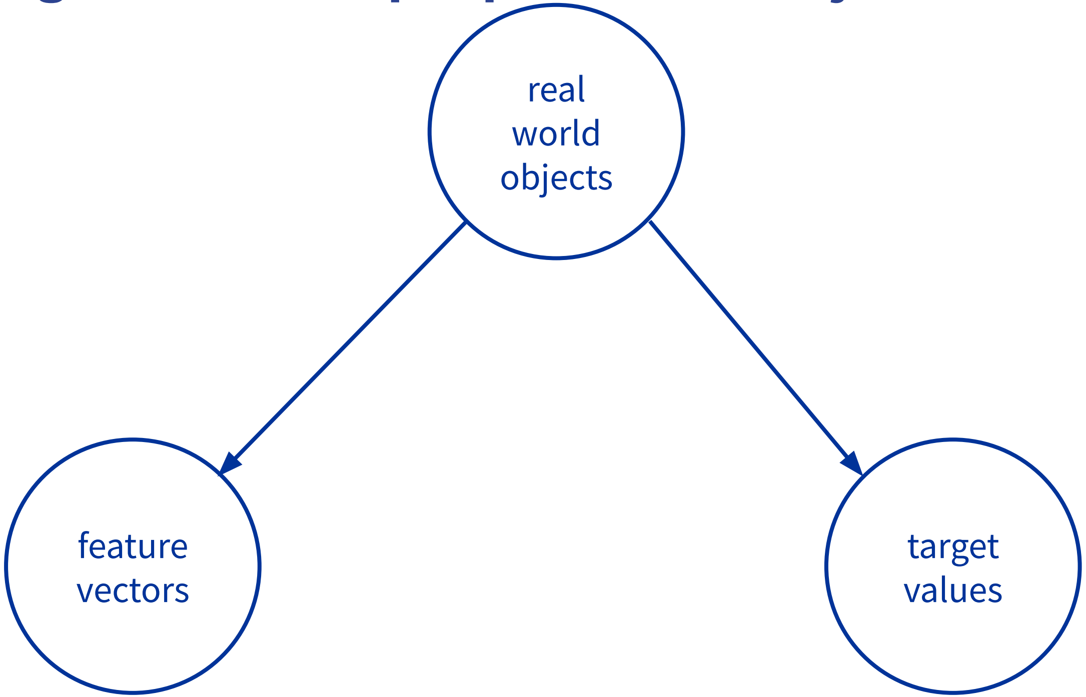
Features :: POS labeling



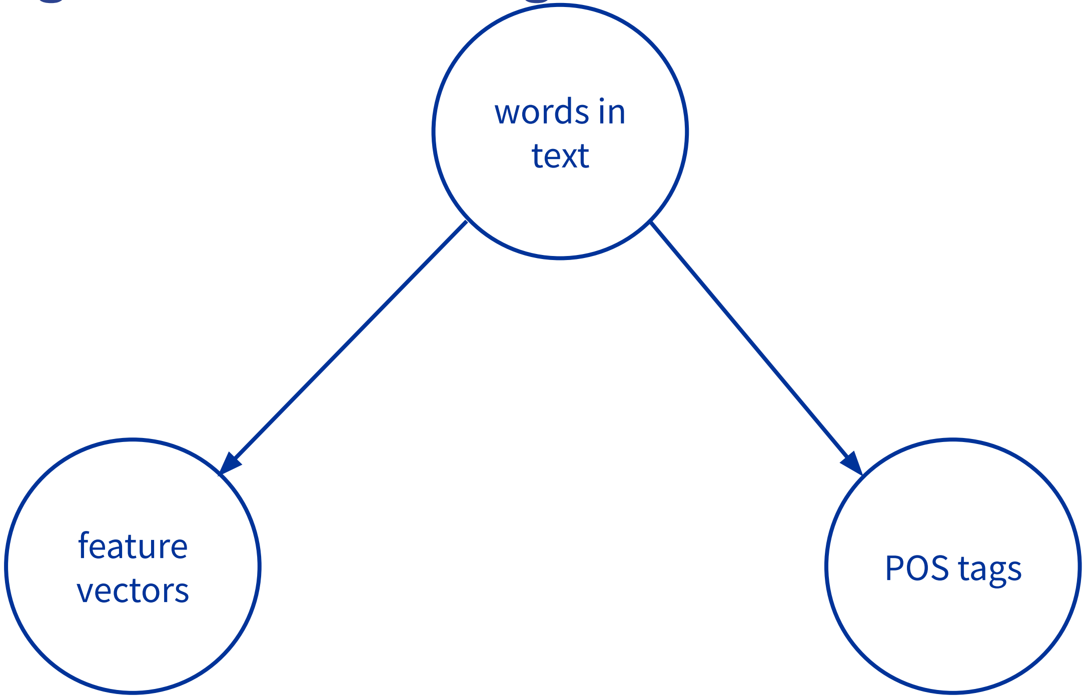
$\mathbf{x} =< x_1, \dots, x_m >$

$\mathbf{x} =< preceding\ word, preceding\ POS\ tag, word\ ending, \dots >$

Target values = properties of objects to be predicted



Target values :: Categorical



Target values :: Categorical :: Binary

passengers

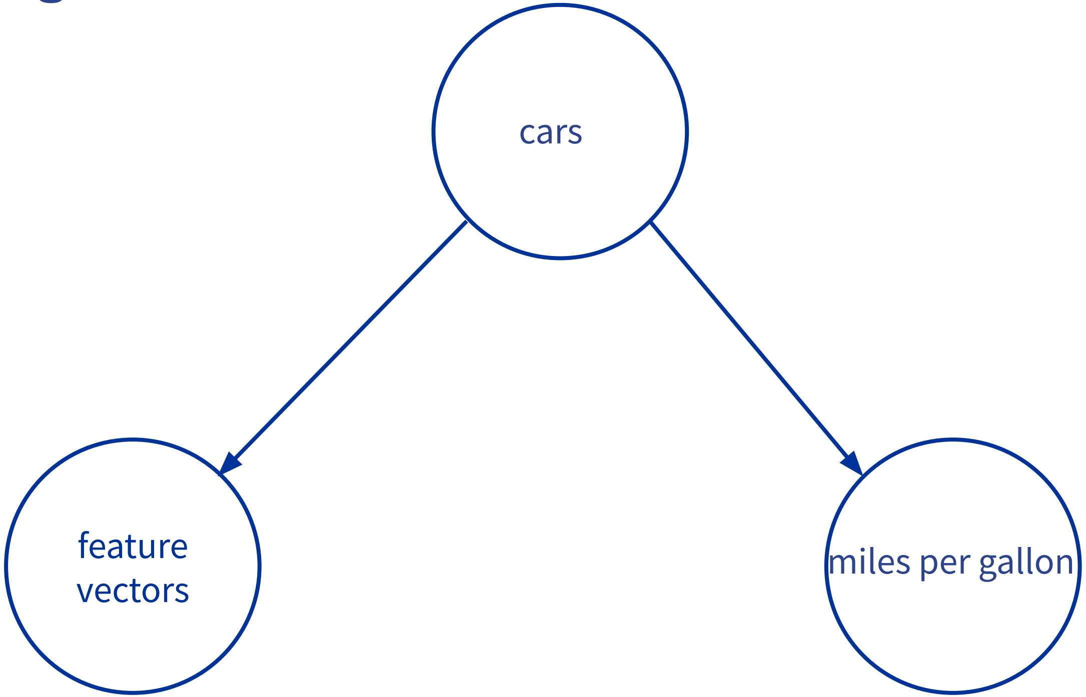
feature vectors

Survived
0/1

Filter

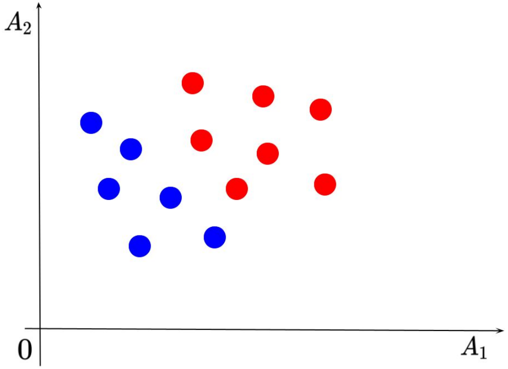
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	[illegible]	0	3	Moran, Mr. James	[illegible]	[illegible]	0	0	330877	8.4583		Q
7	[illegible]	[illegible]	1	McCarthy, Mr. Timothy J	[illegible]	[illegible]	[illegible]	0	17463	51.8625	E46	S
8	[illegible]	[illegible]	3	Palsson, Master. Gosta Leonard	[illegible]	[illegible]	[illegible]	1	349909	21.0750		S
9	[illegible]	[illegible]	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	[illegible]	[illegible]	[illegible]	2	347742	11.1333		S
10	[illegible]	[illegible]	2	Nasser, Mrs. Nicholas (Adele Achem)	[illegible]	[illegible]	[illegible]	0	237736	30.0708		C

