Машинное обучение в эпоху больших данных

Воронцов Константин Вячеславович
ФУПМ МФТИ • ВМК МГУ • Яндекс • FORECSYS

5 июля 2016
Сочи, Сириус • Проектная смена • 1–24 июля 2016

Цели этой лекции

Научиться замечать задачи *машинного обучения* в окружающей нас с вами информационной среде.

Получить первоначальное общее представление о:

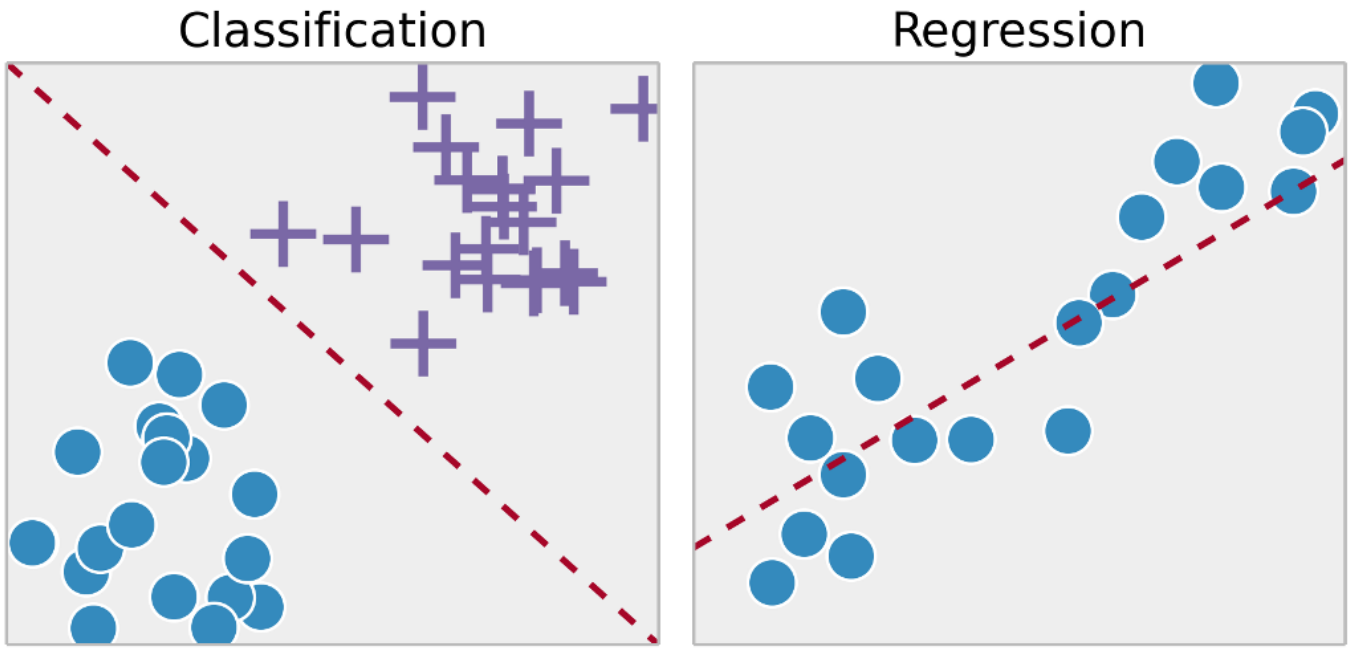
- задачах,
- методах,
- последних достижениях,
- научном сообществе

машинного обучения.

Статистическое (машинное) обучение с учителем

= обучение по прецедентам
= восстановление зависимостей по эмпирическим данным
= предсказательное моделирование
= проведение функции через заданные точки

Два основных типа задач — *классификация* и *регрессия*



Задача статистического (машинного) обучения с учителем

Задача восстановления зависимости $y = f(x)$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: векторы объектов $x_i = (x_i^1, \ldots, x_i^n)$, ответы $y_i = f(x_i)$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: функцию классификации $a(x)$, способную давать ответы на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($y \in \mathbb{N}$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — геологический район; y — наличие месторождения;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($y \in \mathbb{R}$ или $\mathbb{R}^m$):
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура химического соединения; y — его свойство;
  - x — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — характеристики недвижимости; y — цена;

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы — диагноз или способ лечения или исход заболевания.

