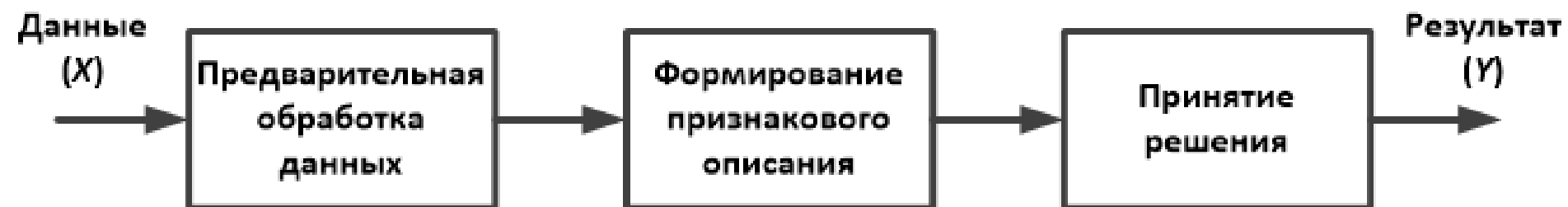


Машинное обучение

Структура системы распознавания



Основные понятия. Задача обучения по прецедентам

1. Множество объектов X
2. Множество допустимых ответов Y
3. Целевая функция $y^* : X \rightarrow Y$ – неизвестная зависимость
4. Пара «объект–ответ» (x_i, y_i) – прецедент
5. Обучающая выборка
6. Найти алгоритм $a: X \rightarrow Y$, который приближает y^* на всём множестве X

Объекты и признаки

$f_j : X \rightarrow D_j, j = 1, \dots, n$ – признаки объектов (features). Типы признаков:

$D_j = \{0,1\}$ – бинарный признак;

$|D_j| < \infty$ – номинальный признак;

$|D_j| < \infty, D_j$ – упорядочено – порядковый признак;

$D_j = \mathbb{R}$ – количественный признак.

Вектор $(f_1(x), \dots, f_n(x))$ – признаковое описание объекта x .

Ответы и типы задач

Задачи классификации (classification)

- $Y = \{-1, +1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – на M непересекающихся классов.
- $Y = \{0, 1\}^M$ – на M классов, которые могут пересекаться

Задачи восстановления регрессии (regression)

- $Y = R$ или $Y = R^m$.

Задачи ранжирования

- Y – конечное упорядоченное множество.

Предсказательная модель

Модель алгоритмов – параметрическое семейство функций:

$$A = \{g\{x, \vartheta\}, \vartheta \in \Theta\},$$

g – некоторая фиксированная функция,

Θ – множество допустимых значений параметра ϑ .

Метод обучения

Метод обучения – ставит в соответствие произвольной конечной выборке $X^l = (x_i, y_i)$, $(i = 1, l)$ некоторый алгоритм $a \in A$.

В задачах обучения по прецедентам есть два этапа:

- этап обучения;
- этап применения.

Обучение с учителем и без учителя

Supervised vs Unsupervised

There is a bunch of different fruits



Supervised

*Based on its
color/shape/weight...*

➤ *Is that "fruit" an apple?*

Unsupervised

➤ *How the different fruits
can be classified inside
your grocery store?*

Функционал качества

Функция потерь – это неотрицательная функция $L(a, x)$, характеризующая величину ошибки алгоритма a на объекте x .

Функционал качества алгоритма a на выборке X^l :

$$Q(a, X^l) = (1/l) \sum L(a, x_i).$$

$$L(a, x) = [a(x) \neq y^*(x)] \text{ – классификация}$$

$$L(a, x) = [a(x) - y^*(x)] \text{ – средняя ошибка}$$

$$L(a, x) = [a(x) - y^*(x)]^2 \text{ – квадратичная функция потерь}$$

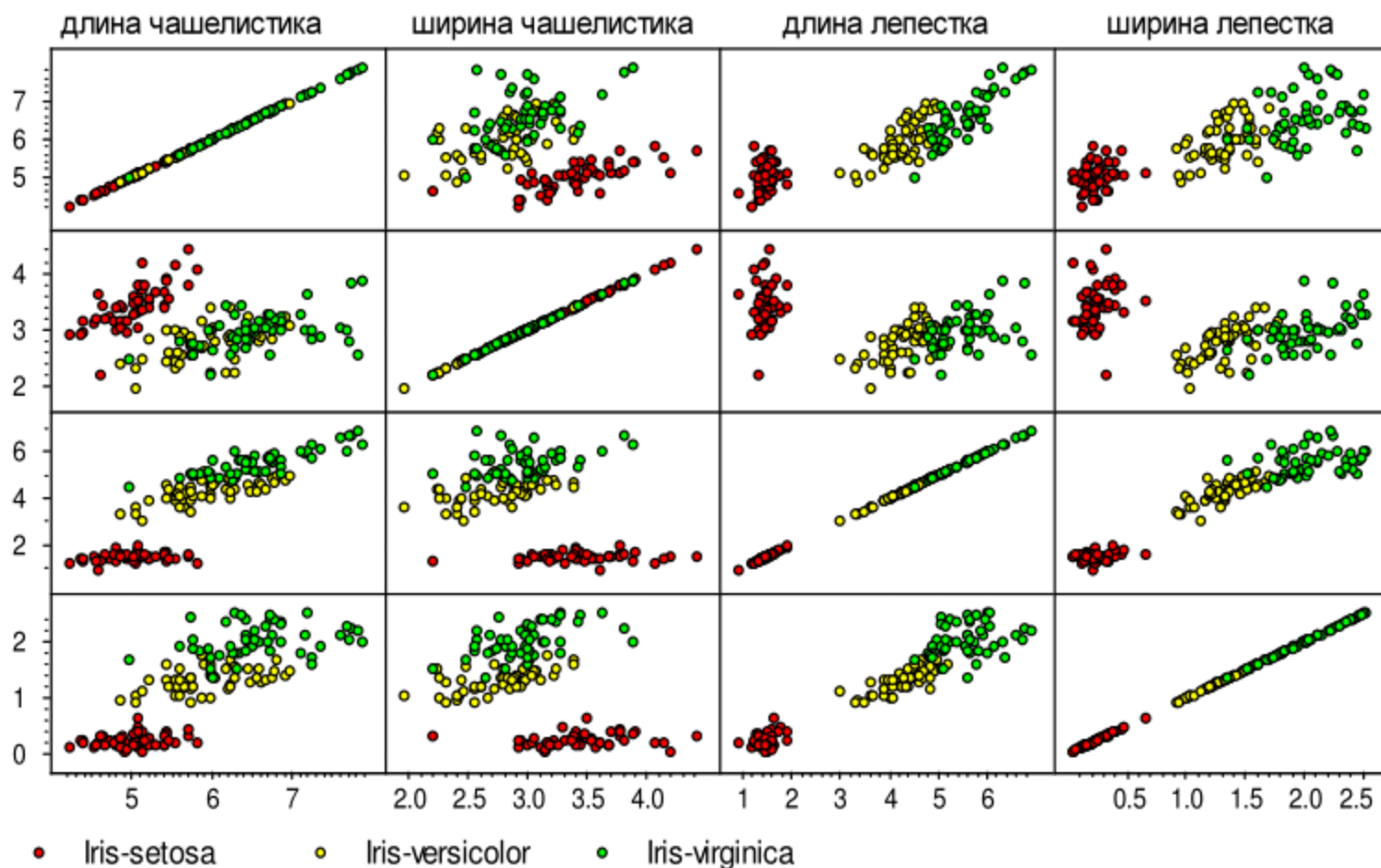
$\arg \min_{a \in A} Q(a, X^l)$ – минимизация эмпирического риска