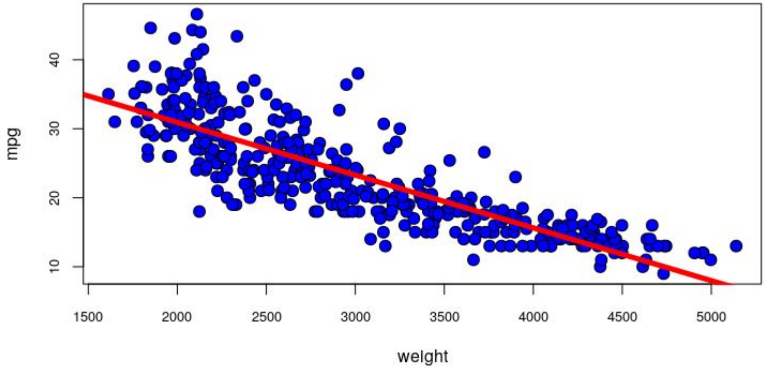
Target values :: Numerical



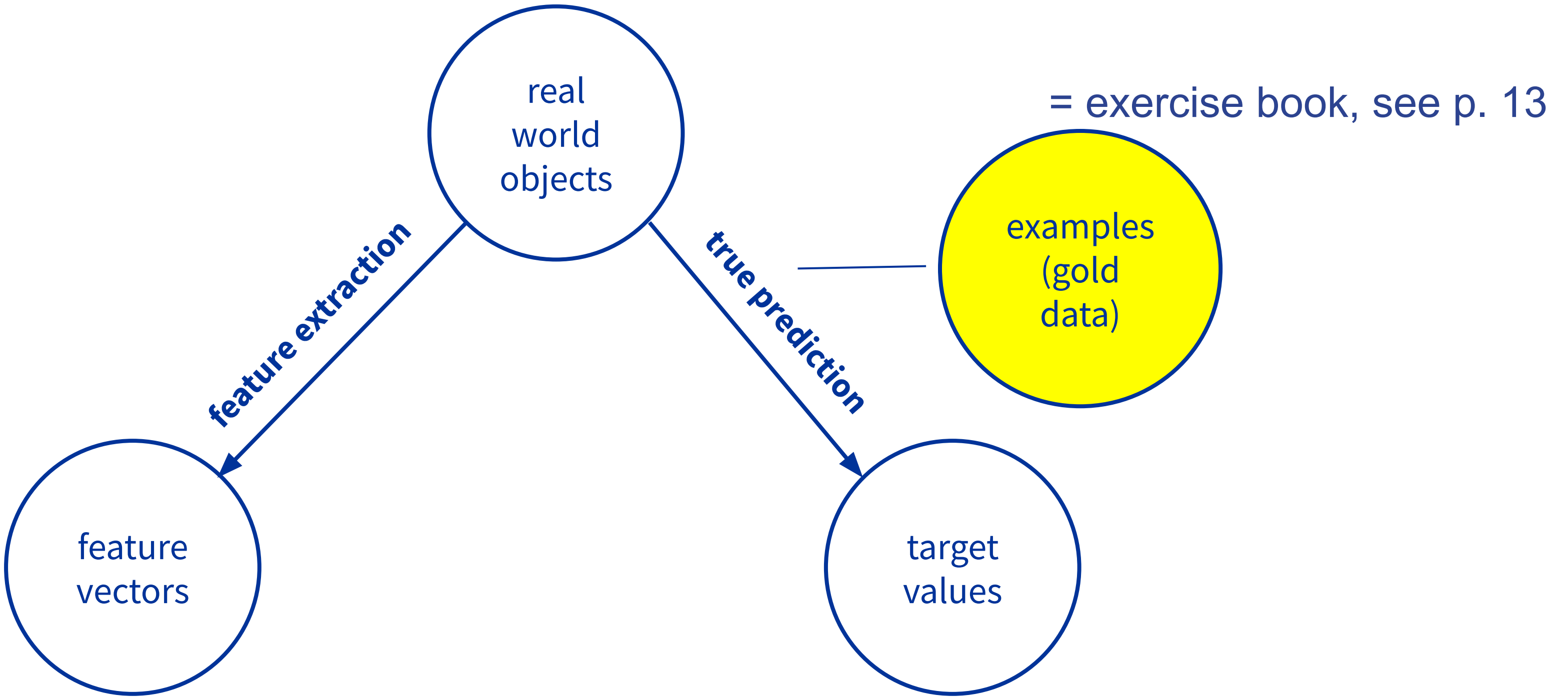
Supervised machine learning

- classification if target values are categorical
- regression if target values are numerical

