Yandex School of Data Analysis

Oleg Shpynov
JetBrains Biolabs
07.10.2013

Conference

The Yandex School of Data Analysis conference Machine learning and Very Large Data Sets, which marks the school's fifth anniversary, will be held in Moscow, Russia, on September 27 – October 2, 2013.

Key topics:

- Machine learning
- Application of machine learning in text analysis, computer vision and signal analysis
- Structured data analysis
- Algorithms for big data and large-scale machine learning

Yandex Data Center

- Map Reduce
- MatrixNet – Gradient boosting with random forest
- Erasure coding
- Data Center

Genome

- IBinom project http://ibinom.com/ с использованием платформы YT от Яндекса.
- Сборка генома
- Поиск мутаций с использованием YT
- 80 часов -> 40 минут

VC-complexity

$$R(\alpha) = E\left[\frac{1}{2}|y - f(x,\alpha))|\right]$$

$$R^{emp}(\alpha) = \frac{1}{R}\sum_{k=1}^{R}\frac{1}{2}|y_k - f(x_k,\alpha))|$$

Power of machine

Machine f can **shatter** a set of points $x_1, x_2, .., x_r$ if and only if

For every possible training set of the form $(x_1, y_1), (x_2, y_2), .., (x_r, y_r)$

There exists some value of α that gets zero training error.

Power of machine is the maximum number of points that can be arranged so that f shatter them.

Theorem

$R(\alpha) \leq R^{emp}(\alpha) + \sqrt{\frac{h(log(2R/h)+1)) - log(\eta/4)}{R}}$ с вероятностью $1 - \eta$, где h - power of machine f

VC-complexity

In general the following scheme is true

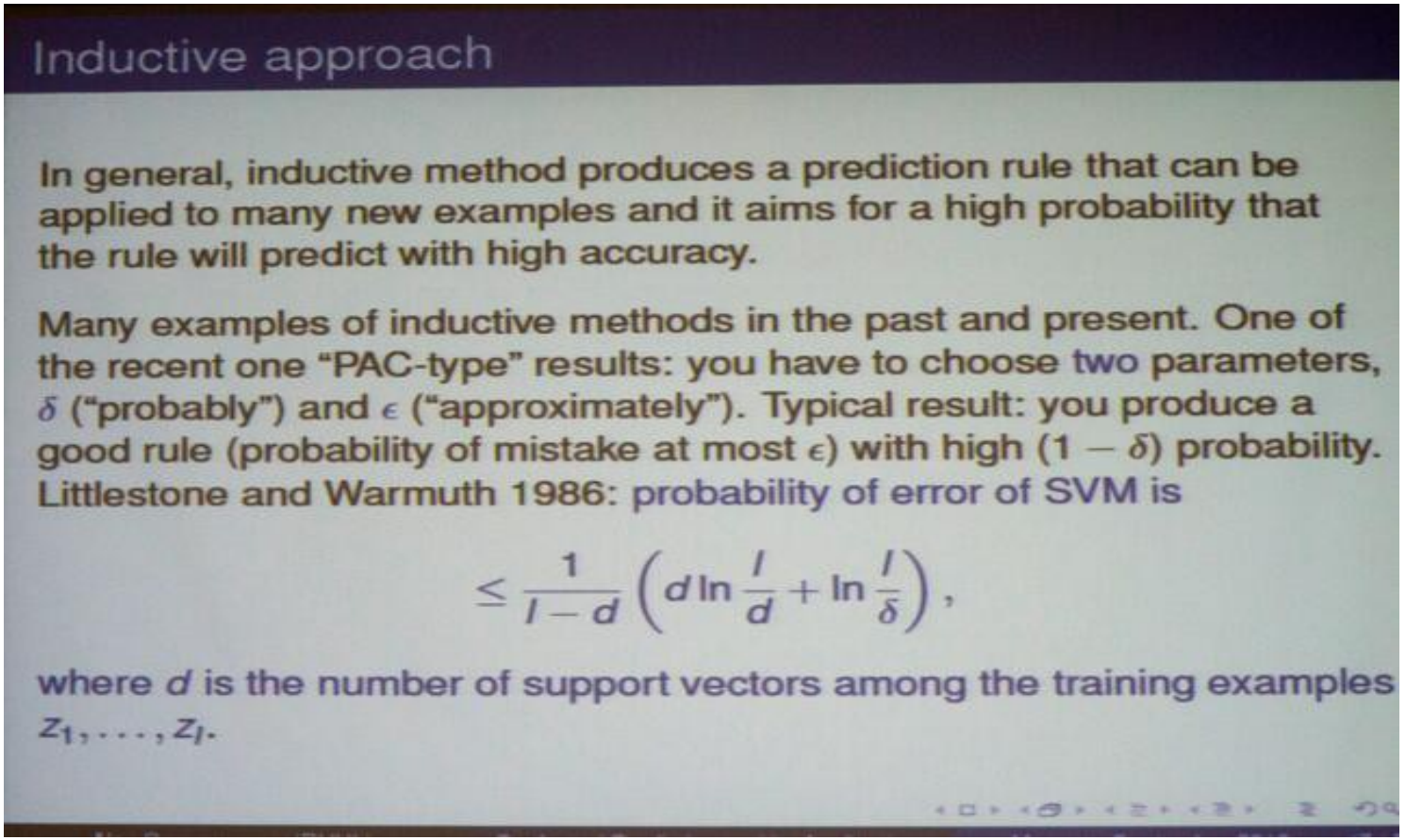
VC-dimension is finite →
Uniform convergence holds →
the system is able to be learned.