Выборка

- Обучающая
- Тестовая
- Контрольная

Классификация цветков ириса [Фишер, 1936]

$n = 4$ признака, $|Y| = 3$ класса, длина выборки $\ell = 150$



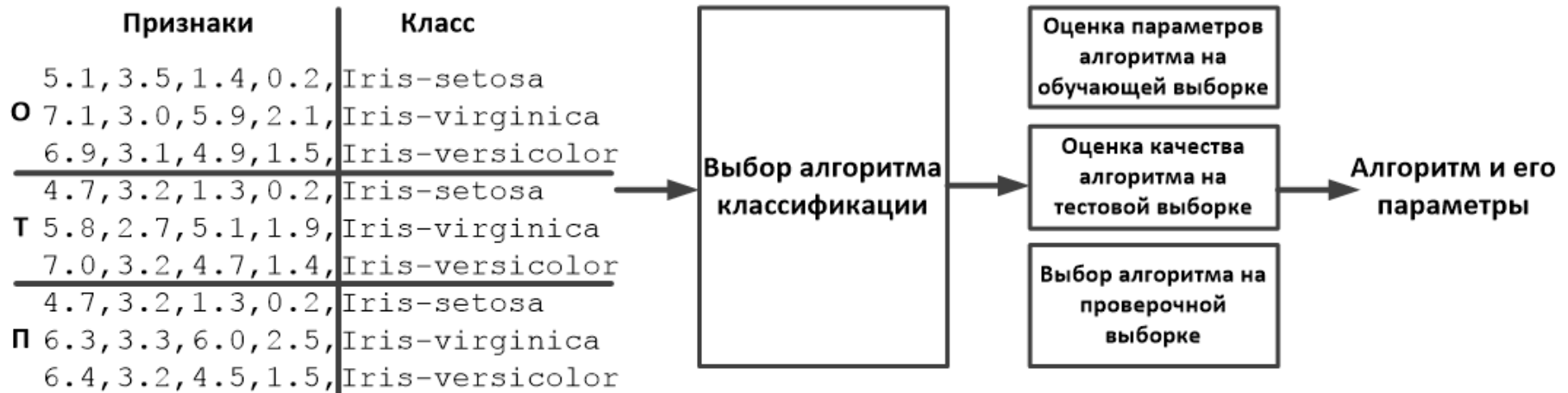
Данные

Ирисы Фишера

Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	Вид ириса
5.1	3.5	1.4	0.2	<i>setosa</i>
4.9	3.0	1.4	0.2	<i>setosa</i>
4.7	3.2	1.3	0.2	<i>setosa</i>
4.6	3.1	1.5	0.2	<i>setosa</i>
5.0	3.6	1.4	0.2	<i>setosa</i>
5.4	3.9	1.7	0.4	<i>setosa</i>
4.6	3.4	1.4	0.3	<i>setosa</i>
5.0	3.4	1.5	0.2	<i>setosa</i>
4.4	2.9	1.4	0.2	<i>setosa</i>
4.9	3.1	1.5	0.1	<i>setosa</i>
5.4	3.7	1.5	0.2	<i>setosa</i>
4.8	3.4	1.6	0.2	<i>setosa</i>
4.8	3.0	1.4	0.1	<i>setosa</i>
4.3	3.0	1.1	0.1	<i>setosa</i>
5.8	4.0	1.2	0.2	<i>setosa</i>
5.7	4.4	1.5	0.4	<i>setosa</i>

