Машинное обучение и анализ данных

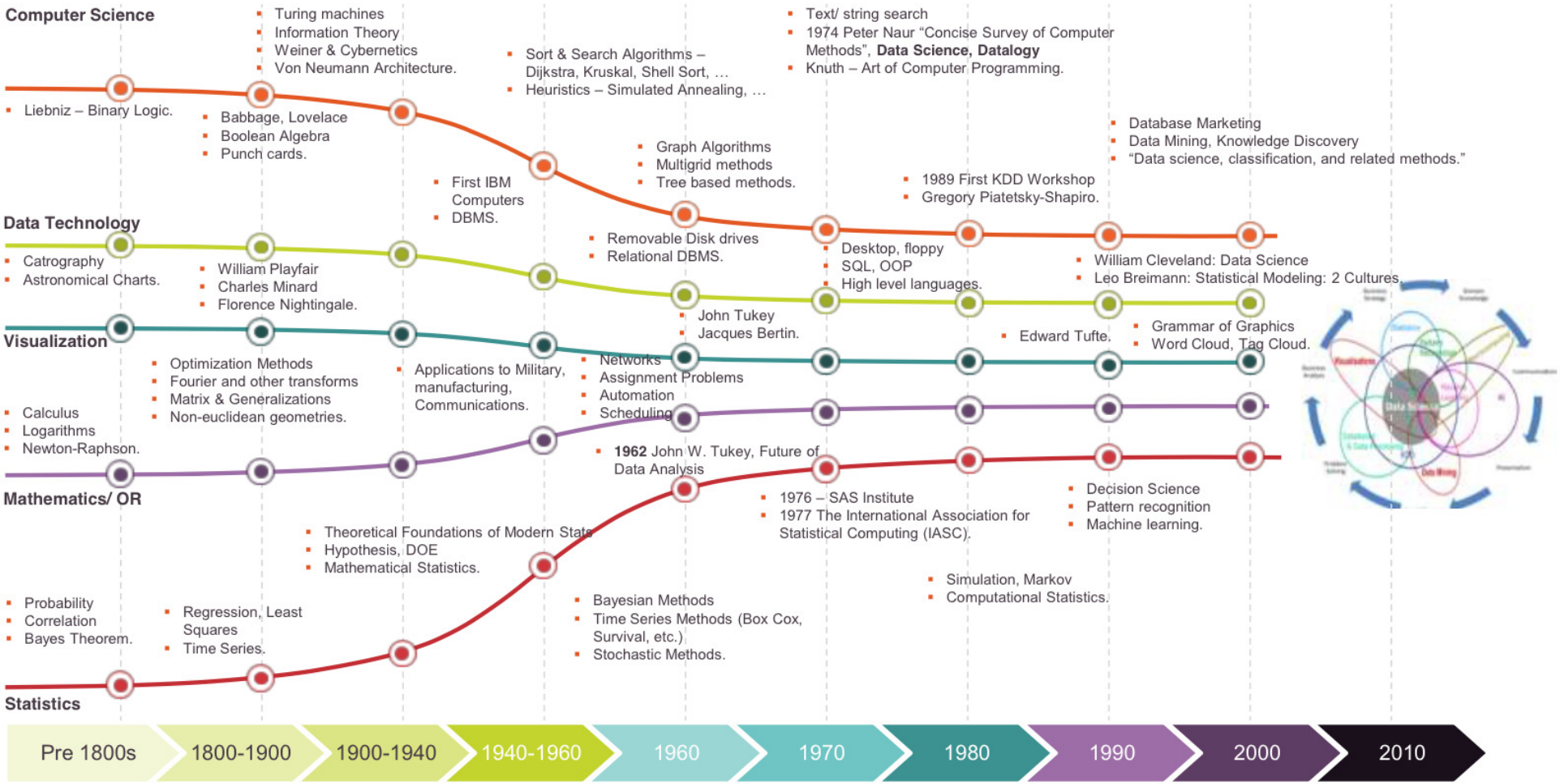
Воронцов Константин Вячеславович
(ФИЦ ИУ РАН • МФТИ • Форексис)