But not on the contrary. A system may be able to learn, though Uniform Convergence does not hold, and uniform convergence may hold, though VC-dimension is infinite.

Conformal Predictors

Probabilistic predictor $P(prediction) \geq 1 - \epsilon$

Microarray analysis
http://www-stat.stanford.edu/~tibs/PAM/pam.pdf



Inductive approach

In general, inductive method produces a prediction rule that can be applied to many new examples and it aims for a high probability that the rule will predict with high accuracy.

Many examples of inductive methods in the past and present. One of the recent one "PAC-type" results: you have to choose two parameters, δ ("probably") and ϵ ("approximately"). Typical result: you produce a good rule (probability of mistake at most ϵ) with high $(1 - \delta)$ probability. Littlestone and Warmuth 1986: probability of error of SVM is

$$\leq \frac{1}{l-d}\left(d \ln \frac{l}{d} + \ln \frac{l}{\delta}\right),$$

where d is the number of support vectors among the training examples $z_1, \ldots, z_l$.

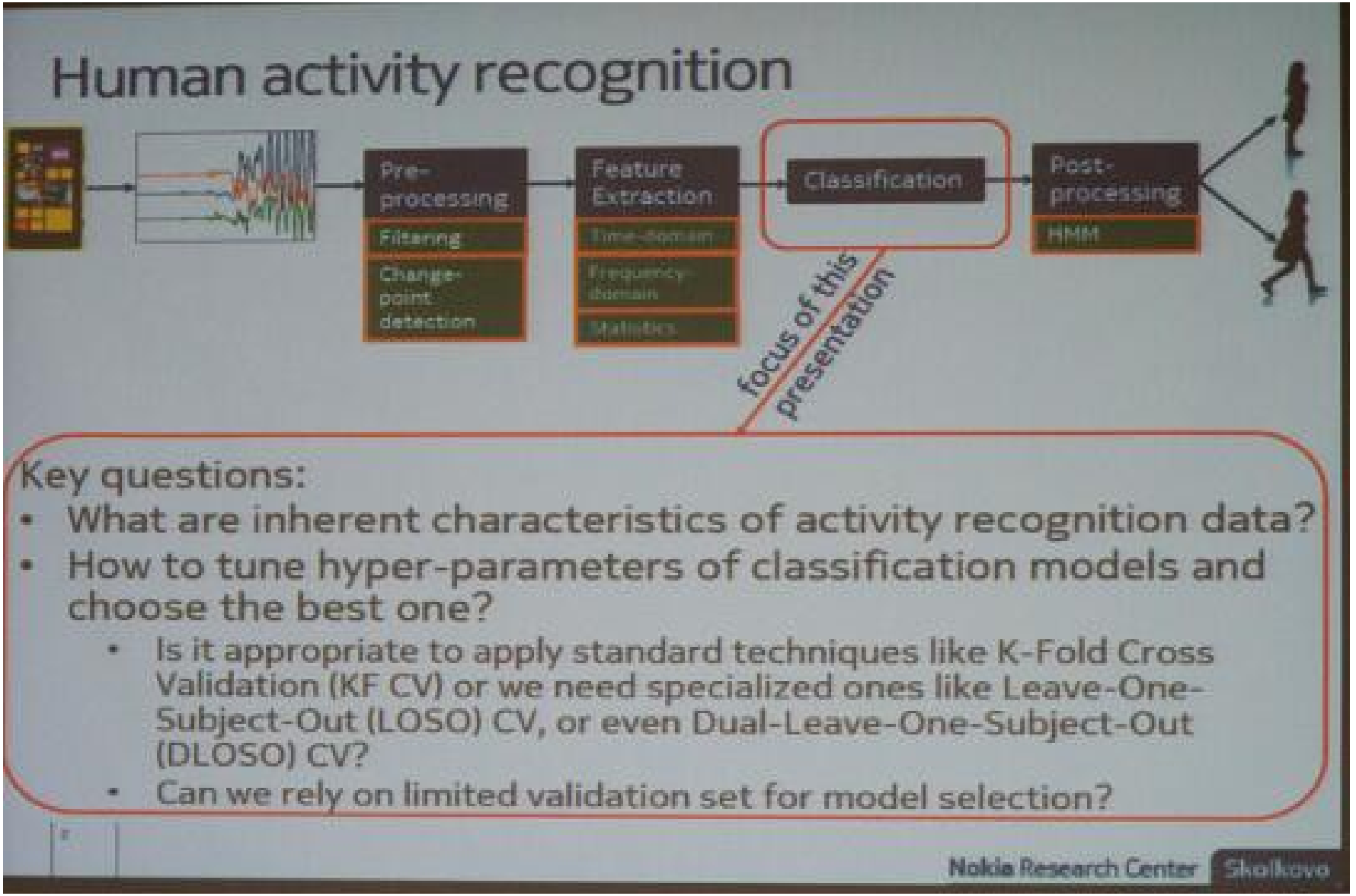
Deep learning

- Gaussian Mixture -> HMM -> Convolution Networks
- **Main idea** P(over-fitting) можно эффективно считать на самых главных классификаторах (нижних слоях)
- Deep Learning Tool: Torch 7 – Neural Networks Library http://www.torch.ch/manual/tutorial/index
- Основная проблема deep learning - очень трудно правильно выбрать классификаторы, и коэффициенты.
- Convolution neural networks explained http://www.youtube.com/watch?v=n6hpQwq7Inw

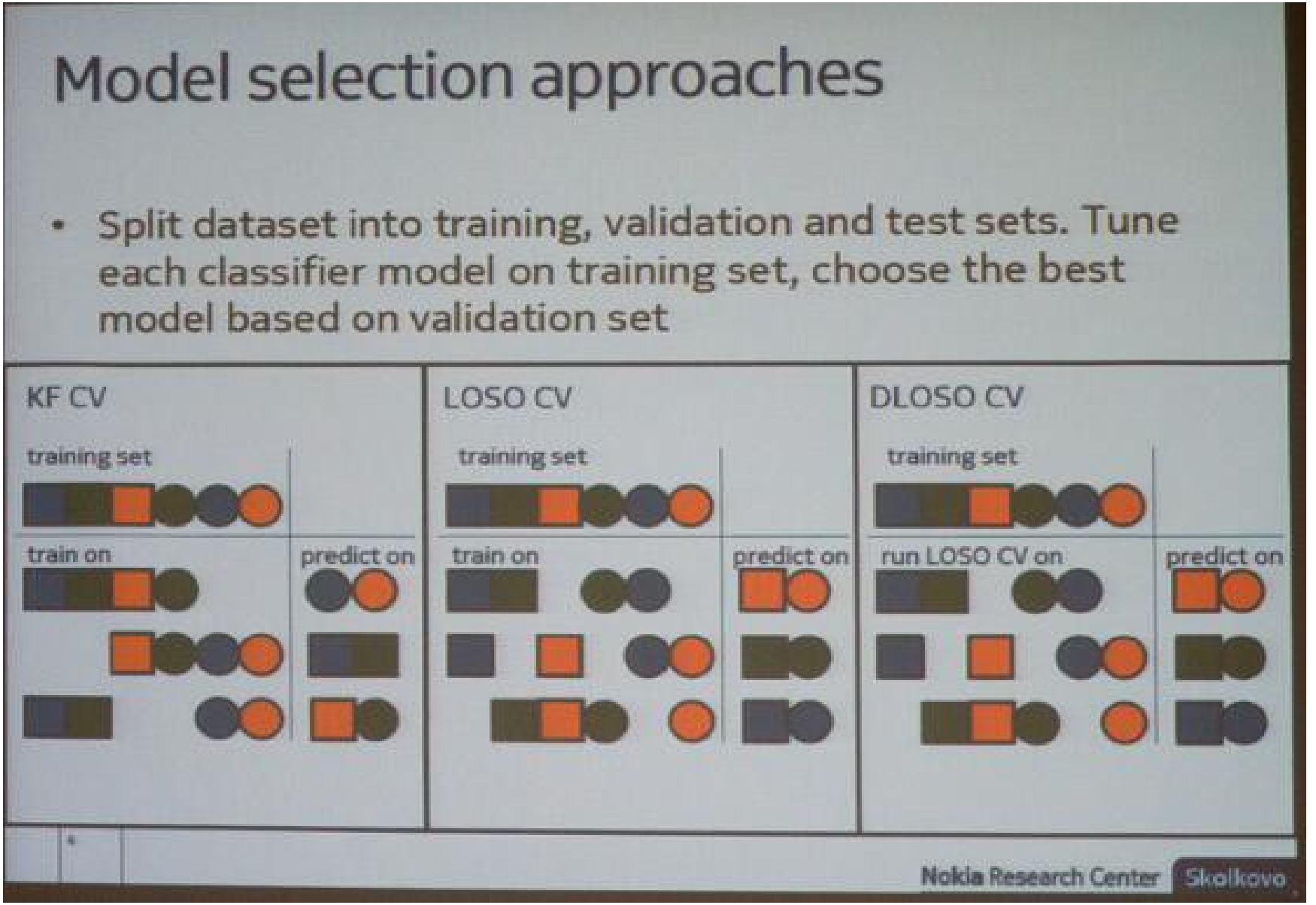
ИДЕЯ?