Примеры признаков:

- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- как правило, недостаточный объём данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности (риска | успеха | исхода).

Задачи распознавания месторождений

Объект — геологический район (рудное поле).

Классы — есть или нет полезное ископаемое.

Примеры признаков:

- **бинарные:** присутствие крупных зон смятия и рассланцевания, и т. д.
- **порядковые:** минеральное разнообразие; мнения экспертов о наличии полезного ископаемого, и т. д.
- **количественные:** содержания сурьмы, присутствие в рудах антимонита, и т. д.

Особенности задачи:

- проблема «малых данных» — для редких типов месторождений объектов много меньше, чем признаков.

Задачи биометрической идентификации личности

Идентификация по отпечаткам пальцев

Идентификация по радужной оболочке глаза

Особенности задач:

- нетривиальная предобработка для извлечения признаков;
- высочайшие требования к точности.

Задача категоризации текстовых документов

Объект — текстовый документ.

Классы — рубрики иерархического тематического каталога.

Примеры признаков:

- **номинальные:** автор, издание, год, и т. д.
- **количественные:** для каждого термина — частота в тексте, в заголовках, в аннотации, и т. д.

Особенности задачи:

- лишь небольшая часть документов имеют метки y_i;
- документ может относиться к нескольким рубрикам;
- в каждом ребре дерева свой классификатор на 2 класса.

Задача ранжирования поисковой выдачи

Объект — пара $\langle$короткий запрос, документ$\rangle$.

Классы — асессорские оценки релевантности.

Примеры признаков:

- **количественные:**
  частота слов запроса в документе,
  число ссылок на документ,
  число кликов на документ: всего, по данному запросу,
- **номинальные:**
  ID пользователя, ID региона, язык запроса.

Особенности задачи:

- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.

Задача ранжирования в рекомендательных системах

Объект — пара ⟨клиент, товар⟩
(товары — книги, фильмы, музыка).

Предсказать: вероятность покупки или рейтинг товара.

Примеры признаков:

- **количественные:**
  рейтинг схожих товаров для данного клиента;
  рейтинг данного товара для схожих клиентов;
  вектор интересов клиента;
  вектор интересов товара;

Особенности задачи:

- сверхбольшие разреженные данные;
- интересы скрыты, их надо сначала выявить.

Разведочный информационный поиск (exploratory search)

Объект — пара $\langle$длинный запрос, документ$\rangle$.

Найти: что ещё известно по этой теме.

Примеры приложений:

- поиск и мониторинг научно-технической информации,
- выявление эпидемий по поисковым логам,
- оценка социальной напряжённости по данным соцсетей.

Особенности задачи:

- темы скрыты, их надо сначала выявить;
- лишь небольшая часть документов может быть размечена;
- плохо формализуемые критерии качества;
- необходимость продвинутой визуализации.

Задача прогнозирования стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, и т. д.
- **номинальные:** район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- **количественные:** число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- выборка неоднородна, стоимость меняется со временем;
- разнотипные признаки;
- для линейной модели нужны преобразования признаков.

