Introduction

Machine Learning

Ivan Smetannikov

1.07.2019

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Parts of ML course

ML course consists of:

1. Theory-oriented track
2. Problem-oriented track

Parts of ML course

ML course consists of:

1. Theory-oriented track
2. Problem-oriented track

Yes, you will have a test in the end!

Introduction track

1. Introduction to machine learning
2. Introduction to data science
3. Introduction to Python

Method-oriented track

Starts **strictly** at 11:00

11:00 – 12:30 — Lecture
12:40 – 14:10 — Lecture
14:10 – 15:00 — Lunch Break
15:00 – ~18:00 — Seminars

Lectures are for learning new methods
Seminars are for applying them in practice

ML general Schedule

1) Supervised learning
2) Model selection and optimization
3) Unsupervised learning
4) Neural networks
5) Data science and semi-supervised learning

And the test in the end!

FITP and ML lab

IT and Programming Faculy

1) More than 400 students on bachelor degree each year

2) ICM ICPC World champions 7 times (more than anyone)

3) More than 200 full-time scientists in programming, machine learning, cyber–physical systems, optimization and other IT fields

FITP and ML lab

Machine Learning lab

1) Almost 20 employees including 2 PhDs and 5 PhD students

2) More than 200 ML-related project for the last 5 years

3) More than 20 Scopus/WoS papers a year

4) ML schools, courses and even English master program

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Machine learning definitions

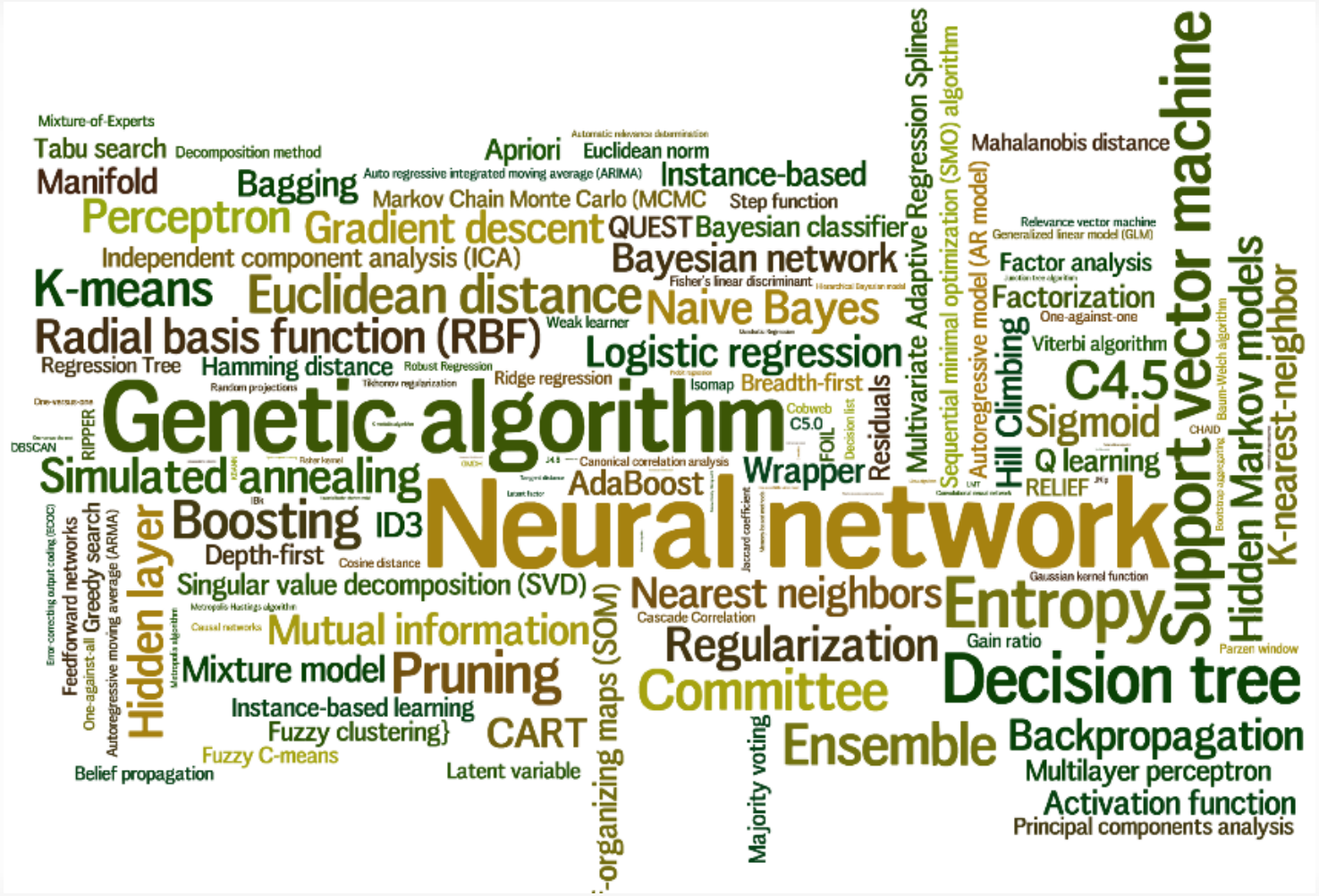
Machine learning is a process (field of study) that gives computers ability to learn without being explicitly programmed.