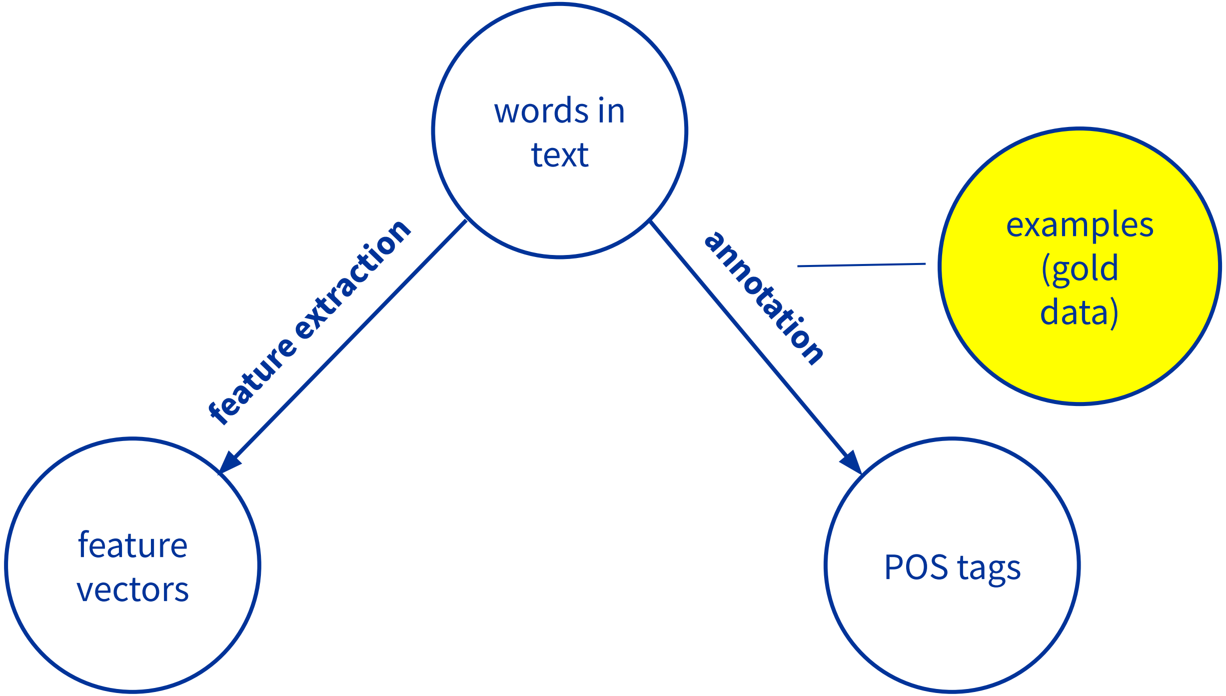




Gold data

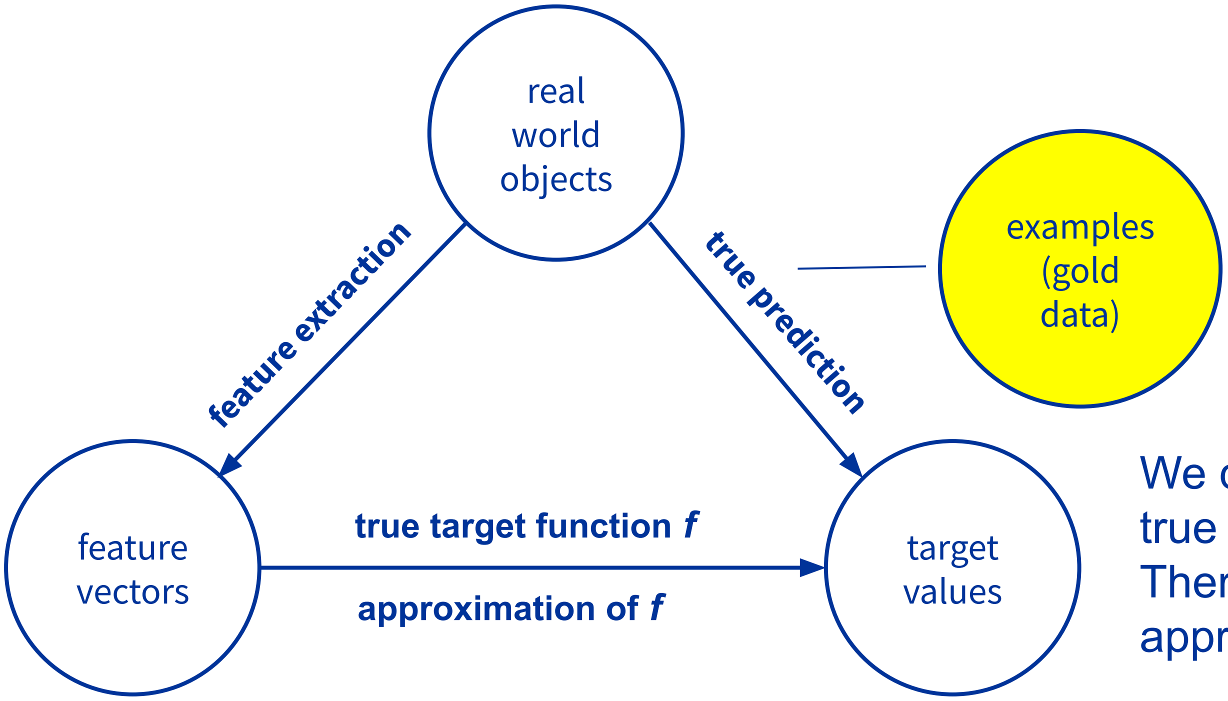

real
world
objects
feature extraction
true prediction
= exercise book, see p. 13
examples
(gold
data)
feature
vectors
target
values