Задача прогнозирования объёмов продаж

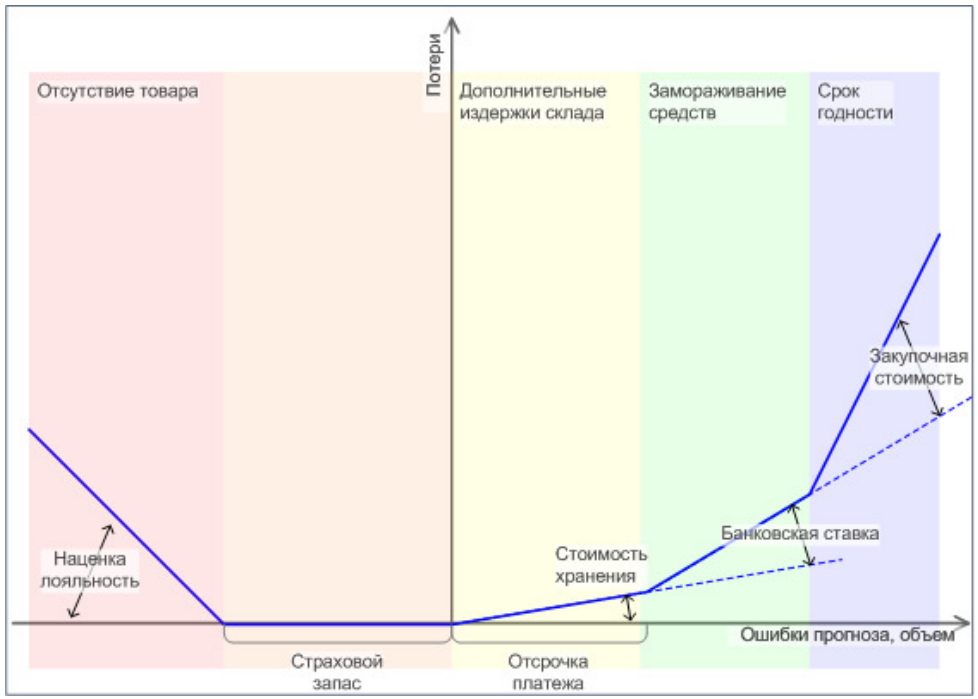
Объект — тройка $\langle$товар, магазин, день$\rangle$.

Примеры признаков:

- **бинарные:** выходной день, праздник, промоакция, и т. д.
- **количественные:** объёмы продаж в предшествующие дни.

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Конкурс kaggle.com: TFI Restaurant Revenue Prediction

Объект — место для открытия нового ресторана.

Предсказать — прибыль от ресторана через год.

Примеры признаков:

- демографическими свойствами района;
- цены на недвижимость поблизости;
- маркетинговые данные: наличие школ, офисов и т.д.

Особенности задачи:

- мало объектов, много признаков;
- разнотипные признаки;
- есть нетипичные объекты — «выбросы»;
- разнородные объекты (возможно, имеет смысл строить разные модели для мелких и крупных городов).

Конкурс kaggle.com: Avito Context Ad Clicks Prediction

Объект — тройка $\langle$пользователь, запрос, объявление$\rangle$.

Предсказать — кликнет ли пользователь по контекстной рекламе, которую показали по его запросу на avito.ru.

Сырые данные:

- все действия пользователя на сайте,
- профиль пользователя (бразуер, устройство и т. д.),
- история показов и кликов других пользователей,
- . . . всего 10 таблиц данных.

Особенности задачи:

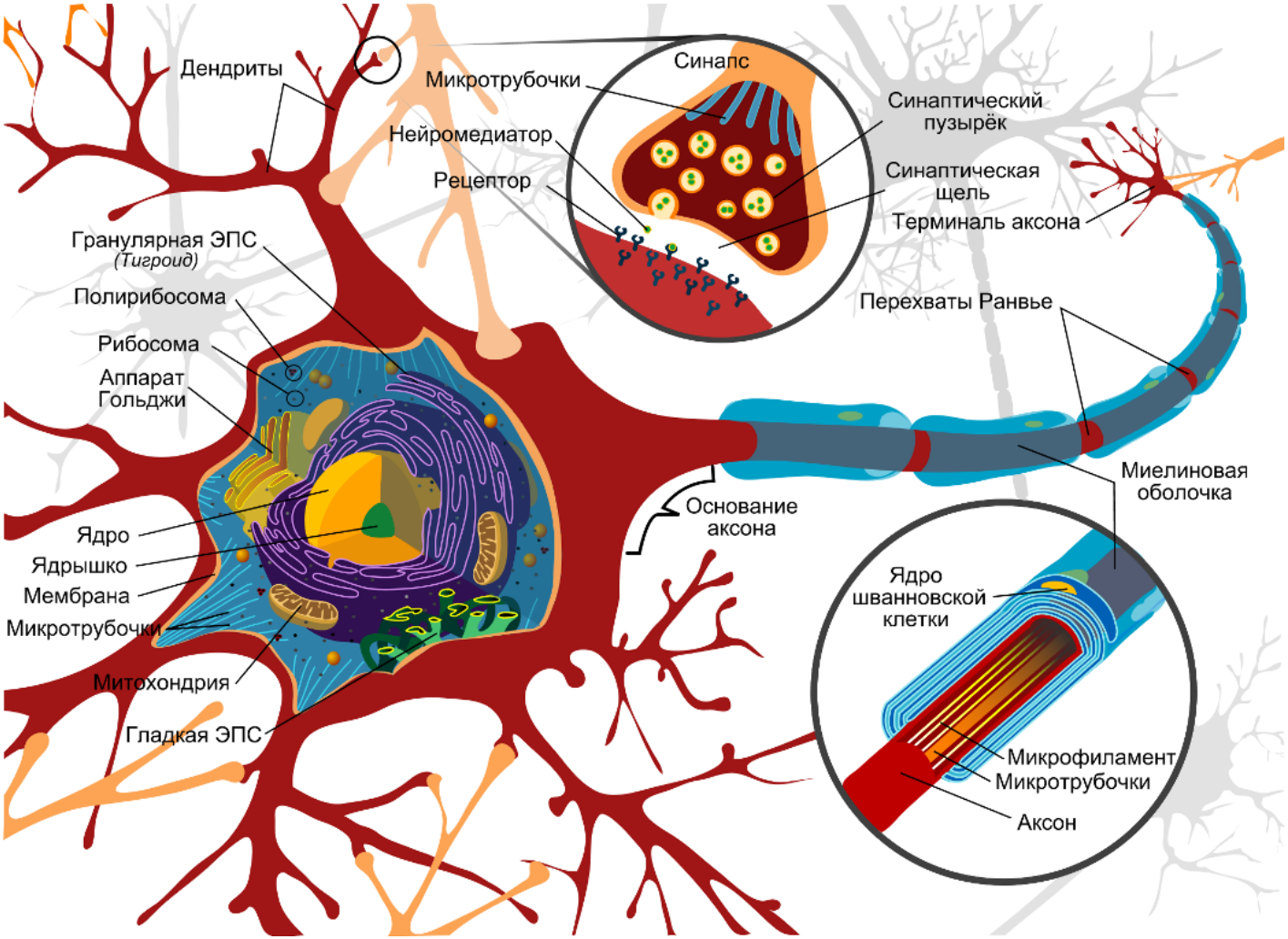
- признаки надо придумывать;
- данных много — сотни миллионов показов;
- основной критерий качества — доход рекламной площадки;
- но имеются и дополнительные критерии.

Особенности данных и постановок прикладных задач

- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- неструктурированные (нет признаковых описаний)

- заказчик не знает точно, чего хочет
- критерии качества нетривиальны или неясны
- заказчик не заботится о качестве своих данных