A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

A computer program is said to be **learnt** from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

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

Machine Learning Approaches



Machine Learning Applications

Related fields

- Pattern recognition
- Computer vision
- Data mining
- Informational Retrieval
- Natural Language Processing
- Neural Computation
- …

Related concepts

- Artificial intelligence
  Strong AI (AGI) vs Weak AI
- Intellectual systems
  Expert system vs ML systems
- Mathematical modelling

- Way of knowledge representation and using

Knowledge vs data

Knowledge ≠ data

Knowledge consists of patterns in a certain domain (principals, regularities, relations, rules, laws), gained with practice and professional experience, which helps to formulate and solve problems in a certain field.

Machine Learning vs Data Mining

Formally, DM is a step in **Knowledge discovery in databases** (KDD). Usually, these two terms are synonyms.

1. Collect data
2. Pre-process data
3. Apply machine learning algorithms

Required background

- Probability theory and mathematical statistics
- Optimization
- Computational science
- Linear algebra
- Discrete math
- Computational complexity theory
- …

Machine learning problems

- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Semi-supervised learning
- Active learning
- Online learning
- Structured prediction
- Model selection and tuning

Supervised learning

Set of examples with answers is given.
A rule for giving answers for all possible examples is required:

- classification;
- regression;
- learning to rank;
- forecasting.

Unsupervised learning

Set of examples is given, but no answers. A rule for finding answers or some regularity is required:

- clustering;
- dimension reduction;
- association rules learning;
- model selection (very general problem).

Model selection and tuning

How to choose an algorithm?

There are many parametrized families of algorithms. You should choose both a family (model selection) and its parameters (tuning).

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Examples (1/3)

1. Medical diagnosis problem
For a patient, decide, what is his/her illness, risks and treatment.

2. Credit scoring
For an applicant, decide if he or she return a credit.

3. Spam filtering
For a letter, decide if it is spam or not.

4. Documents categorization
For a document, pick categories, to which it belongs, or topics that are represented in it.

Examples (2/3)

5. Immobile property cost forecasting
For a house or land, predict its cost or factors that have impact on its cost.

6. Sales rate forecasting
For history of sales, predict how much a certain shop will sell goods or how many certain goods will be sold.

7. Search engine results ranking
For a search query, return the most relevant links.

8. Collaborative filtering
For a user, determine his preferences (movies, books, music, goods).

Examples (3/3)

9. Detecting consumers categories
For a set of consumers, find groups with similar degree of interest in a certain product.

10. Signature authentication
For someone's signature, define if it is real or fake.

11. Forecasting stock indices
Predict values and dynamics of stock indices.

12. Computational synthesis of drugs
Predict if a molecule can be used in a certain drug.

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Books

1. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer, 2009

2. Bishop C.M. Pattern recognition and machine learning. Springer, 2006.

3. Mitchell T. Machine learning. McGraw Hill, 1997.

4. Vapnik V.N. The nature of statistical learning theory. NY: Springer, 1995.

5. Rassel S., Norvig P. Artificial Intelligence: Modern Approach. Prentice Hall Inc., 1995.

6. Givens G.H., Hoeting J.A. Computational Statistics, 2nd Edition. Wiley, 2012

Web sources

MOOC courses (coursera.org):