Научный семинар ВНИИА • 17 ноября 2016

Машинное обучение «и много других страшных слов»

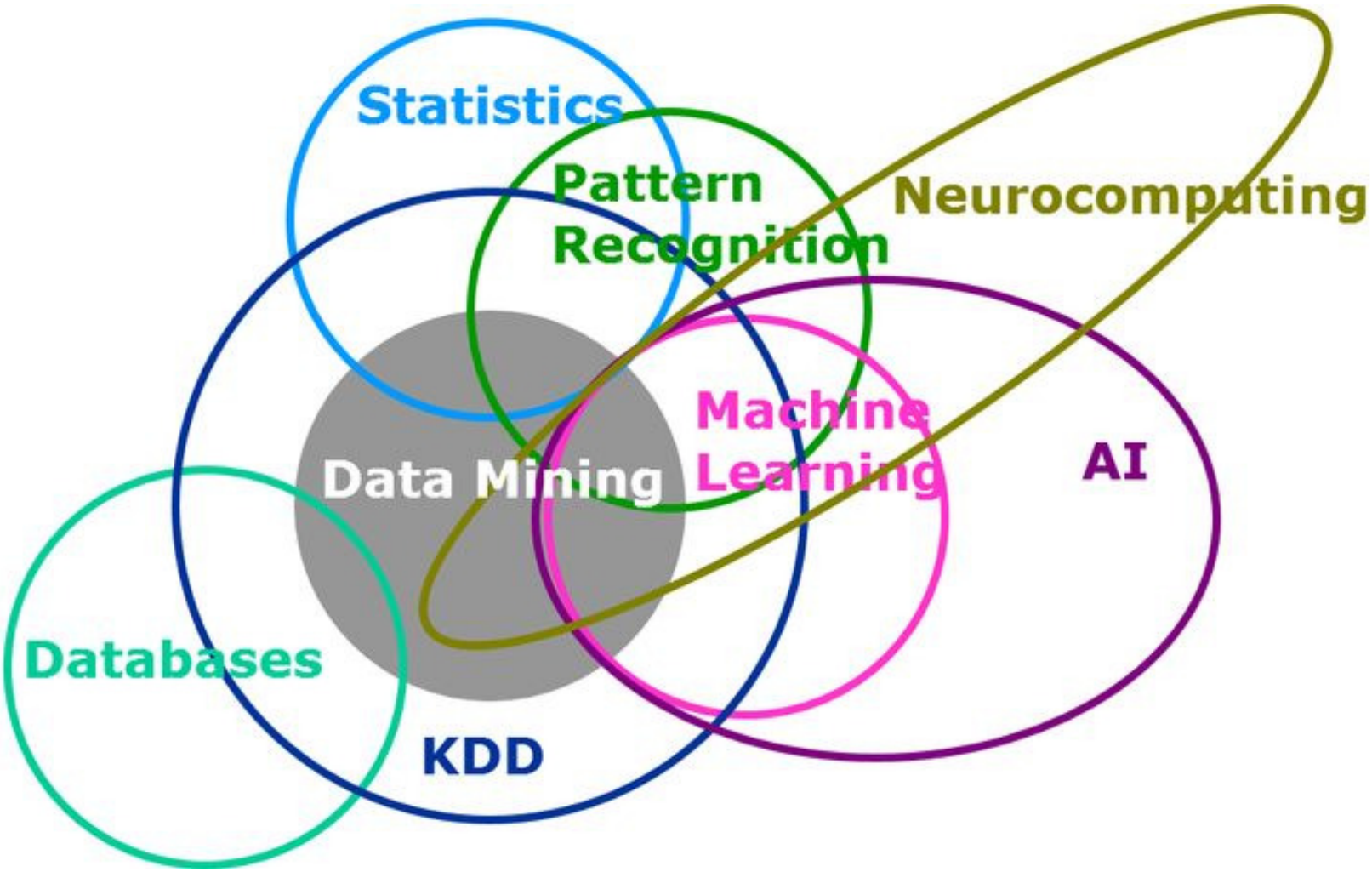
- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) — 1955
- Распознавание образов (Pattern Recognition)
- Машинное обучение (Machine Learning) — 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) — 1989
- Knowledge Discovery in Databases — 1989
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) — 2007
- Большие данные (Big Data) — 2008
- Аналитика больших данных (Big Data Analytics)
- Науки о данных (Data Science) — 2011

