Lecture 1

Introduction

Intellectual systems
(Machine Learning)

Andrey Filchenkov

06.09.2018

Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

- The presentation is prepared with materials of the K.V. Vorontsov's course "Machine Leaning".
- Slides are available online: **goo.gl/BspjhF**

Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

ML research lab

- Part of Computer technologies centre
- Research interests:
  - meta-learning
  - social network analysis, user profiling
  - natural language processing
  - feature selection
  - image recognition and processing
  - applications to medicine, finances, etc.
  - …

Preliminary lectures plan

- Introduction (1 lecture)
- Classification and regression (6 lectures)
- Optimization (1 lecture)
- Deep networks (3 lectures)
- Unsupervised learning (3 lectures)
- Reinforcement learning (1 lecture)
- Data mining workflow (1 lecture)

Note: it may be changed somehow

Course schedule

Starts at 13:30

Usually, one class is lecture class and another class is practical seminar.

Today you have two lectures.

Many shrifts and changes are expected.

“How can I get grade?”

- Labs
- Quick tests each lecture
- Wiki abstracts on course / course works
- Final test

Ways you cannot improve you grade

You cannot improve your grade by:

- writing a coursework / thesis on machine learning / data mining
- attending lectures
- giving money to teaching staff (we consider only blue chips and immobile property)
- being a relative or a friend of teachers

Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

Machine learning definitions

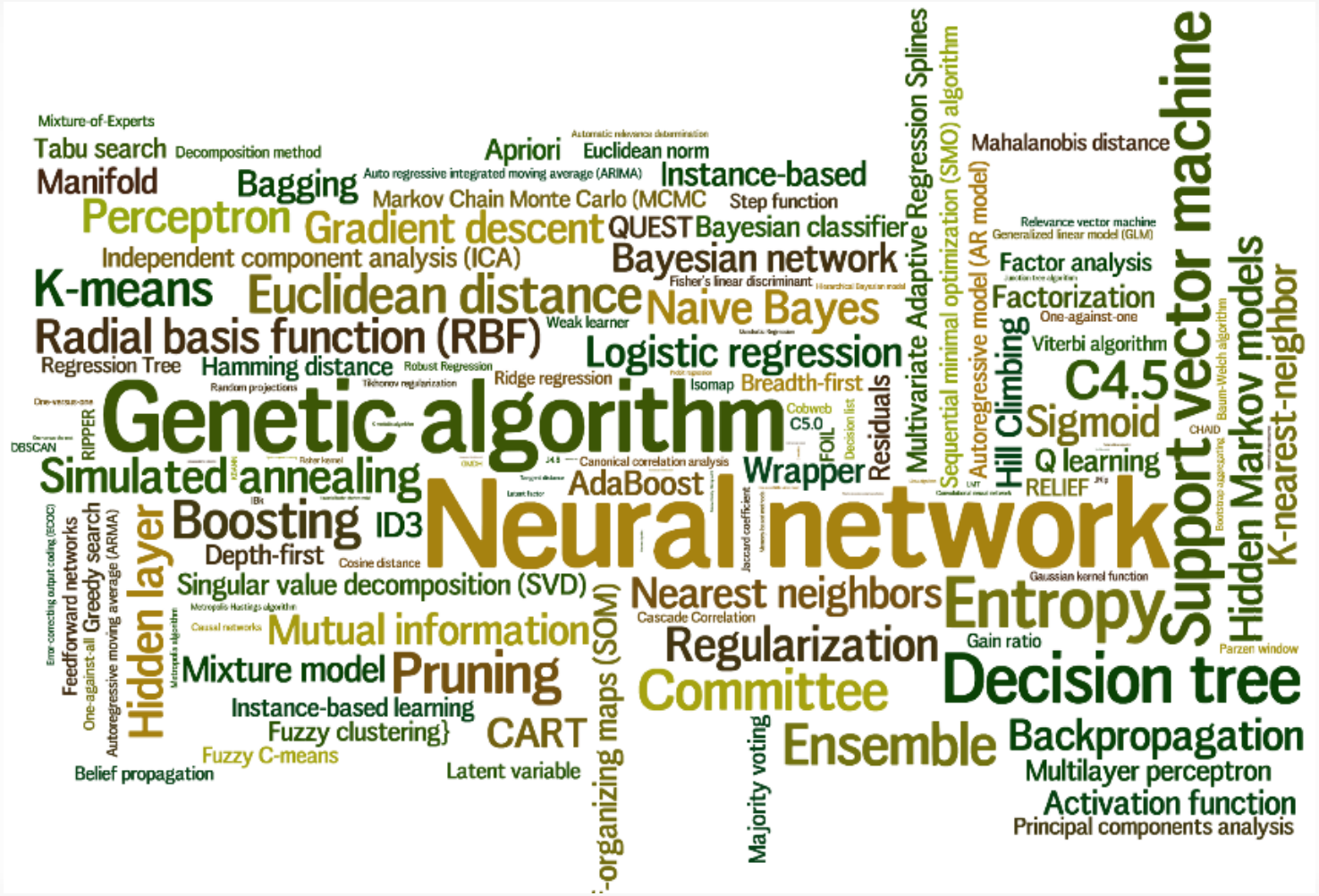
Machine learning is process (field of study) that gives computers ability to learn without being explicitly programmed.