- A. Ng “Machine Learning”
- D. Koller “Probabilistic Graphical Model”
- G. Hinton “Neural Networks for Machine Learning”

YouTube courses:

- N. de Freitas “Deep Leaning” (at Oxford, 2015)

https://www.youtube.com/playlist?list=PLE6Wd9FR--EfW8dtjAuPoTuPcqmOV53Fu

- N. de Freitas “Machine Learning” (at UBC, 2013)

https://www.youtube.com/playlist?list=PLE6Wd9FR--EdyJ5lbFl8UuGjecvVw66F6

- A. Ng “Machine Learning (at Stanford, 2014)

https://www.youtube.com/playlist?list=PLA89DCFA6ADACE599

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Optimization problem (simple)

Optimization is the process of finding the minima or the maxima of a function.

A point x^* is a **global maximum** if $f(x) \leq f(x^*)\ \forall x$ and is a **global minimum** if $f(x) \geq f(x^*)\ \forall x$.

Consider $f: \mathbb{R} \to \mathbb{R}$, such that f' and f'' are continuous.
Then necessary conditions for maximum are:

1. $f'(x) = 0$.
2. $f''(x) \leq 0$.

Optimization problem (general)

1. It is not necessary defined on $\mathbb{R}$.
2. It is not univariate.
3. Solutions for $f'(x) = 0$ may be hard to compute.
4. It may be not smooth.
5. We may lack time.

Optimization problem (general)

Assume f, g and h_j are defined on a variable space X, $x \in X$. Then, the problem is stated as follows:

$$\begin{cases} f(x) \to \min_x \\ g_i(x) \leq 0, \\ h_j(x) = 0. \end{cases} \qquad i = 1, \dots, m; j = 1, \dots, k.$$

Optimization methods

- Bisection method
- Fixed point integration
- Gradient descent
- Newton's method
- Coordinate descent
- Maximum likelihood
- …

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Supervised learning

We are going to talk about supervised learning most of the time

Prediction task

- The most popular task in machine learning and data analysis that includes classification and regression
- In most of the cases, it can be solved without expertise in domain
- In philosophy of science, verification and falsification are about prediction-making
- Data mining is data analysis are strongly associated with predictions because of data mining platforms

Data mining platforms

АлгоМост

CodaLab

INNOCentive

CHALEARN

KDDCup

kaggle

DRIVENDATA

Интернет-математика

topcoder

Challenge.gov

datascience.net

TunedIT

Consequences of focusing on prediction task

Positive

- Very effective algorithms and useful tricks
- Understanding of what usually works

Negative

- Zoo of models, methods and tricks that are extremely costly to be learnt
- Interpretability suffers
- Provability suffers

The problem

X is **object set**, or input set;
Y is **label set**, or **answer set,** or output set;
$y : X \to Y$ is unknown **target function (dependency)**.
$\{x_1, \ldots, x_\ell\} \subset X$ is **training sample**;
$y_i = y(x_i),\ i = 1, \ldots, \ell$ are **known values** of the function.

Problem: find $a : X \to Y$, **solving function** (decision function), which approximates y on X.

We are going to speak only about **algorithms**.
What is the difference between algorithms and functions?

Main questions

1. How are the objects described?
2. How do the answers look like?
3. What is algorithm set from which a is being chosen?
4. How to measure quality of how a approximates y?

How are the objects described?

$f_j : X \to D_j$, $j = 1, \dots, n$ are **features** or **attributes**.

Feature types:

- **binary**: $D_j = \{0, 1\}$;
- **categorical**: D_j is finite;
- **ordinal**: D_j is finite and ordered;
- **numerical**: $D_j = \mathbb{R}$.

Features data

$\big(f_1(x), \ldots, f_n(x)\big)$ is feature description of an object x.
Object is its feature description.

Data is usually represented with matrix objects-features:

$$F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \cdots & \cdots & \cdots \\ f_1(x_\ell) & \cdots & f_n(x_\ell) \end{pmatrix}$$

How do the answers look like?

Classification:

- $Y = \{-1, +1\}$ — binary;
- $Y = \{1, \ldots, M\}$ — M non-overlapping classes;
- $Y = \{0, 1\}^M$ — M classes that can overlap.

Ranking:

- Y — finite (partially) ordered set.

Regression:

- $Y = \mathbb{R}$ or $Y = \mathbb{R}^m$.