Предпосылки Data Science



http://www.kdnuggets.com/2015/02/history-data-science-infographic.html

Структура смежных областей Data Science



Машинное обучение — это...

- одна из ключевых технологий будущего
- наиболее успешное направление искусственного интеллекта, часто противопоставляемое инженерии знаний
- симбиоз математической статистики и численных методов оптимизации
- математическое моделирование в условиях, когда данных много, знаний мало
- проведение функции через заданные точки в сложно устроенных пространствах

Задачи обучения с учителем

Восстановление зависимости $y(x)$ по точкам (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: выборка ℓ объектов $x_i = (f_1(x_i), \ldots, f_n(x_i))$ и ответов y_i,
$f_j(x)$ — признаки объекта x, $j = 1, \ldots, n$

$$\begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную давать правильные ответы на *тестовых объектах* $x'_i = (f_1(x'_i), \ldots, f_n(x'_i))$, $i = 1, \ldots, k$:

$$\begin{pmatrix} f_1(x'_1) & \ldots & f_n(x'_1) \\ \ldots & \ldots & \ldots \\ f_1(x'_k) & \ldots & f_n(x'_k) \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(x'_1) \\ \ldots \\ a(x'_k) \end{pmatrix}$$

Типология задач машинного обучения

1. Обучение с учителем (supervised learning)
   - классификация (classification)
   - регрессия (regression)
   - ранжирование (learning to rank)
   - прогнозирование (forecasting)
2. Обучение без учителя (unsupervised learning)
   - кластеризация (clustering)
   - поиск ассоциативных правил (association rule learning)
   - восстановление плотности (density estimation)
   - одноклассовая классификация (anomaly detection)
3. Частичное обучение (semi-supervised learning)
   - трансдуктивное обучение (transductive learning)
4. Предварительная обработка (data preparation)
   - извлечение признаков (feature extraction)
   - отбор признаков (feature selection)
   - восстановление пропусков (missing values)

Типология задач машинного обучения (продолжение)

5. Обучение представлений (representation learning)
   - обучение признаков (feature learning)
   - обучение многообразий (manifold learning)
   - анализ главных компонент (principal component analysis)
   - матричные разложения (matrix factorization)
   - коллаборативная фильтрация (collaborative filtering)
   - тематическое моделирование (topic modeling)
6. Обучение выявлению связей (relational learning)
7. Динамическое обучение (online/incremental learning)
8. Обучение с подкреплением (reinforcement learning)
9. Активное обучение (active learning)
10. Привилегированное обучение (privilege learning)
11. Обучение с переносом опыта (transfer learning)
12. Мета-обучение (meta-learning)

Основные школы машинного обучения

1. символизм
   - Decision Tree, Rule Induction
2. коннекционизм
   - BackPropagation, Deep Belief Nets, Deep Learning
3. эволюционизм
   - Genetic Algorithms, Genetic Programming
4. байесионизм
   - Naive Bayes, Bayesian Networks, Graphical Models
5. аналогизм
   - kNN, RBF, SVM, Kernel Regression, Kernel Density Estimation
6. композиционизм (+)
   - Weighted Voting, Boosting, Bagging, Stacking, Random Forest, Яндекс.MatrixNet, xgboost

Домингос П. Верховный алгоритм. 2016. 336 с.