- Можно рассматривать любой регион метилирования, гистонных модификаций и т.д. как вектор из средних по вложенным областям в разных масштабах - после этого по cosine distance можно уже проводить кластеризацию, оценивать похожесть итд.
- Можно попробовать использовать convolution analysis для треков с разным уровнем детализации. Строить классификаторы разных уровней, соответствующие детализации.

Model assessment and selection for human activity recognition



Cross Validation Techniques



Statistics and Causation

Correlations have predictive value

- *"It is raining" ⇒ "People probably carry open umbrellas."*
- *"People carry open umbrellas" ⇒ "It is probably raining."*

Interventions

- Hypothetical: *"Will it rain if we ban umbrellas?"*
- Counterfactual: *"Would have it rained if we had banned umbrellas?"*

Causation

- Causal relations let us to reason on the outcome of interventions.

Recent advances in causal inference

- (Rubin, 1986) (Spirtes et al., 1993, 2011, ...)
- (Pearl, 2000, 2009, ...)

Structure -> Functions

- Structural Equation Model -> Модель с зависимыми функциями
- При переходе от функций к наблюдениям приходим к Bayesian Network с известной структурой. И уже пытаемся оценить Распределения
- Вопрос: Как быть с параметрами?
- Пример с Equilibrium Analysis из физики

Inference of Causal Direction

Amazon Better Together

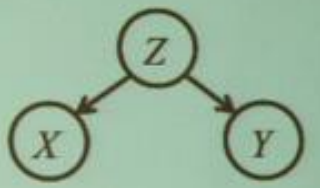
- Сейчас не смог повторить
- Предлагает купить ноутбук к рюкзаку для ноутбуков.
- Не учитывается причинно-следственная связь

Common Cause Principle

Statistical Implications of Causality

Reichenbach's
Common Cause Principle
links **causality** and **probability:**

(i) if X and Y are statistically dependent, then there is a Z causally influencing both;



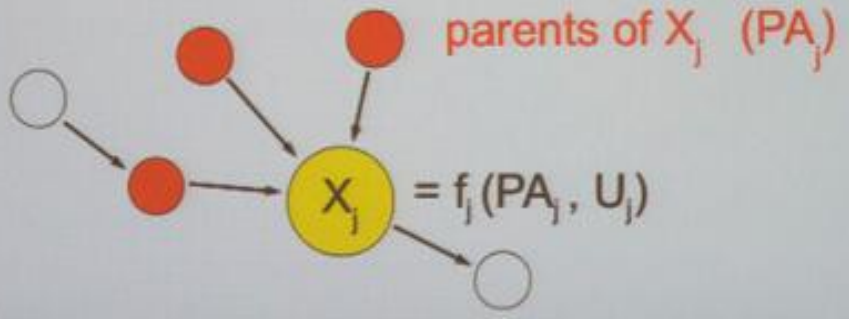
special cases:

(ii) Z screens X and Y from each other (given Z, the observables X and Y become independent)

Functional Causal Model

Functional Causal Model *(Pearl et al.)*

- Set of observables $X_1, \ldots, X_n$
- directed acyclic graph G with vertices $X_1, \ldots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, with independent $\text{Noise}_1, \ldots, \text{Noise}_n$.
- "Noise" means "unexplained" (or "exogenous"), we use U_i
- Can add requirement that $f_1, \ldots, f_n, \text{Noise}_1, \ldots, \text{Noise}_n$ "independent" (cf. *Lemeire & Dirkx 2006, Janzing & Schölkopf 2010* — more below)



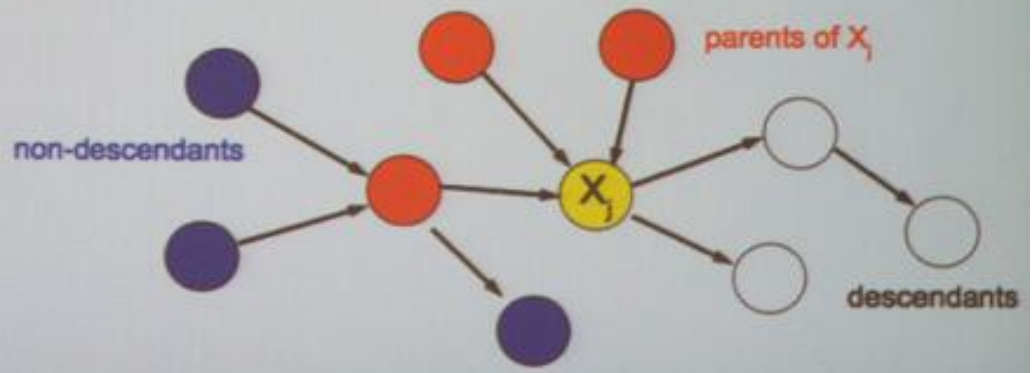
Theorem

Functional Model, ctd. *(Lauritzen, Pearl,...)*

Theorem: the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition: X_j *statistically independent* of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: d-separation (characterizes the set of independences implied by local Markov condition)
- Factorization $P(X_1, \ldots, X_n) = \prod_j P(X_j \mid \text{Parents}_j)$ (conditionals as causal mechanisms generating statistical dependence)

(subject to technical conditions)



Counterfactuals

Counterfactuals and Interventions

- David Hume (1711–76): *"... we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed."*
- Jerzy Neyman (1923): consider m plots of land and ν varieties of crop.