Нервная клетка — естественный нейрон

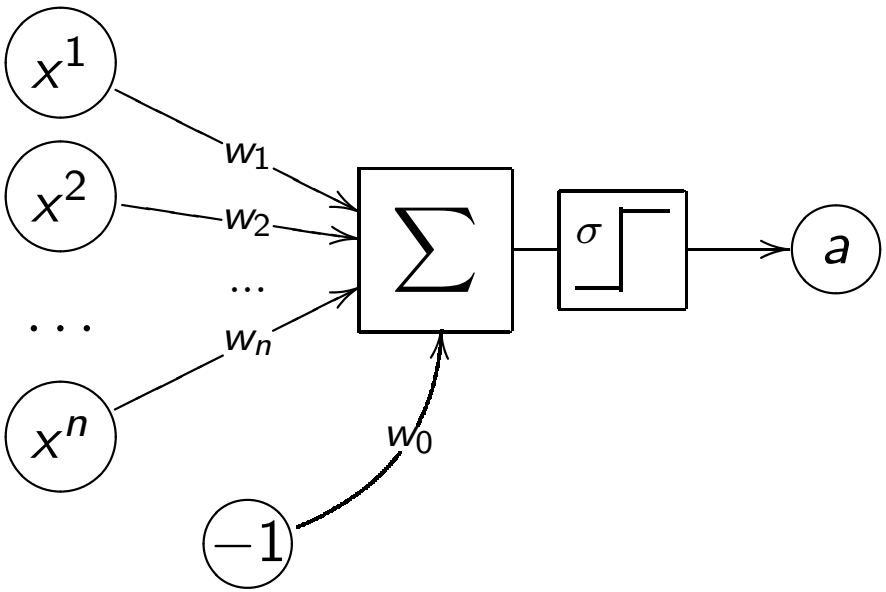


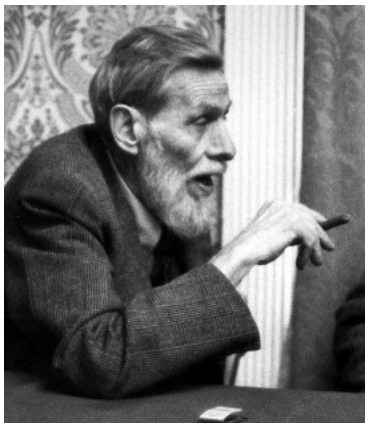
Модель МакКаллока–Питтса — искусственный нейрон

Линейная модель нейрона (1943):

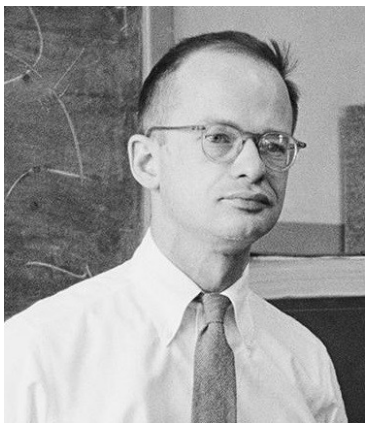
$$a(x, w) = \sigma\left(\sum_{j=1}^{n} w_j x^j - w_0\right),$$

где $\sigma(z)$ — функция активации,
например, $\text{sign}(z)$ или $\text{arcth}(z)$





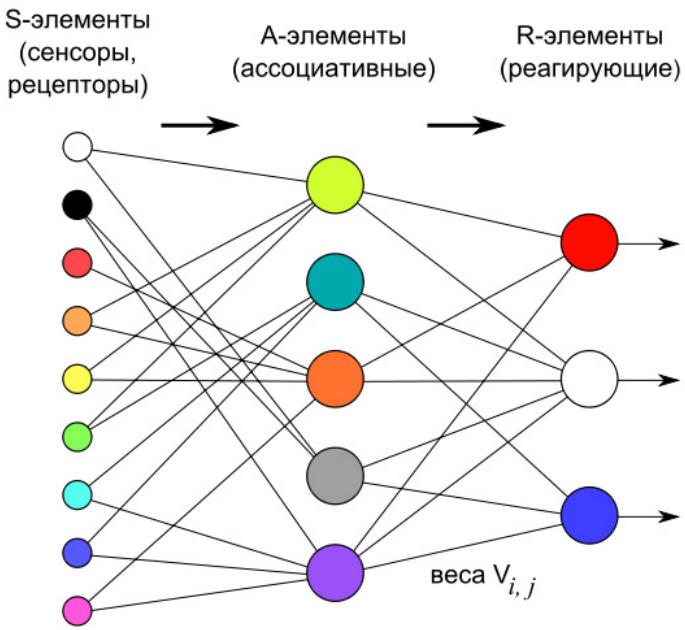
Уоррен МакКаллок
(1898–1969)



Вальтер Питтс
(1923–1969)

Персептрон Розенблатта (1957)

Mark-1 — первый нейрокомпьютер (1960)
Обучение — метод коррекции ошибки
Архитектура — двухслойная сеть



Фрэнк Розенблатт
(1928–1971)

Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга. 1965 (1962).