What is algorithm set from which a is being chosen?

Algorithms model is a parametric family of mappings

$$A = \{g(x, \theta) | \theta \in \Theta\},$$

where $g : X \times \Theta \to Y$ is a fixed function, Θ is a set of possible values of the parameter θ.

Example: **linear model** with parameter vector $\theta = (\theta_1, \ldots, \theta_n)$, $\Theta = R^n$.

Which type of problem is the one, where the function is

$$g(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x) \, ?$$

Learning Method

Learning method is a mapping

$$\mu\colon (X \times Y)^{\ell} \to A,$$

which for a certain training set $T^{\ell} = \{(x_i, y_i)\}_{i=1}^{\ell}$ returns an algorithm $a \in A$.

Two steps:

1. **Training**:

with method μ on training set T^{ℓ} build $a = \mu(T^{\ell})$.

2. **Testing**:

apply a for new object x to find answer $a(x)$.

How to measure quality of how a approximates y?

Loss function $L(a, x)$ is the error size of algorithm a on object x

- for classification problem:

$$L(a, x) = [a(x) \neq y(x)],$$

- for regression problem:

$$L(a, x) = d\big(a(x) - y(x)\big),$$

usually quadratic loss function:

$$d(x) = x^2, L(a, x) = \big(a(x) - y(x)\big)^2.$$

Empirical risk is a quality measure of algorithm a on T^ℓ:

$$Q(a, T^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(a, x_i).$$

Empirical risk minimization

Empirical risk minimization method

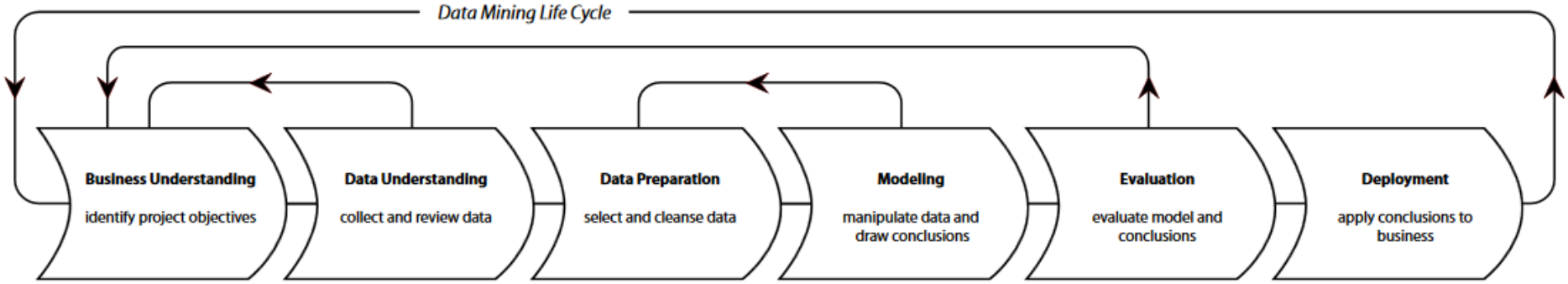
$$\mu\left(T^{\ell}\right) = \operatorname{argmin}_{a \in A} Q(a, T^{\ell}).$$

Decreasing error on train set can lead to a certain problem of lack of generalization.

Lecture plan

- Course organization
- Concept of machine learning
- Examples of ML problems
- Books and materials
- Optimization problem
- Supervised learning
- CRISP-DM methodology

Data Mining Life Cycle



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

Business Understanding

- Business Objectives Determination
- Assessment of Situation, Risks and Costs
- Determination of Data Mining Goals
- Determination of Quality Criteria
- Project Plan Production

Data Understanding

- Initial Data Collecting
- Data Description and Visualization
- Data Exploration
- Data Quality Verification

Data Preparation

- Data Selection and Fusion
- Data Tiding
- Missing Values Completion
- Outlier Detection
- Data Standardization
- Feature Engineering
- Dimensionality Reduction

Modeling

- Model Assessment Design
- Model Selection
- Hyperparameter Optimization
- Overfitting/Underfitting Control and Model Assessment

Evaluation

- Results Evaluation
- Model Approval
- Determination of Next Steps

Deployment

- Deployment
- Monitoring and Maintenance
- Final Report Production
- Review of Project