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

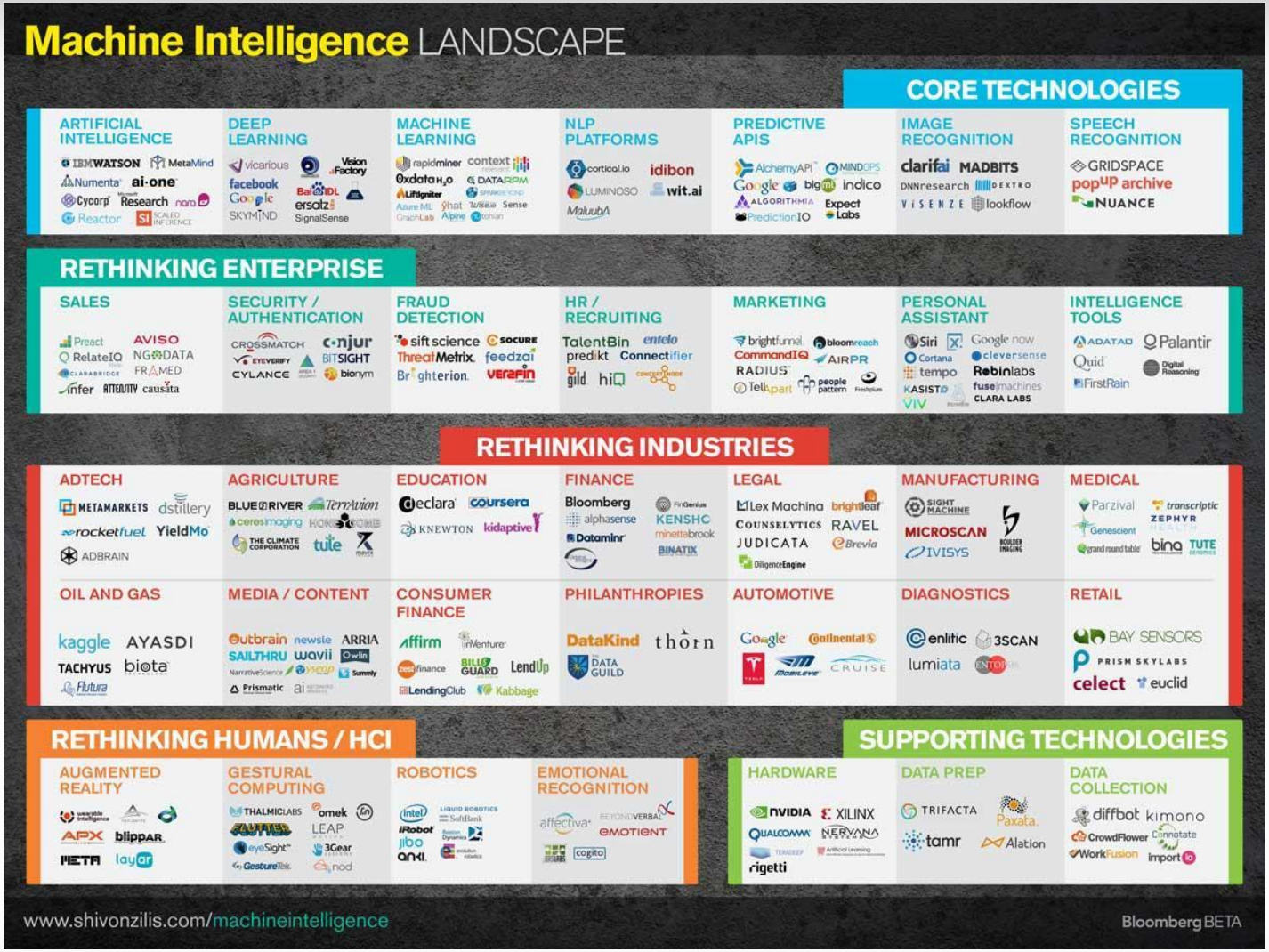
A computer program is said to be **learn** from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