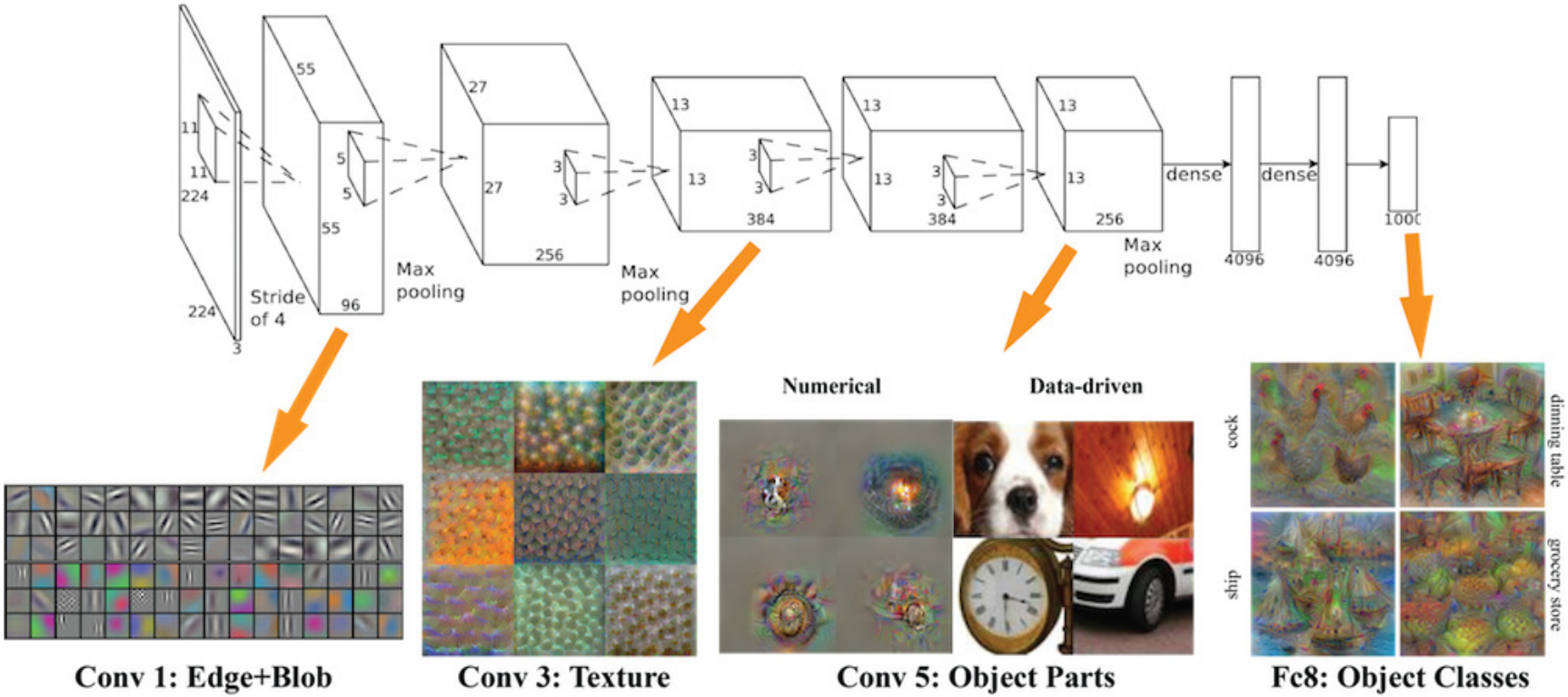
Глубокие нейронные сети для обработки изображений

Цель — извлечение признаков из сырых данных.