Работа алгоритма классификации

Обучение



Применение



Цветки ириса



Iris setosa



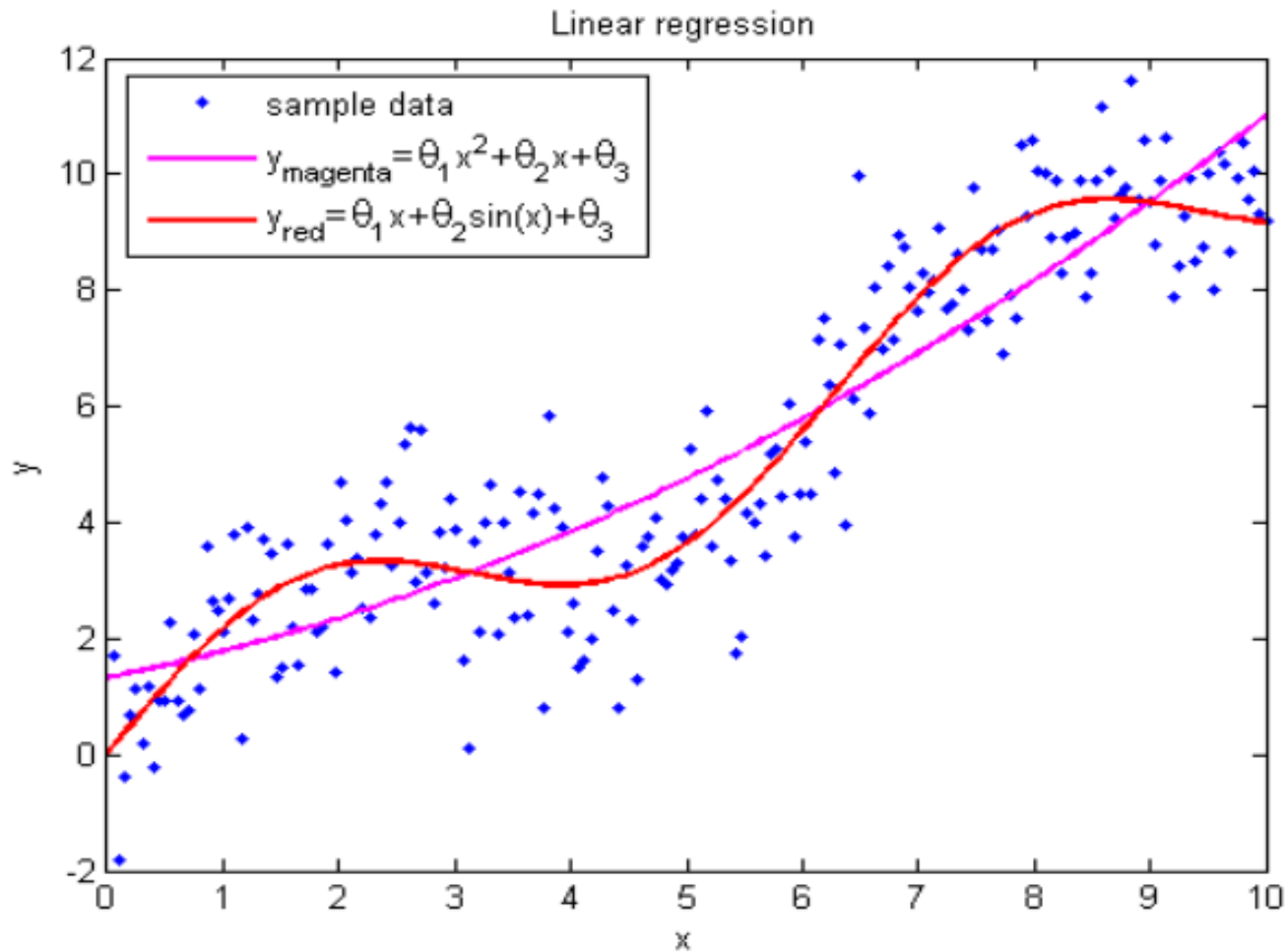
Iris virginica



Iris versicolor

Задача регрессии

$X = Y = \mathbb{R}$, $I = 200$



Проблема переобучения

Аппроксимация вещественной функции:

$$y(x) = 1 / (25 + x^2)$$

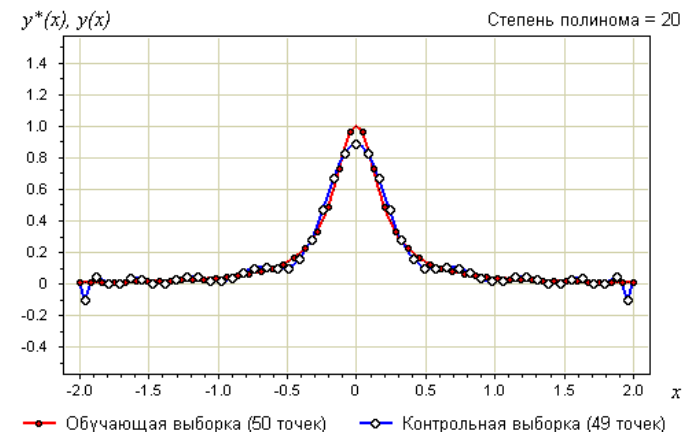
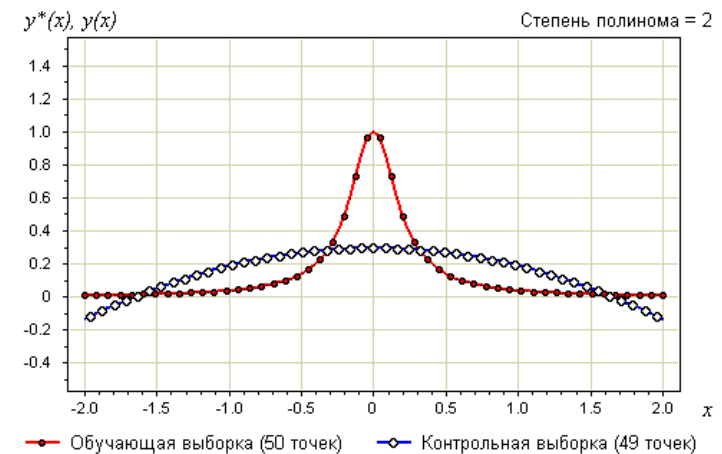
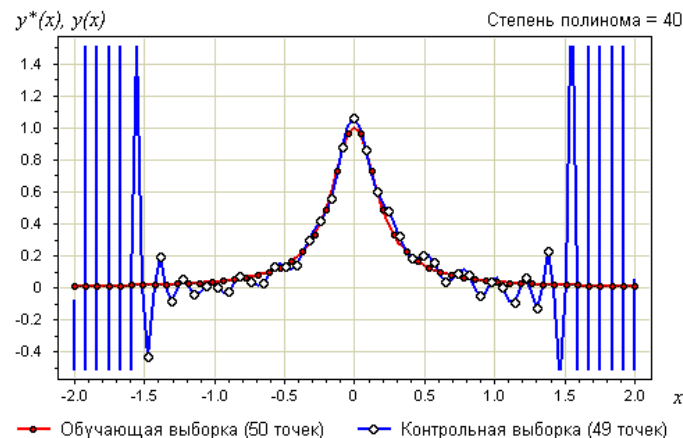
Модель для аппроксимации – полином:

$$a(x, w) = w_0 + w_1x + \dots + w_px^p$$

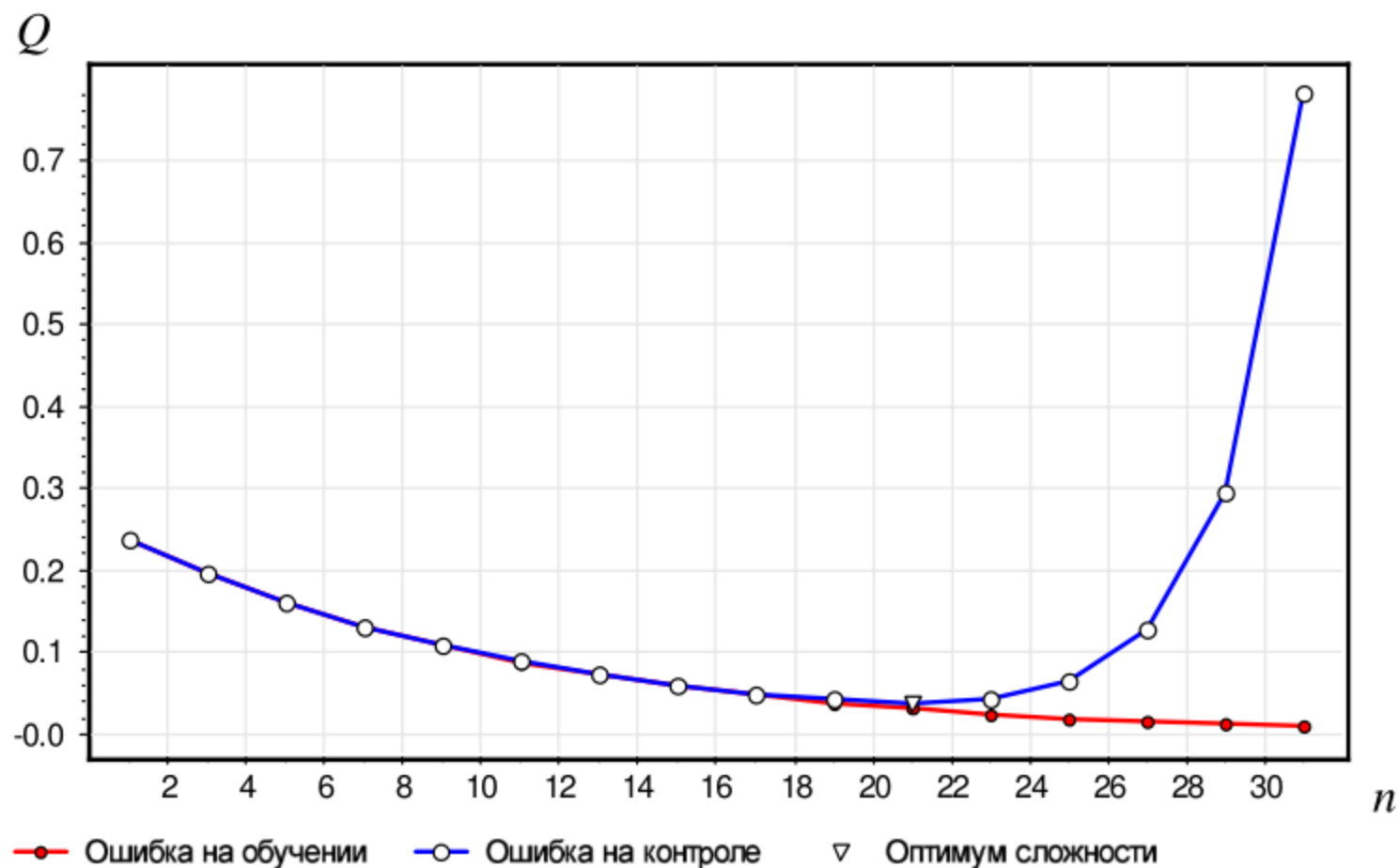
Обучающая выборка: $x_i = (4 * (i-1) / (m-1)) - 2$

Тестовая выборка: $x_i = (4 * (i-0.5) / (m-1)) - 2$

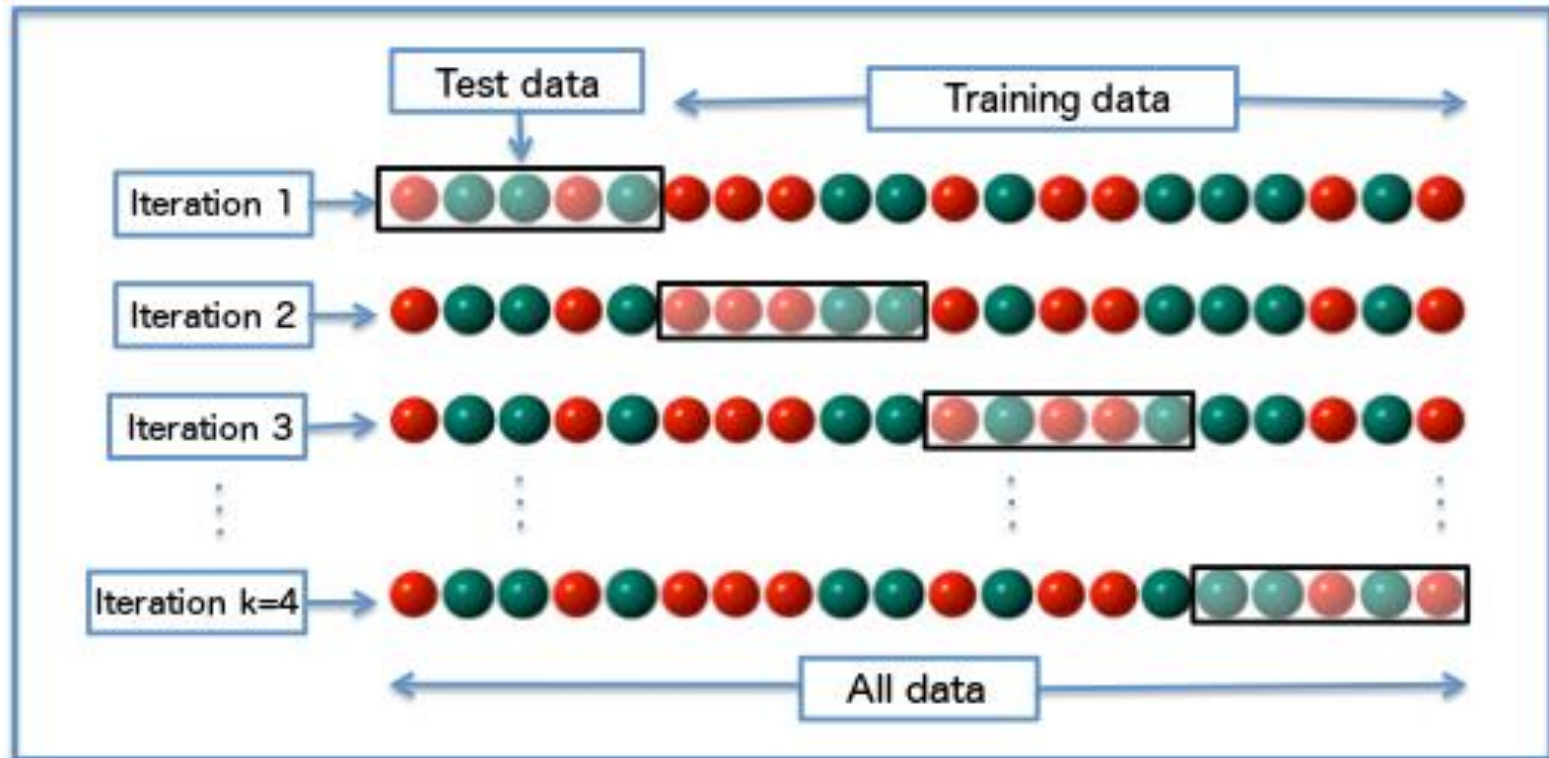
Узлы тестовой выборки находится между узлами обучающей выборки.