Типология методов машинного обучения

- минимизация эмпирического риска
  MVR, Linear Regression, Logistic Regression
- регуляризация эмпирического риска
  SVM, RLR, ElasticNet, LASSO, Least Angle Regression
- метрические методы
  kNN, RBF, Kernel Regression, Kernel Density Estimation
- логические методы и отбор признаков
  DT, DL, DF, Rule Induction, Add, Add-Del, BFS, МГУА, СПА, GA
- байесовские методы
  Naive Bayes, Linear Discriminant, Bayesian Networks
- нейросетевые методы
  BackPropagation, Deep Belief Nets, Deep Learning
- композиционные методы
  Boosting, Bagging, Stacking, Random Forest, Яндекс.MatrixNet

Особенности данных и постановок прикладных задач

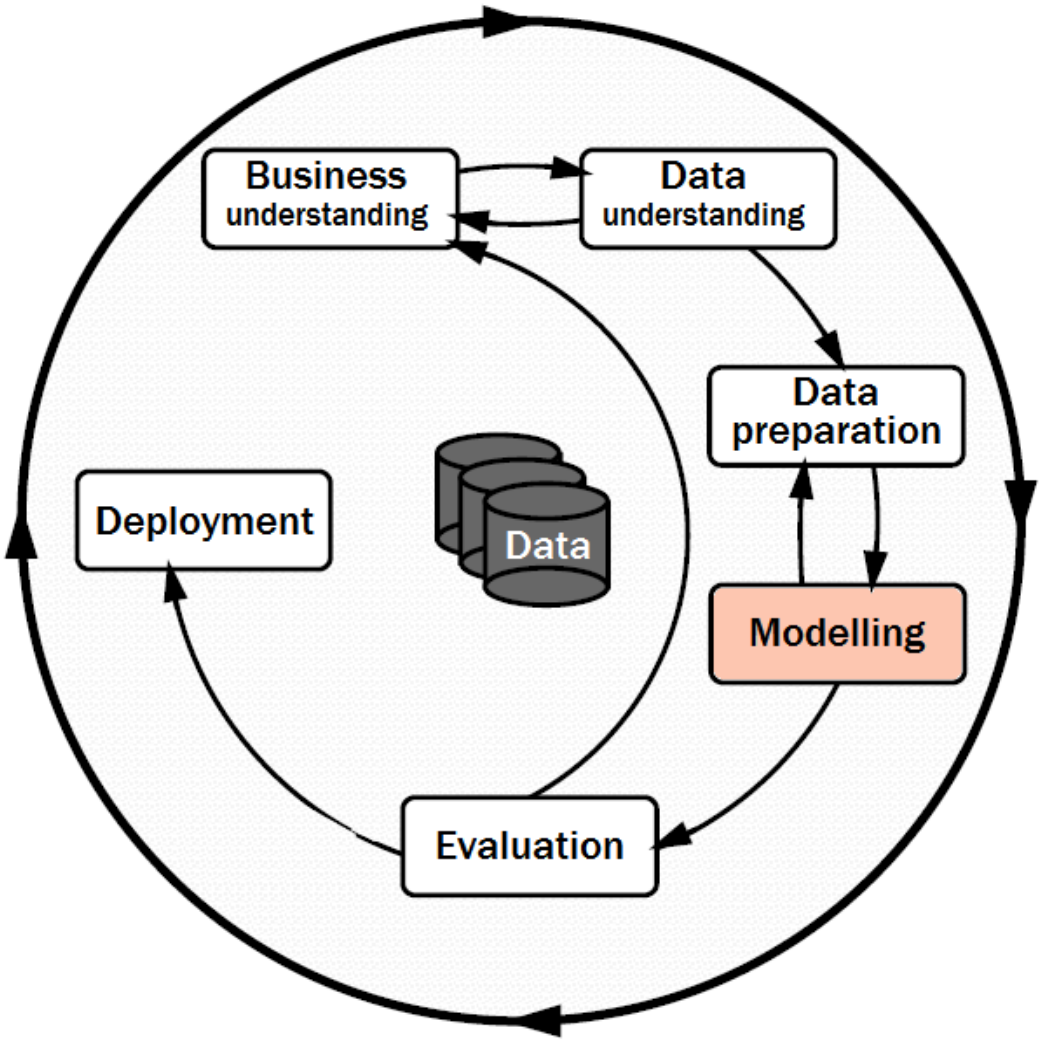
Данные...