  Denote U_{ij} the crop yield that *would be* observed if variety $i = 1, \ldots, \nu$ were planted in plot $j = 1, \ldots, m$

  For each plot j, we can only experimentally determine *one* U_{ij} in each growing season.

  The others are called "counterfactuals".
- this leads to the view of causal inference as a missing data problem — the "potential outcomes" framework (Rubin, 1974)
- in $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, the equality sign is interpreted as an assignment ":=" — interventions can only take place on the right hand side

Causal Inference Method

Causal Inference Method

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

- Compute a function f as non-linear regression of X on Y
- Compute the residual

$$E := Y - f(X)$$

- check whether E and X are statistically independent (uncorrelated is not enough)

Independence

Independence-based Regression *(Mooij et al., 2009)*

- Problem: many regression methods assume a particular noise distribution; if this is incorrect, the residuals may become dependent
- Solution: minimize dependence of residuals rather than maximizing likelihood of data in regression objective
- Use RKHS distance between kernel mean embeddings/Hilbert-Schmidt-norm of cross-covariance operator between two RKHSes as a dependence measure

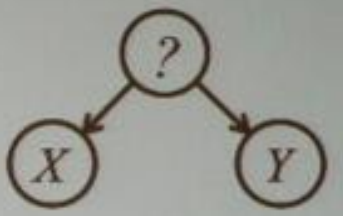
Mooij, Janzing, Peters, Schölkopf: Regression by dependence minimization and its application to causal inference. *ICML* 2009.
Yamada & Sugiyama: Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise. *AAAI* 2010.

Confounders

Detection of Confounders

Given $p(X, Y)$, infer whether

- $X \to Y$
- $Y \to X$
- $X \leftarrow T \to Y$ for some (possibly) unobserved variable T



- Confounded additive noise (CAN) models

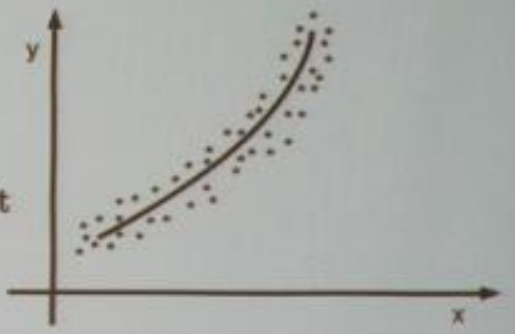
$$X = f_X(T) + U_X$$
$$Y = f_Y(T) + U_Y$$

with functions f_X, f_Y and U_X, U_Y, T jointly independent

Note: includes the case

$$Y = f(X) + U$$

by setting $f_X = id$ and $U_X = 0$.



- Estimate $(f_X(T), f_Y(T))$ using dimensionality reduction
- If U_X or U_Y is close to zero, output 'no confounder'
- Identifiability result for small noise

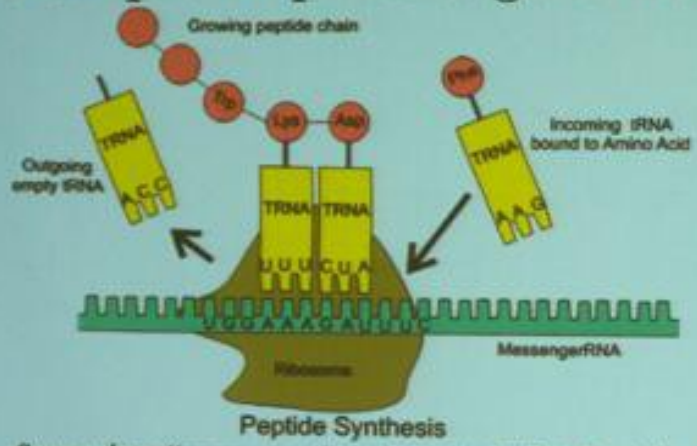
Janzing, Peters, Mooij, Schölkopf: Identifying latent confounders using additive noise models. *UAI 2009*