Выбор степени полинома



Перекрёстная проверка



Задачи машинного обучения

- Классификация
- Регрессия
- Кластеризация
- Ранжирование
- Поиск ассоциативных правил

Языки программирования

- R
- Python
- Библиотеки для R, Python

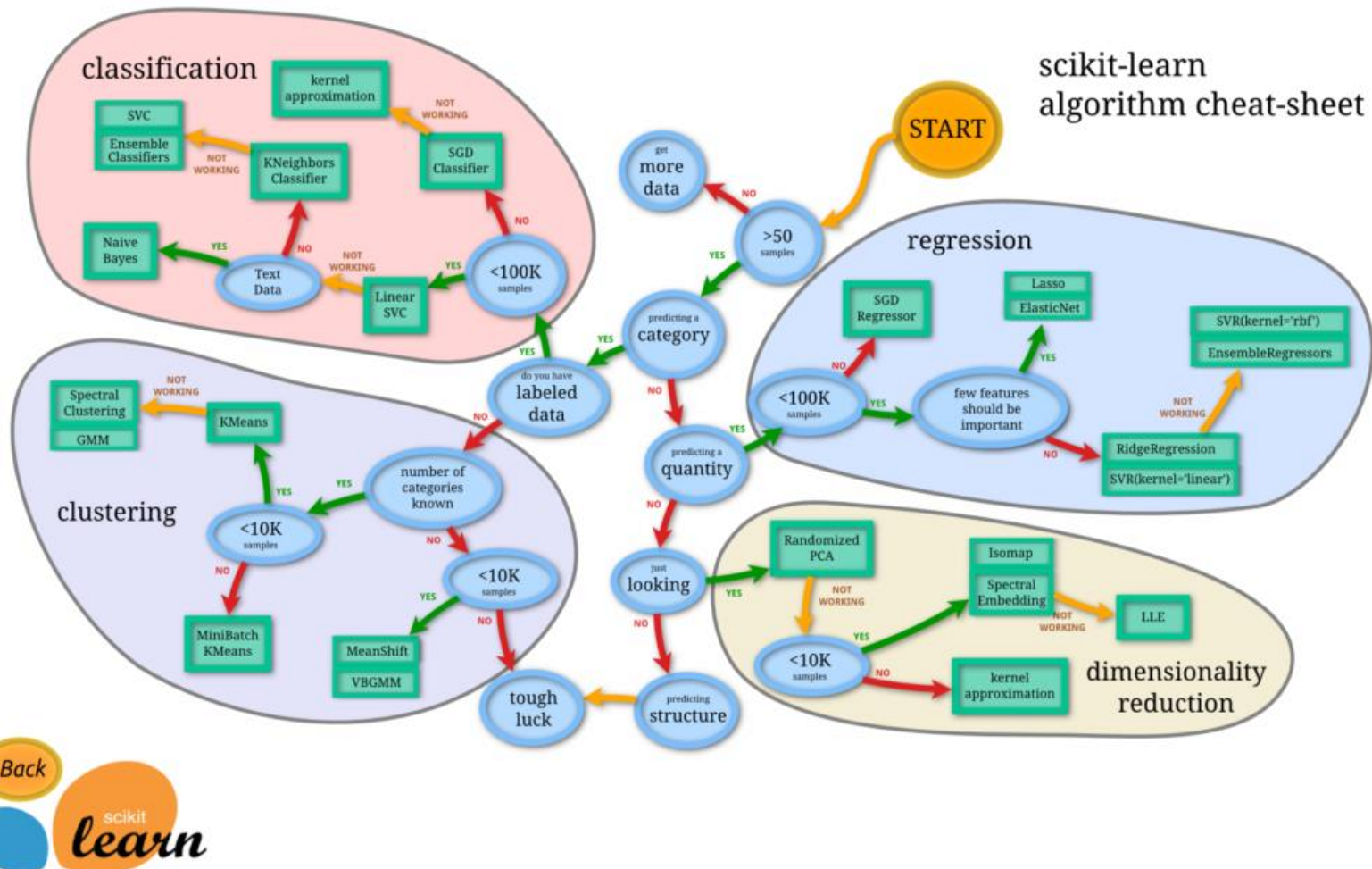
Библиотеки машинного обучения

1. Tensorflow (Google)
2. CNTK (Microsoft)
3. Caffe (Berkeley Vision and Learning Center)
4. Keras (Франсуа Шолле)
5. Theano (Монреальский университет)
6. OpenCV (Intel, ITSeez)
7. CatBoost (Яндекс)

Актуальность (показать видео)

1. Распознавание дорожных знаков
2. Детектор дорожных полос
3. GTA, Mario, Flappy Bird
4. FindFace
5. Оценка эмоций человека
6. Кредитный скоринг
7. Системы поиска



Список алгоритмов




Список алгоритмов




Kaggle



[Competitions](#) [Datasets](#) [Kernels](#) [Discussion](#) [Jobs](#) [...](#)


Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems




New to Data Science?