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

Заказчик...

- не знает точно, чего хочет
- не имеет чётких критериев качества (KPI)
- не заботится о качестве своих данных
- не понимает роль математики и ожидает чуда

CRISP-DM (1999) — межотраслевой стандарт Data Mining



Этапы решения задач машинного обучения

- понимание задачи и данных
- предобработка данных и изобретение признаков
- построение модели
- сведение обучения к оптимизации
- решение проблем оптимизации и переобучения
- оценивание качества решения
- внедрение и эксплуатация

Арсенал технологий

Экосистемы машинного обучения:

- Python + SciPy + SciKit-Learn
- Java + Weka + RapidMiner
- R
- Deductor — аналитическая платформа BaseGroup Labs

Инструменты для хранения и обработки больших данных:

- Hadoop — распределённое хранение данных
- Spark — распределённые вычисления

Инструменты для обучения нейронных сетей:

- TensorFlow
- Theano
- Torch

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 — геологич. объект; y — наличие полезного ископаемого;
  - 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 — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - 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 (болезни | 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 кредитного скоринга

Объект — заявка на выдачу банком кредита.

Классы — bad или good.

Примеры признаков:

- **бинарные:** пол, наличие телефона, и т. д.
- **номинальные:** место проживания, профессия, работодатель, и т. д.
- **порядковые:** образование, должность, и т. д.
- **количественные:** возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

Особенности задачи:

- нужно оценивать вероятность дефолта $P(\text{bad})$
- и риск всего кредитного портфеля банка

Задача предсказания оттока клиентов

Объект — абонент в определённый момент времени.

Классы — уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- **бинарные:** корпоративный клиент, включение услуг, и т. д.
- **номинальные:** тарифный план, регион проживания, и т. д.
- **количественные:** длительность разговоров (входящих, исходящих, СМС, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- сложно идентифицировать факт ухода;
- нужно оценивать вероятность ухода;
- сверхбольшие выборки;
- не ясно, какие признаки вычислять по «сырым» данным.

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 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

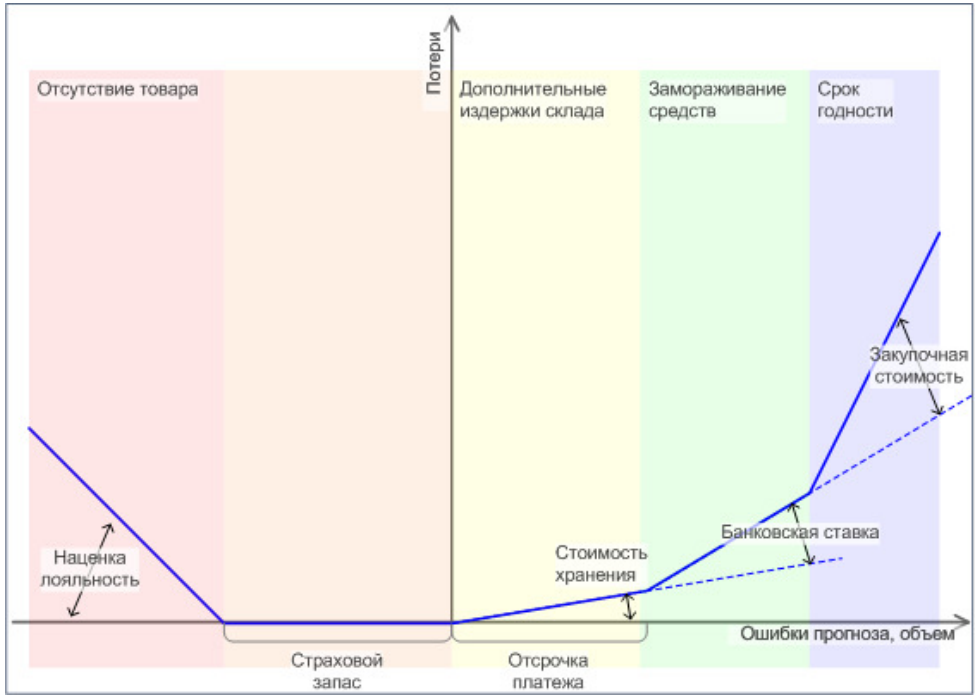
Объект — тройка $\langle$товар, магазин, день$\rangle$.