Causal learning

Causal Learning and Anticausal Learning

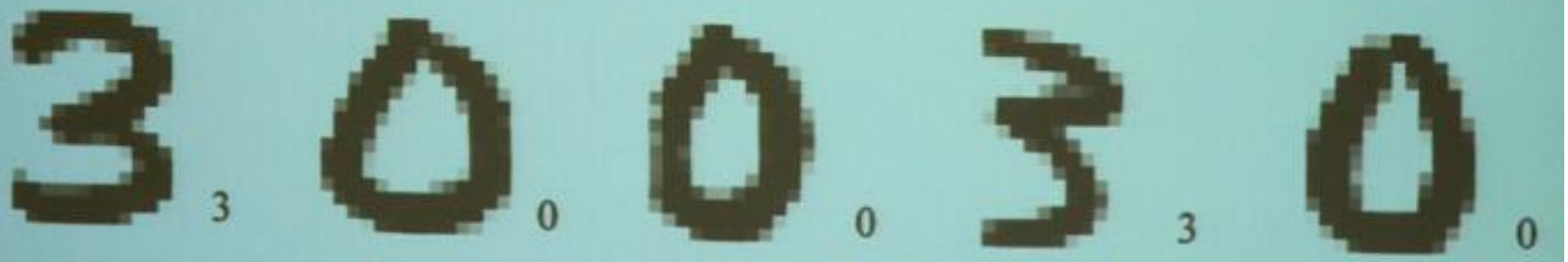
Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, *ICML* 2012

- example 1: predict gene from mRNA sequence



Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png

- example 2: predict class membership from handwritten digit



Mergeable Summaries

Approximation Motivation

- Why use approximate when data storage is cheap?
  - Parallelize computation: partition and summarize data
    - Consider holistic aggregates, e.g. median finding
  - Faster computation (only work with summaries, not full data)
    - Less marshalling, load balancing needed
  - Implicit in some tools
    - E.g. Google Sawzall for data analysis requires mergability
  - Allows computation on data sets too big for memory/disk
    - When your data is "too big to file"

Summaries

- Hash Kernels
- Sort + Quantiles
- Online summaries. Пересчет при поступлении нового измерения
- Random Summary. E(error) может равняться 0 в отличие от determined.

Yandex Data Factory

Classical Machine Learning

Pros

- It works

Cons

- Specific Data Scientists
- Desynchronized formulas/experiment
- Hand made formula quality assurance outside of the formula development process
- Potential over-fitting formula
- Machine learning pipeline