Get started with a tutorial on our most popular competition for beginners, [Titanic: Machine Learning from Disaster](#).



Build a Model

Get the data & use whatever tools or methods you prefer to make predictions.



Make a Submission

Upload your prediction file for real-time scoring & a spot on the leaderboard.

[Learn more](#)

[Dismiss](#)

Active

All

Entered

Sort by Prize

15 active competitions

All Categories

Search competitions



Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Featured · 3 months to go

\$1,500,000

238 teams



Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Featured · 4 months to go

\$1,200,000

2,619 teams



Carvana Image Masking Challenge

Automatically identify the boundaries of the car in an image

Featured · 23 days to go

\$25,000

537 teams



Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

Research · 8 days to go

\$25,000

1,007 teams




Personalized Medicine: Redefining Cancer Treatment

Predict the effect of Genetic Variants to enable Personalized Medicine

Research · a month to go

\$15,000

1,050 teams

 Featured Prediction Competition

Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

\$1,500,000

Prize Money



Department of Homeland Security · 238 teams · 3 months to go (3 months to go until merger deadline)

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Join Competition](#)

Overview

Description

Evaluation

Prizes

Timeline

While long lines and frantically shuffling luggage into plastic bins isn't a fun experience, airport security is a critical and necessary requirement for safe travel.


No one understands the need for both thorough security screenings and short wait times more than U.S. Transportation Security Administration (TSA). They're responsible for all U.S. airport security, screening more than two million passengers daily.

As part of their Apex Screening at Speed Program, DHS has identified high false alarm rates as creating significant bottlenecks at the airport checkpoints. Whenever TSA's sensors and algorithms predict a potential threat, TSA staff needs to engage in a secondary, manual screening process that slows everything down. And as the number of travelers increase every year and new threats develop, their prediction algorithms need to continually improve to meet the increased demand.

Currently, TSA purchases updated algorithms exclusively from the manufacturers of the scanning equipment used. These algorithms are proprietary, expensive, and often released in long cycles. In this competition, TSA is stepping

Additional Files


 stage1_labels.csv

 stage1_sample_submis...


 body_zones.png

 sample.tar.gz

 stage1_a3daps.tar.gz

 stage1_aps.tar.gz

stage1_aps.tar.gz 10.21 GB

 Download

Data Introduction

This dataset contains a large number of body scans acquired by a new generation of millimeter wave scanner called the High Definition-Advanced Imaging Technology (HD-AIT) system. The competition task is to predict the probability that a given body zone (out of 17 total body zones) has a threat present.

The images in the dataset are designed to capture real scanning conditions. They are comprised of volunteers wearing different clothing types (from light summer clothes to heavy winter clothes), different body mass indices, different genders, different numbers of threats, and different types of threats. Due to restrictions on revealing the types of threats for which the TSA screens, the threats in the competition images are "inert" objects with varying material properties. These materials were carefully chosen to simulate real threats.

The volunteers used in the first and second stage of the competition will be different (i.e. your algorithm should generalize to unseen people). In addition, you should not make assumptions about the number, distribution, or location of threats in the second stage.















This is a two-stage competition. Scores on the leaderboard below may be the result of leaderboard probing and not indicative of final competition performance.

In the money


Gold

Silver

Bronze

#	Δ1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	▲2	5 a day			0.00865	137	2h
2	—	Kevin H			0.01658	14	44m
3	▼2	teedrz			0.01734	3	18d
4	▲2	idle_speculation			0.02956	5	2h
5	▼1	dhammack			0.03422	3	15d
6	▼1	I don't know what I'm doing			0.03535	25	5d
7	—	bmci			0.03781	26	1mo
8	—	SuarezAteMyTacos			0.03956	24	1d
9	▲4	waves		 	0.05161	46	16h
10	▼1	Yusaku Sako			0.05342	4	9d
11	▼1	Joseph Chui			0.06206	59	19h
12	▼1	AsymptoticallyFree			0.09695	12	5d
13	▼1	Roland Luethy			0.09737	25 29	8d

Stack Overflow

 Questions Developer Jobs Documentation BETA Tags Users

Top Questions

interesting 321 featured hot week month

Ask Question

0 votes

0 answers

2 views

jQuery .append() <tr> and <td> add's <td> outside of <tr>

javascript jquery html css

asked 45 secs ago Hunter 46

0 votes

1 answer

6 views

Show ng-model length using controllerAs

angularjs

answered 1 min ago Ajinkya Dhote 1

2 votes

1 answer

10 views

Blocking multiprocessing pool.map called in a process

python linux parallel-processing multiprocessing

modified 2 mins ago noxdafox 3,667

1 vote

2 answers

50 views

Parse inequality (character) expressions to numeric ranges

r

modified 3 mins ago d.b 11.9k

1 vote

1 answer

5 views

Using Canny edge to create a mask

python opencv image-processing

answered 3 mins ago Martin Beckett 74.2k

0 votes

1 answer

38 views

Multiple issues with ggplot2 (discrete X_axis, error bar not correctly aligned horizontally)

r ggplot2

modified 4 mins ago vestland 1,011

0 votes

0 answers

4 views

Binning data into non-grid based bins

python numpy vectorization voronoi

asked 4 mins ago J.Warren 55


0 votes


1 answer

0 views

How to save an image to SQL server with java? when the image is in

FEATURED ON META

 [Sunsetting Documentation](#)

 [Documentation is read-only. What's next?](#)

HOT META POSTS

16 [Is that plagiarism by Community♦?](#)

15 [Duplicates dichotomy: \[excel-vba\] vs \[excel\] + \[vba\]](#)

Favorite Tags [edit](#)

r


matlab


image-processing

opencv

Want a python job?