Gold data



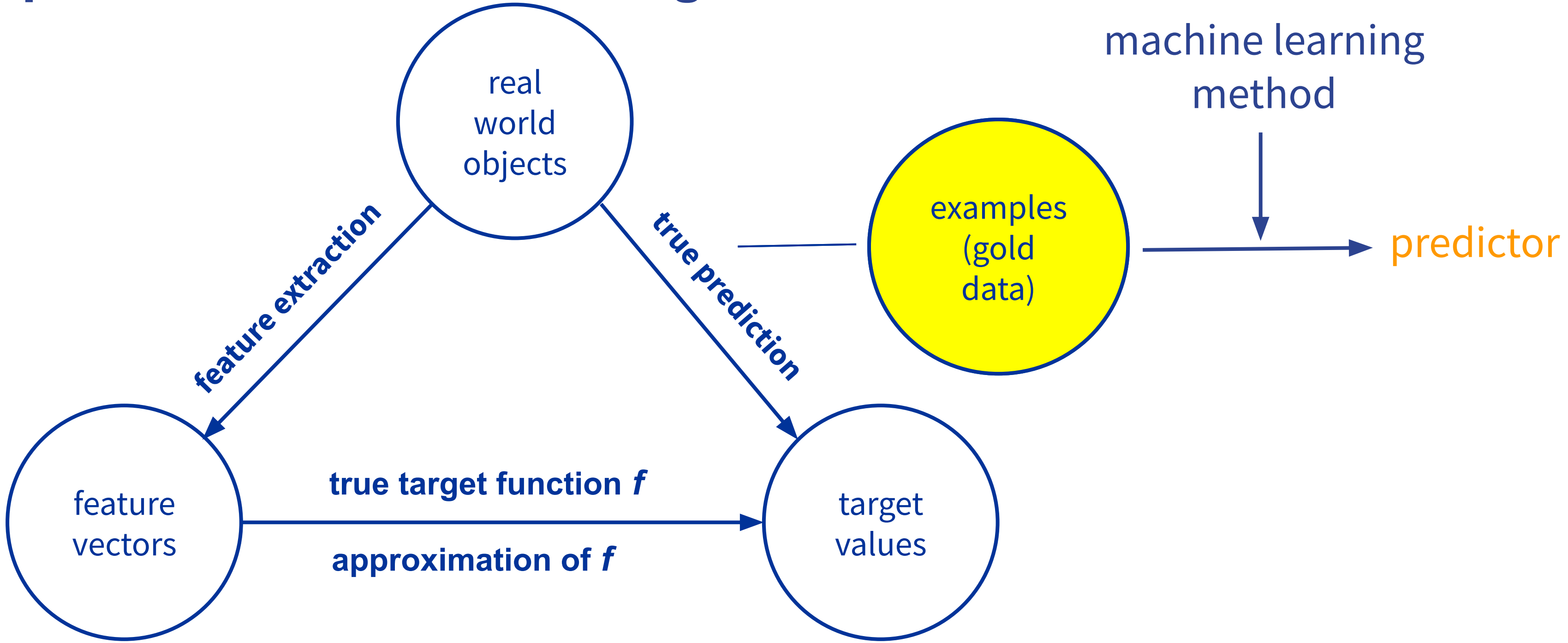
Nadal	NS-
plays	VSP
with	R--
control	AS-
aggression	NS-
…	…

Predictor



We do not know the true target function *f*. Therefore we approximate it.

Supervised machine learning

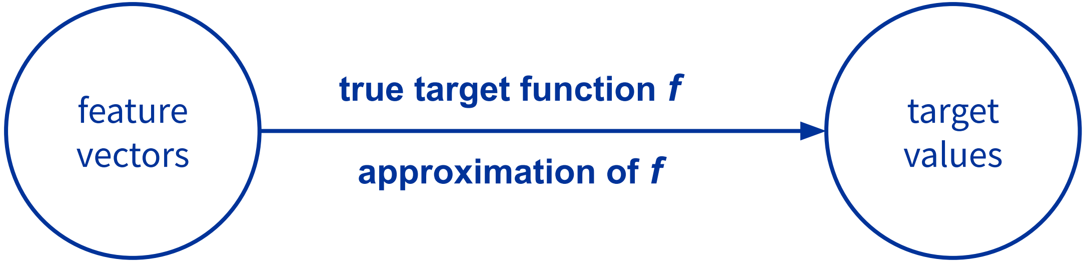


Machine learning process

1. Formulating the task
2. Getting examples (gold data)
   Splitting them into training and test subset
3. Learning from training examples using ML algorithms > model/predictor
4. Testing the learned knowledge on test examples
5. Evaluation