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

Machine Learning Approaches



Machine Learning Applications



Related fields

- Pattern recognition
- Computer vision
- Data mining (DM)
- Information Retrieval (IR)
- Natural Language Processing (NLP)
- Neural Computation
- …

Machine Learning vs Data Science

1. Data collecting
2. Data integration
3. Data warehousing
4. Data analysis
5. High performance computing

Machine Learning vs Data Analysis

Also known as Business Intelligence

1. Exploratory data analysis
2. Confirmatory data analysis (statistical hypothesis testing)
3. Predictive data analysis
4. Data visualization

Related concepts

- Artificial intelligence
  - Strong AI vs Weak AI
- Intellectual systems
  - Expert system vs ML systems
- Mathematical modeling

- Way of knowledge representation and using

Knowledge vs data

Knowledge ≠ data

Knowledge is 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. Engineer features
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
- Semi-supervised learning
- Reinforcement learning
- Active learning
- Online learning
- Structured prediction
- Model selection and validation

Supervised learning

A 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

A set of examples without answers is given. A rule for finding answers or some regularity is required:

- clustering;
- association rules learning;
- recommender systems*;
- dimension reduction**.

Machine learning problems classification

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

Machine learning problems we are to discuss in this course

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

Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

Supervised learning

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

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 set**;
$y_i = y(x_i),\ i = 1, \ldots, \ell$ are **known values** of the function.

Problem: find $a : X \to Y$ that is **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 the algorithm set, from which we chose a?
4. How to measure quality of how well a approximates y?

How are the objects described?

$f_j : X \to D_j, j = 1, \ldots, n$ 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 a feature description of object x. Object is its feature description.

Data are usually represented with "objects-features" matrix (feature data):

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

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 is 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 parameter θ.

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

Which type of problem is the following, where we chose a function from the following set parametric set:

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

Learning method

Learning method is mapping

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

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

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 an 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 the train set can lead to lacking of generalization.

Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

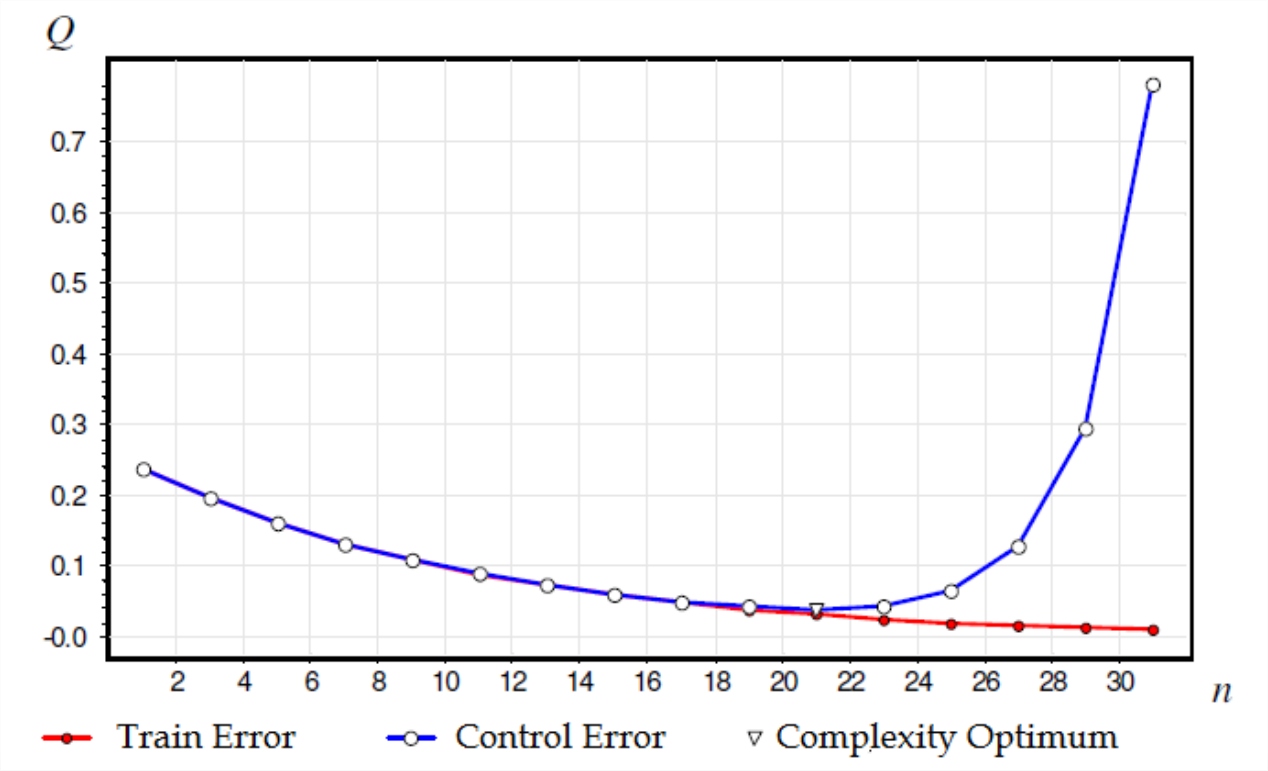
Overfitting problem

Overfitting problem: starting from a certain model complexity level, the better an algorithm performs on train set X^ℓ, the worse it performs on real world objects.