Data Science Consultant (m/w)


Comma Soft AG  Bonn, Deutschland


€50K - €90K  REMOTE

r

python

Senior Back-end Developer (Remote / Well versed in scaling)

Hotjar  No office location

€75K - €95K  REMOTE

Курсы по машинному обучению

coursera

Каталог

Поиск в каталоге



Поделиться



Машинное обучение и анализ данных

Типовые задачи машинного обучения и анализа данных и методы их решения

О специализации

Курсы

Авторы

Часто задаваемые вопросы

Специализация
Машинное обучение и анализ данных

Зарегистрироваться

Начался Sep 04

Об этой специализации

Мы покажем, как проходит полный цикл анализа, от сбора данных до выбора оптимального решения и оценки его качества. Вы научитесь пользоваться современными аналитическими инструментами и адаптировать их под особенности конкретных задач.

В рамках специализации вы освоите основные темы, необходимые в работе с большим массивом данных, в т.ч. современные методы классификации и регрессии, поиск структуры в данных, проведение экспериментов, построение выводов, базовая фундаментальная математика, основы программирования на Python.

Мы разберём, как построить рекомендательную систему, оценить эмоциональную окраску текста,

Математика и Python для
анализа данных
Московский физико-технический
институт, Yandex

Главная страница курса

Неделя 1 📅

Неделя 2

Неделя 3

Неделя 4

Отметки

Форумы обсуждений

Ресурсы

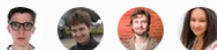
Информация о курсе

Понравился этот курс? Станьте специалистом, присоединившись к [специализация Машинное обучение и анализ данных](#).

Оплатить

Математика и Python для анализа данных

от Московский физико-технический институт & Yandex



Добро пожаловать в удивительный мир машинного обучения и анализа данных!

👇 Еще



Эта сессия завершилась 12 сентября 2016 г.

Не расстраивайтесь! Вы можете зарегистрироваться на следующую сессию. Прогресс будет сохранен, и вы

Switch Sessions

ШКОЛЕ АНАЛИЗА ДАННЫХ ИСПОЛНИЛОСЬ 10 ЛЕТ!



Школа анализа данных

Высокие технологии и современная наука.
Будет сложно, вам понравится.

Вход для участников

Есть вопросы?

Выпуски

Контакты



[О проекте](#)

[Как поступить](#)

[Программа](#)

[Расписание](#)

[Видео](#)

[Мероприятия](#)

[Основная программа](#) → [Первый семестр](#)

Введение в анализ данных

Цель курса — познакомить слушателей с сферой анализа данных, основными инструментами, задачами и методами, с которыми сталкивается исследователь данных в работе

Длительность

15 занятий

60 ак. часов

Курс преподают




Евгений Завьялов

Смешанное
занятие №1

[Введение в python. ipython notebook.](#)

4 часа
+ 2 часа СР

MachineLearning.Ru



Распознавание

статья | обсуждение | просмотр | история

MachineLearning.ru

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

Сейчас ресурс содержит 963 статьи на русском языке.

Poligon.MachineLearning.ru — Полигон алгоритмов классификации

Классификация

Регрессионный анализ

Прогнозирование

Прикладная статистика

Обработка сигналов

Распознавание образов

Анализ и понимание изображений

Обработка и анализ текстов

Прикладные системы анализа данных

Все направления

Концепция

Инструктаж

Все статьи

Ненаписанные статьи

Полезные ссылки

Частые вопросы

Справка

Конференция «Математические методы распознавания образов» (ММРО-2017)

Федеральный исследовательский центр «Информатика и управление» РАН, Московский физико-технический институт и Южный федеральный университет объявляют о проведении с 9 по 13 октября 2016 года 18-й Всероссийской конференции с международным участием «Математические методы распознавания образов» (ММРО-2017).

Место проведения конференции – г. Таганрог. Рабочие языки конференции — русский, английский.

Конференция ММРО является ведущим форумом исследователей и профессионалов, работающих в области интеллектуального анализа данных, площадкой для обсуждения, распространения и продвижения передовых идей, достижений и разработок. Конференция призвана способствовать обмену идеями между представителями науки и индустрии. Конференция организована представителями российской научной школы машинного обучения и нацелена на расширение взаимодействия между российскими и зарубежными исследователями и представителями высокотехнологичного IT бизнеса. Конференция проводится при поддержке Российского фонда фундаментальных исследований, компании Форексис, Центра систем распознавания и прогнозирования.

Тематика конференции: интеллектуальный анализ данных; машинное обучение; анализ больших данных; глубокое обучение; компьютерное зрение; анализ текстов и социальных сетей; промышленные приложения науки о данных.

Подать доклад можно на [странице участника конференции](#).

Цели Ресурса

- Сконцентрировать информацию о достижениях ведущих российских научных школ в области машинного обучения, распознавания образов, анализа данных.
- Способствовать обмену опытом, накоплению и распространению научных знаний в этой области.
- Предоставить площадку для виртуальных научных семинаров и обсуждений.
- Предоставить доступ к Полигону алгоритмов классификации — распределенной системе тестирования алгоритмов классификации на реальных прикладных задачах.

Основные принципы

Ресурс строится по принципам Википедии — свободной энциклопедии.

Содержимое Ресурса создаётся всеми его пользователями и является общественным достоянием. Каждый пользователь ресурса может создать или модифицировать статью или раздел (категорию), в любое время, в любом месте, располагая только доступом в Интернет.

Главное отличие от Википедии — профессиональная направленность тематики. Допускается (и поощряется) пополнение Ресурса специальными, полемическими и учебными материалами, информацией о незавершенных исследованиях, исходными кодами алгоритмов и программ. По этим причинам Ресурс не может являться частью Википедии. В то же время, не исключается возможность обмена материалами с Википедией и другими сетевыми энциклопедиями.