Yandex Data Factory

Machine Learning Pipeline

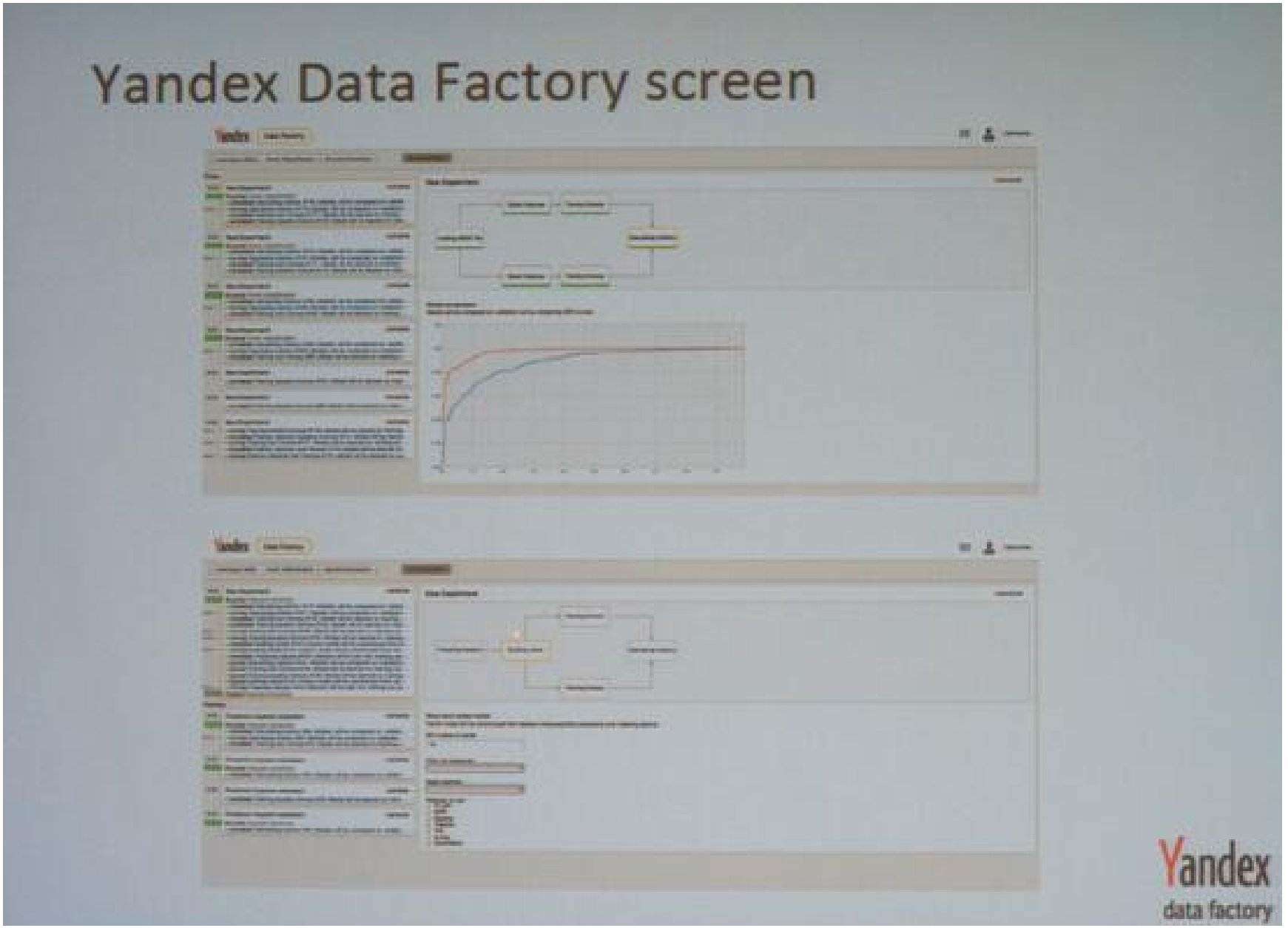
Pros

- Less time consuming
- Results of previous learning cycles are preemptive
- Builtin looks for automating QA
- Features quality monitoring
- Complete workflow automation
- Enables effective teamwork
- Specific Data scientists are not "must have"

Cons

- Supports only framework-implemented algorythms

Yandex Data Factory



Explaining AdaBoost

- Boosting Weak Classifiers -> Combination
- AdaBoost – Adaptive Boosting Algorithm
- Weak classifier should be slightly better random coin tossing!

References:

- Robert Schapiro and Yoav Freund. Boosting: Foundations and Algorithms. MIT 2012.
- Robert Schapiro. Explaining AdaBoost. to appear.

Algorithm

AdaBoost

[with Freund]

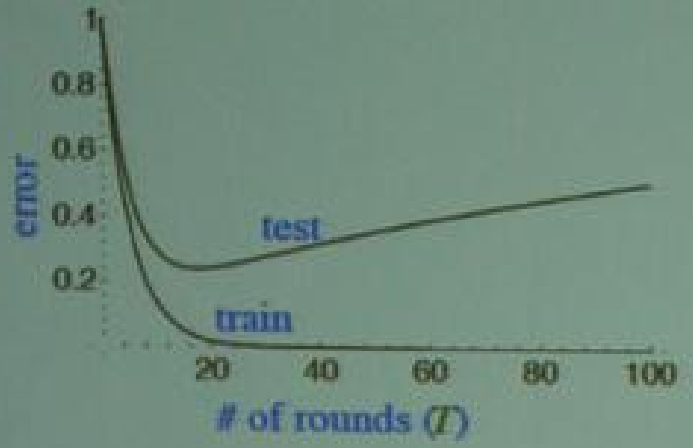
- given m training examples (x_i, y_i) where $y_i \in \{-1, +1\}$
- initialize $D_1 =$ uniform distribution on training examples
- for $t = 1, \ldots, T$:
  - train weak classifier h_t on D_t
  - choose $\alpha_t =$ [some formula] > 0
  - compute new distribution D_{t+1}:
    - for each example i:

      multiply $D_t(i)$ by $\begin{cases} e^{-\alpha_t} & (< 1) \quad \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & (> 1) \quad \text{if } y_i \neq h_t(x_i) \end{cases}$
    - renormalize

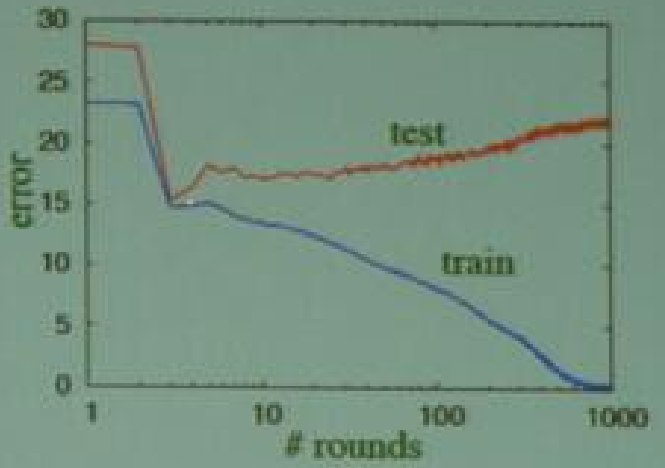
VC theory

Predicted Behavior

predicted



actual example



- complexity increases with # rounds ⇒ overfitting
- can happen

Margins approach

Margins

#(+1) - #(-1). Computed on Training Set

Theorem

Larger Margins on test -> Better bounds on Test.

Theorem

AdaBoost -> Increase Margins.

Loss function approach

Theorem

AdaBoost -> Minimization Exponential Loss.

Но на самом деле минизации Loss Function не хватает. Дает очень плохие оценки на регуляризацию!