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:

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

Особенности задачи:

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

Задача ранжирования поисковой выдачи

Объект — пара $\langle$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.

1. **Анализ данных жидкостной хроматографии**

$$z(t,\lambda) = \sum_i X_i(t)\,Y_i(\lambda)$$

дано: $z(t,\lambda)$ — выход сканирующего УФ-детектора;
найти: $X_i(t)$ — хроматограмма i-го вещества, t — время
$Y_i(\lambda)$ — спектр i-го вещества, λ — длина волны.

2. **Анализ данных ДНК-микрочипов**

$$I(p,k) = \sum_g a_{pg}\,C_{gk}$$

дано: $I(p,k)$ — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
C_{gk} — концентрация g-го гена на k-м чипе.

3. **Рекомендательные системы**

$$R_{iu} = \sum_t p_i(t) q_u(t)$$

дано: R_{iu} — рейтинги товаров i, поставленные пользователем u;
найти: $p_i(t)$ — профиль интересов товара i;
$q_u(t)$ — профиль интересов пользователя u.

4. **Тематическое моделирование текстовых коллекций**

$$p(w|d) = \sum_t p(w|t) p(t|d)$$

дано: $p(w|d) = \frac{n_{dw}}{n_d}$ — частоты слов w в документах d;
найти: $\phi_{wt} = p(w|t)$ — распределения слов w в темах t,
$\theta_{td} = p(t|d)$ — распределения тем t в документах d.

дано: $p(w|d) = \frac{n_{dw}}{n_d}$ — доля расходов клиента d по категории w;
найти: $\phi_{wt} = p(w|t)$ — распределения категорий w в темах t,
$\theta_{td} = p(t|d)$ — распределения тем t у клиента d.

Разведочный информационный поиск (exploratory search)

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

Предсказать — тематическую близость документа запросу.

Примеры приложений:

- поиск и мониторинг научно-технической информации,
- выявление эпидемий по поисковым логам,
- выявление 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 ранжирования в рекомендательных системах

Объект — пара $\langle$клиент, товар$\rangle$
(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;
  оценки интересов клиента;
  оценки интересов товара;

Особенности задачи:

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

Конкурс kaggle.com: Avito Context Ad Clicks Prediction

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

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

Сырые данные:

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

Особенности задачи:

- признаки надо придумывать;
- данных много — сотни миллионов показов;
- основной критерий качества — доход рекламной площадки;
- несколько дополнительных критериев и ограничений.

Полезные ссылки

- www.MachineLearning.ru — русскоязычная вики
- www.kdnuggets.com — главный сайт датамайнеров
- www.datasciencecentral.com — 72 000 датамайнеров
- www.kaggle.com — конкурсы анализа данных
- 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

Рекомендуемая литература

- *Домингос П.* Верховный алгоритм. 2016. 336 с.
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016. 302 с.
- *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
- *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
- Машинное обучение (курс лекций, К. В. Воронцов). www.MachineLearning.ru. 2004–2016.
- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014. 739 p.
- *Bishop C. M.* Pattern Recognition and Machine Learning. - Springer, 2006. 738 p.