Новые статьи

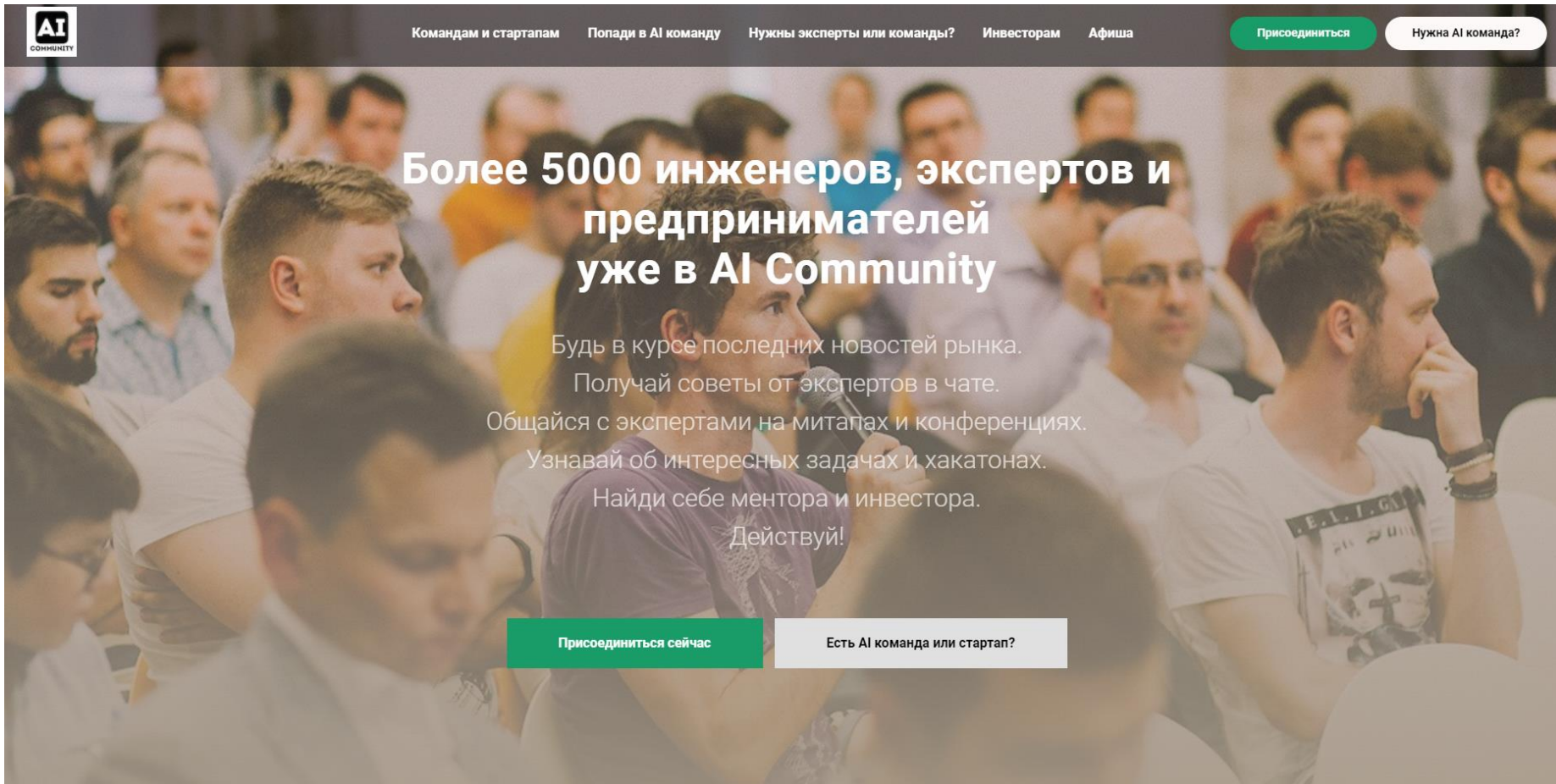
- Практикум на ЭВМ (417)/2017 (Sergey Ivanov) – [15.25, 4 сентября 2017]

Последние новости

- 15 апреля 2017 года — Конференция «Математические методы распознавания образов» (ММРО-2017), проводимая в Таганроге с 9 по 13 октября 2017 г., объявляет о приеме докладов. Срок подачи докладов до 1 сентября.
- 14 марта 2017 года — Принимаются работы в специальный выпуск Special Issue on Advances in Evolutionary Multi-objective Optimization журнала Swarm and Evolutionary Computation. Срок подачи работ до 30 мая, подробнее.
- 14 декабря 2016 года — Программный комитет конференции Интеллектуализация обработки информации объявляет о приеме исследовательских статей в сборник Intelligent Data Processing: Springer, Communications in Computer and Information Science series. Срок подачи работ до 20 февраля, подробнее.
- 14 ноября 2016 года — Сайт IFORS Developing Countries OR Resources, посвященный проблемам оптимизации и анализа данных в экономике, приглашает авторов к размещению публикаций.
- 23 августа 2016 года — В рамках конференции AINL-FRUCT-2016 проводится секция Business Intelligence, принимаются статьи в сборник, индексируемый Web of Sciences. Срок до 15 сентября.
- 13 августа 2016 года — Python Data Science meetup – Avito проводит встречу специалистов по Data Science и машинному обучению. В числе докладов будет рассказ победителя прошедшего конкурса Avito-2016 по распознаванию марки и модели автомобилей по изображениям.
- 1 августа 2016 года — стартует второй этап конкурса по распознаванию категории объявления, проводимого при информационной поддержке 11-й Международной конференции «Интеллектуализация обработки информации».
- 19 июля 2016 года — Специализация МФТИ «Машинное обучение и анализ данных» — серия онлайн курсов на сайте coursera.org приглашает слушателей, желающих быстро освоить практику и теорию профессии, научиться решать типовые промышленные задачи. Уже сейчас курс слушают несколько тысяч человек.

Все новости

ai-community.com

The image shows the header and main banner of the AI Community website. The header is a dark grey bar with the AI Community logo on the left and navigation links in the center. The main banner features a background image of a diverse group of people at a conference, with a man in the foreground speaking into a microphone. Overlaid on this image is promotional text and two call-to-action buttons.

AI
COMMUNITY

Командам и стартапам Попади в AI команду Нужны эксперты или команды? Инвесторам Афиша

[Присоединиться](#) [Нужна AI команда?](#)

Более 5000 инженеров, экспертов и предпринимателей уже в AI Community

Будь в курсе последних новостей рынка.
Получай советы от экспертов в чате.
Общайся с экспертами на митапах и конференциях.
Узнавай об интересных задачах и хакатонах.
Найди себе ментора и инвестора.
Действуй!

[Присоединиться сейчас](#) [Есть AI команда или стартап?](#)