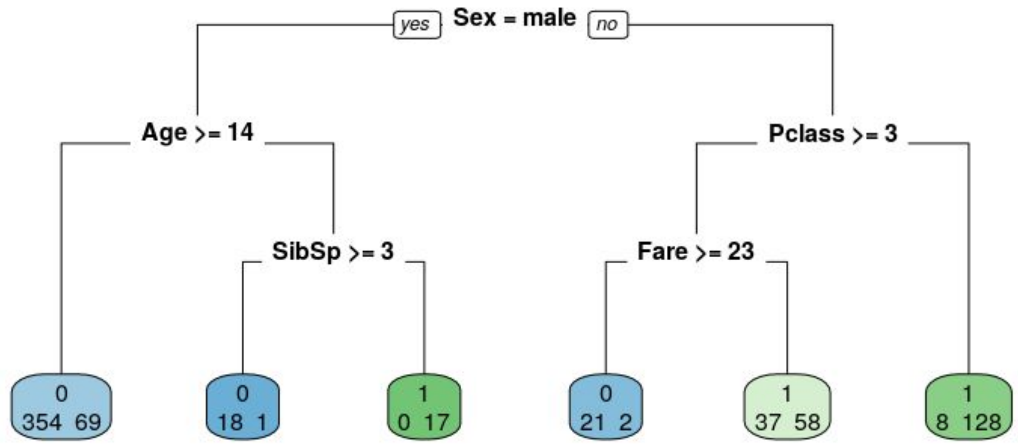
Decision tree learning algorithm

- Traditional supervised machine learning algorithm
- Decision trees represent functions that map feature vectors to target values



Decision tree learning algorithm

training → model m2.1.pred



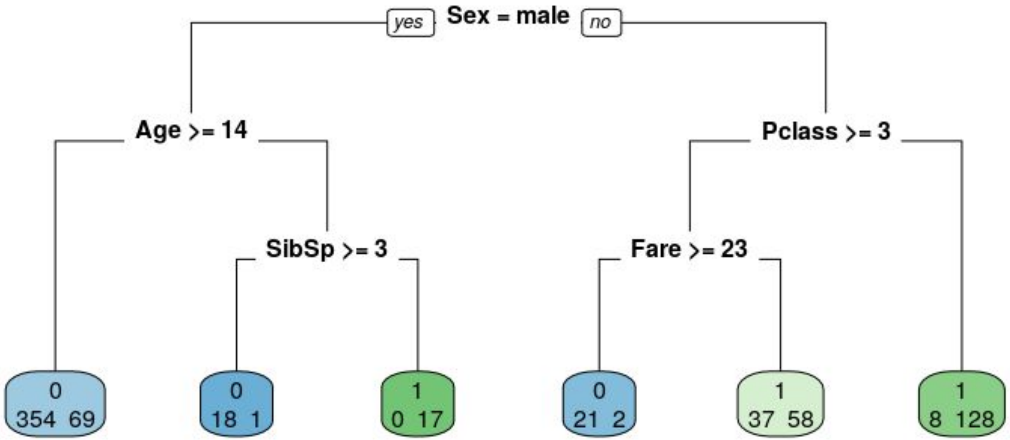
Titanic

Tree structure description

- Nodes
  - Root node
  - Internal nodes
  - Leaf nodes with target class values
- Decisions
  - Binary questions on a single feature, i.e. each internal node has two child nodes

Decision tree learning algorithm

training → model m2.1.pred



Titanic

evaluation on test data

```
m2.1.pred  0  1
         0 98 20
         1 13 47
```

confusion matrix

Evaluation of binary classifiers :: Confusion matrix

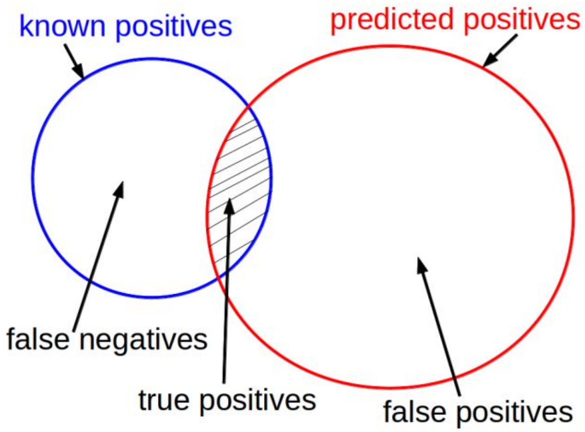
		Predicted class	
		Positive	**Negative**
True class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Explanation

- **'Trues'** are examples correctly classified
- **'Falses'** are examples incorrectly classified
- **'Positives'** were predicted as positives (correctly or incorrectly)
- **'Negatives'** were predicted as negatives (correctly or incorrectly)

Confusion matrix :: Sensitivity TP/(TP+FN)

= ability of a classifier to correctly identify positive examples (e.g., survivors, patients with a disease)



Generalization

= model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to train the model

HW #5 assignment :: Subject labeling in an EU regulation

- https://ufal.mff.cuni.cz/courses/npfl134/subjann