Example of overfitting

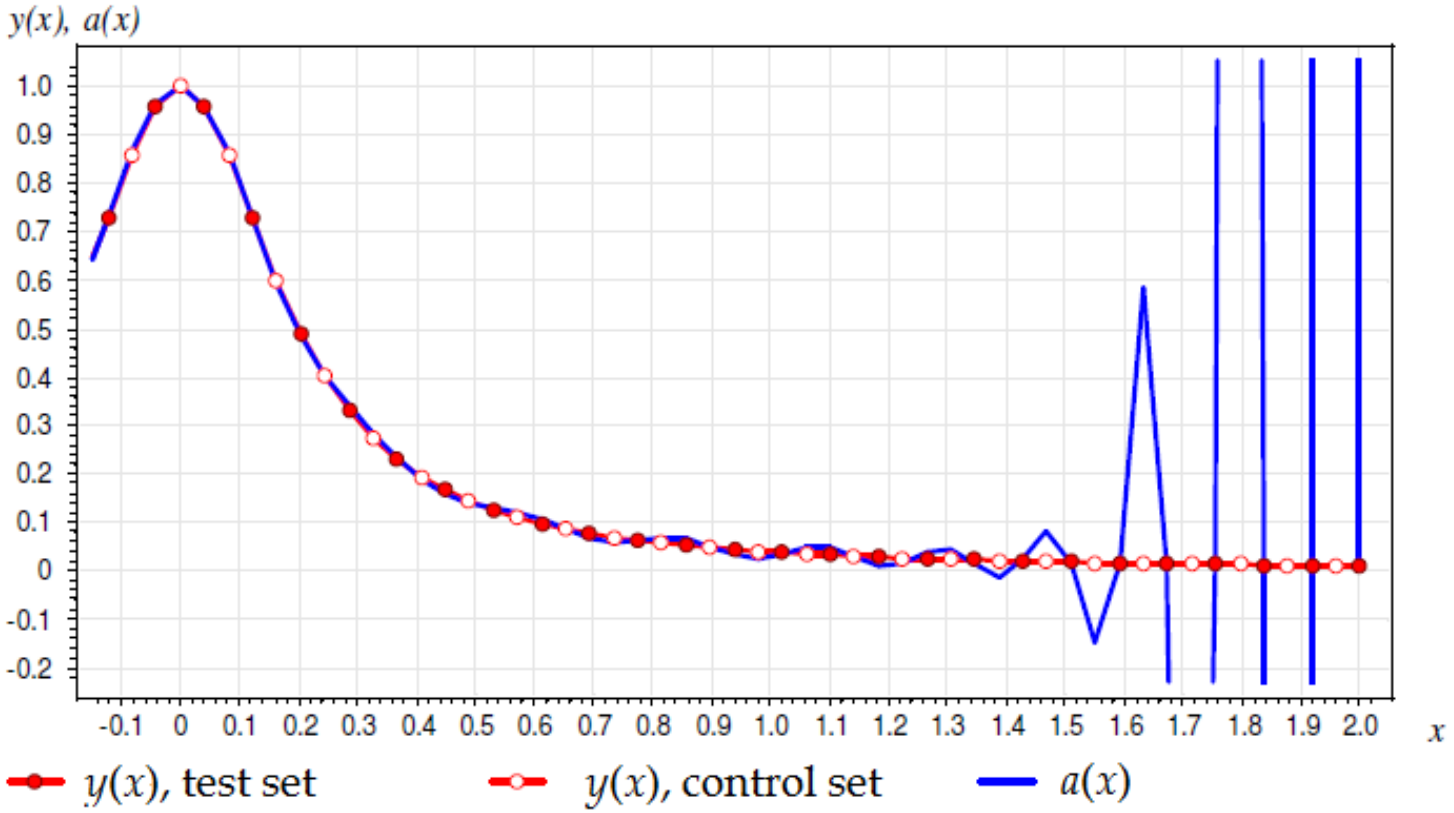
Dependency $y(x) = \frac{1}{1 + 25x^2}$ defined on $x \in [-2, 2]$.

Let search a function among polynomials with degree n.



Overfitted algorithm

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) - \text{polynomial of degree } n = 38$$



Lecture plan

- Organizational questions
- Concept of machine learning
- Supervised learning
- Overfitting and model validation
- Examples

Examples (1/3)

1. Medical diagnosis problem

Decide given a patient, what is his/her illness, risks and possible treatment.

2. Credit scoring

Decide given an applicant, if he/she will return a credit.

3. Spam filtering and malware detection

Decide given a letter, if it is spam or not / Decide given a program if it is a malware

4. Documents categorization and user profiling

Pick categories given documents, to which these documents belong, or topics, which are represented in them. Or which of users will behave in the same way and what are their interests.

Examples (2/3)

5. Learn a model to play Minecraft

Given rules, learn a model to win a game with these rules.

6. Sales rate forecasting

Predict given a history of sales, how much a certain shop will sell goods or how many certain goods will be sold.

7. Search engine results ranking

Given a search query, return the most relevant links.

8. Collaborative filtering

Predict given a user, what is his/her preferences (movies, books, music, goods).

Examples (3/3)

9. Dialog system

Build a system that can conduct a dialog on a given topic or even about everyting.

10. Signature authentication

Define given smb's signature it if is real or fake.

11. Forecasting stock indices

Predict values and dynamics of stock indices.

12. Computational synthesis of drugs or images

Synthesize a drug molecule given constraints or a certain image.

References (1/2)

1. Smola A., Vishwanathan S.V.N. Introduction to Machine Learning http://alex.smola.org/drafts/thebook.pdf

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

3. Courville A., Goodfellow I., Bengio Y. Deep Learning, 2016

4. Nikolenko S.I., Kadurin A.A., Arkhangelskaya E.O. Deep Learning, 2017 (in Russian)

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

References (2/2)

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

2. Duda R. O., Hart P. E., Stork D. G. Pattern classification. New York: JohnWiley and Sons, 2001.

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

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

References (2/2)

MOOC courses (coursera.org):

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

MOOC in Russian

- K.V. Vorontsov “Machine Leaning”.