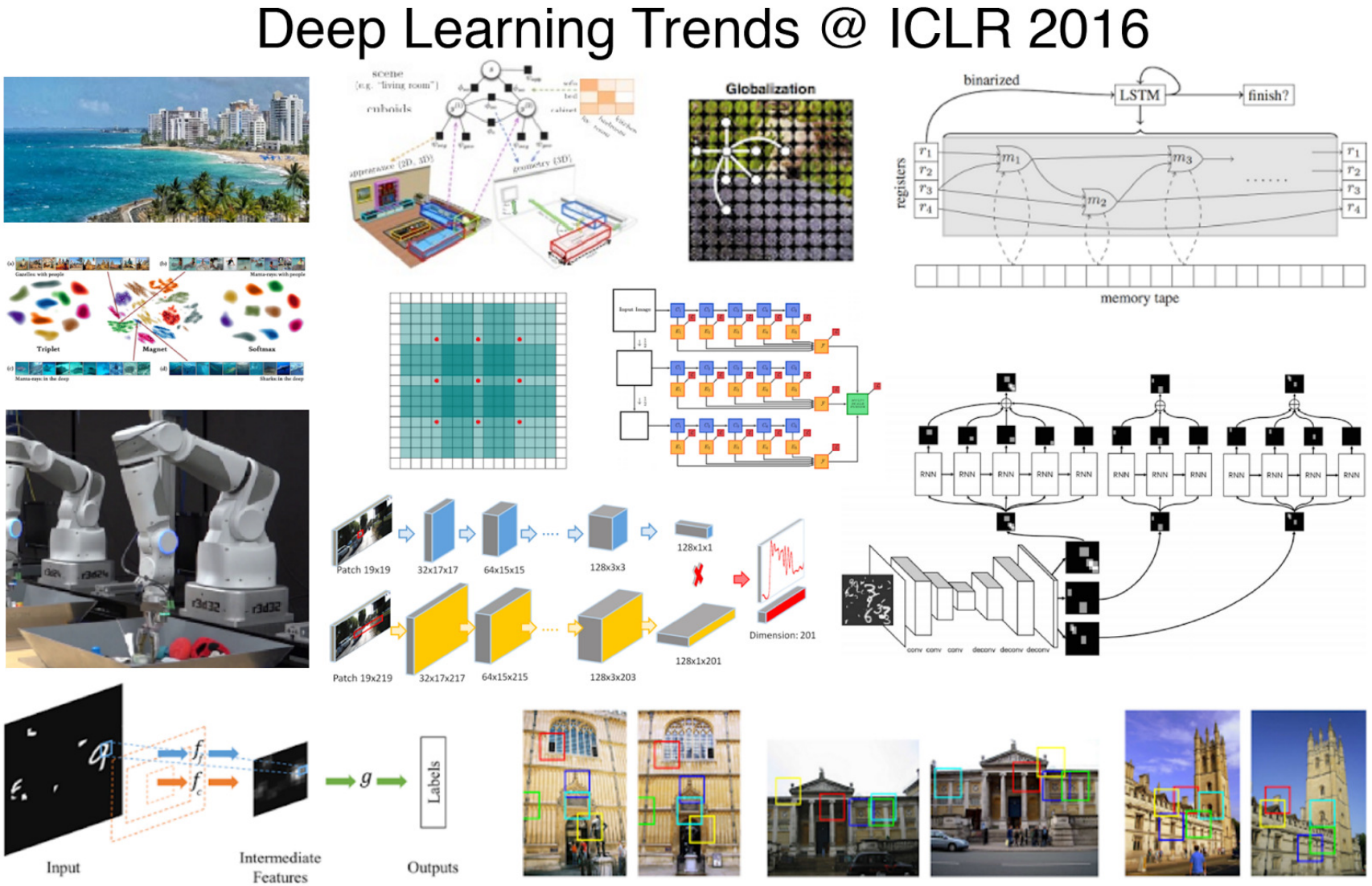
Нейрон — это линейная комбинация признаков.

Свёрточный нейрон комбинирует признаки соседних пикселей.

Выход нейрона — это признак для нейронов следующего слоя.



Разнообразие приложений глубокого обучения



Глубокие нейронные сети для обработки изображений

Нейронная сеть, обученная распознаванию художников/стилей, позволяет синтезировать изображения, смешивая форму заданного изображения с любым из выученных стилей.

Арсенал технологий

Экосистемы машинного обучения:

- Python + SciPy + SciKit-Learn
- Java + Weka
- R

Инструменты для хранения и обработки больших данных:

- Hadoop — распределённое хранение данных
- Spark — распределённые вычисления

Инструменты для обучения нейронных сетей:

- TensorFlow
- Theano
- Torch

Полезные ссылки

- www.kaggle.com — конкурсы анализа данных
- www.kdnuggets.com — главный сайт датамайнеров
- www.MachineLearning.ru — русскоязычная вики
- www.datasciencecentral.com — 72 000 датамайнеров
- archive.ics.uci.edu/ml — UCI ML Repository (349 datasets)
- ru.coursera.org/learn/machine-learning — курс Эндрю Ына
- ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie — курс Воронцова от ВШЭ и ШАД Яндекс
- ru.coursera.org/specializations/machine-learning-data-analysis — специализация от МФТИ и ШАД Яндекс

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Если что-то было не понятно,
не стесняйтесь подходить и спрашивать :)