Есть пример, который ломает все существующие алгоритмы, построенные на минимизации Loss Function.

Experiment

- Data: x - равномерно из {0,1}^10000
- y - majority vote of 3 coords
- weak classifier - single coordinate or negation
- training set = 1000

Experiment

An Experiment

- data:
  - instances x uniform from $\{-1, +1\}^{10,000}$
  - label y = majority vote of three coordinates
  - weak classifier = single coordinate (or its negation)
  - training set size $m = 1000$
- algorithms (all provably minimize exponential loss):
  - standard AdaBoost
  - gradient descent on exponential loss
  - AdaBoost, but in which weak classifiers chosen at random
- results:

exp. loss	% test error [# rounds]		
	stand. AdaB.	grad. desc.	random AdaB.
10^{-10}	0.0 [94]	40.7 [5]	44.0 [24,464]
10^{-20}	0.0 [190]	40.8 [9]	41.6 [47,534]
10^{-40}	0.0 [382]	40.8 [21]	40.9 [94,479]
10^{-100}	0.0 [956]	40.8 [70]	40.3 [234,654]

AdaBoost

AdaBoost avoids over-fitting, using little alpha values (implicit regularization) and if it stops.

Theorem

AdaBoost -> Bayes Optimal Fit if

- Enough data
- Run for many, but not too many rounds
- Weak Classifiers space is "sufficiently rich" (classifier set is enough to minimise loss function). (In turn weak classifiers give better margins)

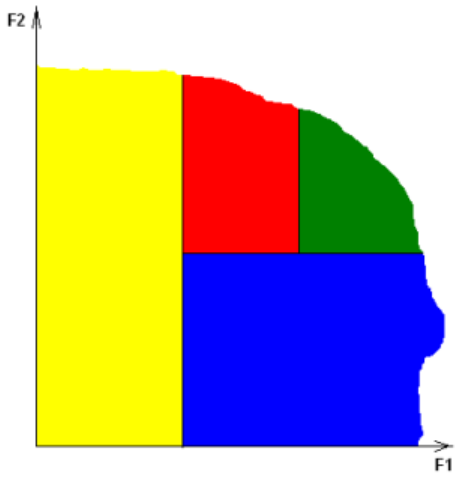
Noise & AdaBoost

Handling Noise or Outliers

- on "real-world" datasets, AdaBoost often works anyway
- various theoretical algorithms based on "branching programs" (e.g., [Kalai & Servedio], [Long & Servedio])
- different approach based on boost-by-majority [Freund]
  - exactly optimally efficient boosting algorithm
  - does not appear to minimize any (convex) loss function
  - like AdaBoost, puts more weight on hard examples
  - unlike AdaBoost, "gives up" on very hard examples
    - may make more robust
  - not adaptive, but can be made so by moving to continuous time $\rightarrow$ "BrownBoost"

ИДЕЯ?

- Можно попробовать посчитать margins для наших классификаторов
- Можно попробовать отображать результат работы AdaBoost на плоскости, как дерево решений, но изменяя площадь, согласно значению соответсвующего alpha коэффициента



Bioinformatics & Linguistics

3. Conclusion.

3.3. Summary. From bioinformatics to linguistics

- Gap penalties, substitution scores, form of gap penalty function, etc – *pay attention on this issue; note that local transformations, e.g. symbol transpositions can be allowed with the same time complexity*
- Tagging of symbols in the text based on knowledge of the text's nature – *can help mutatis mutandis*
- Tree alignment and simultaneous alignment of texts and trees adjusted to the texts – *IMHO common situation. See work of Galitsky et al.*

Linguistics

- Linguistics and topic models. Intrudor word - word inserted in topic to estimate how consistent is the topic. Topic is good **most** of tests showed this word.
- Words distribution ~ Latent Dirichlet allocation

  Approaches
  - Old: Topic -> Word Distribution.
  - New: Topic -> Concept -> Words. Improves model by Human Feedback tutor learning.

- Cognition map & sentiments - System built for Sony corporation to analyse customers reviews (Nuance Communications)
  - wikipedia.cognition.com
  - sentinel.cognition.com

Regularization etc.

- Regularization techniques:
  - Ridge regression
  - Logistic regression
  - Logistic Ridge regression
  - Lasso L1 regularization

    Вывод: Lasso Logistic regression сравним с SVM с линейным ядром.

- **Вопрос** Как искать параметр alpha?
  Ответ Итеративный алгоритм с использованием предыдущего решения как начального приближения.
- **Вопрос** Как оценивать параметр регуляризации?
  Ответ
  - BIC - ассимптотическую оценку.
  - CV - дает более точную оценку, но есть и проблемы.

Issues with cross validation. Решение - Adaptive search, но он не работает для нескольких параметров - открытый вопрос. ВАЖНО: теоретических условий на adaptive search нет!

Statistics Tutorials

- http://www.machinelearning.ru/wiki/index.php?title=%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29
- http://www.autonlab.org/tutorials/