СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 24. Введение в машинное обучение. Часть 2

Обучение без учителя

Машинное обучение без учителя (unsupervised learning): немаркированные данные.

Цель обучения без учителя – обнаружение в наборе данных явных и скрытых шаблонов, общих черт, что позволяет обнаруживать схожесть, необходимую для классификации необработанных данных.

Примеры использования обучения без учителя:

- работа с транзакционными данными (имеется набор данных о клиентах и их покупках и требуется обнаружить схожие атрибуты в профилях клиентов и их типах покупок);
- обнаружение аномалий (например, выявление мошенничества с кредитными картами);
- создание систем рекомендаций, которые советуют пользователю, какие продукты купить/ какой фильм посмотреть и т.п. на основе его предпочтений.

Задачи, возникающие при обучении без учителя

1. **Задача кластеризации** (clustering): группировка объектов в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.
2. **Задача поиска ассоциативных правил** (association rules learning): исходные данные представляются в виде признаковых описаний и требуется найти такие наборы признаков и их значения, которые неслучайно часто встречаются в описаниях объектов.
3. **Задача фильтрации выбросов** (outliers detection) : обнаружение в обучающей выборке небольшого числа нетипичных объектов. В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества). В других приложениях эти объекты являются следствием ошибок в данных или неточности модели, то есть шумом, мешающим настраивать модель, и должны быть удалены из выборки.

Задачи, возникающие при обучении без учителя

4. **Задача построения доверительной области** (quantile estimation) — области минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки.
5. **Задача сокращения размерности** (dimensionality reduction) заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки. В классе линейных преобразований наиболее известным примером является метод главных компонент. Другие методы отбора существенных признаков или факторов были рассмотрены нами ранее в лекции, посвященной методам регуляризации в задаче множественной линейной регрессии. Это гребневая регрессия и метод лассо.
6. **Задача заполнения пропущенных значений** (missing values) — замена недостающих значений в матрице объекты–признаки их прогнозными значениями.

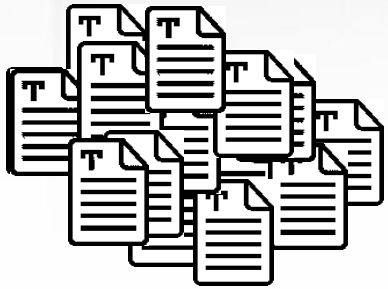
Постановка задачи кластеризации

Рассматривается множество объектов (ситуаций) X. Задано **подмножество прецедентов** $X^l = \{x_1, \dots, x_l\} \subset X$, по каждому из которых собраны (измерены) некоторые данные. Задана функция расстояния между объектами $\rho(x, x')$, где $x, x' \in X$.

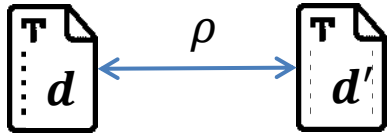
Задача. Разбить выборку X на непересекающиеся подмножества, называемые **кластерами**, так, чтобы каждый кластер состоял из объектов, близких по метрике $\rho(x, x')$, а объекты разных кластеров существенно отличались.

Пример постановки задачи кластеризации

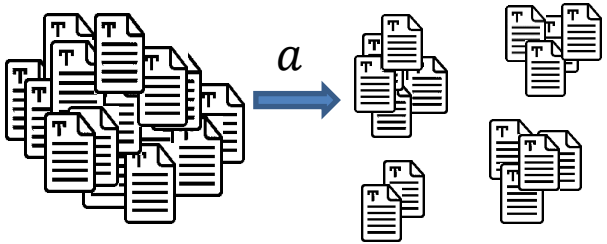
Конечное **множество прецедентов**:
$X = \{d_i\}$ – документы:



Метрика $\rho(d, d')$ оценивает схожесть документов по выделенному множеству ключевых слов



Решающая функция a – алгоритм кластеризации:

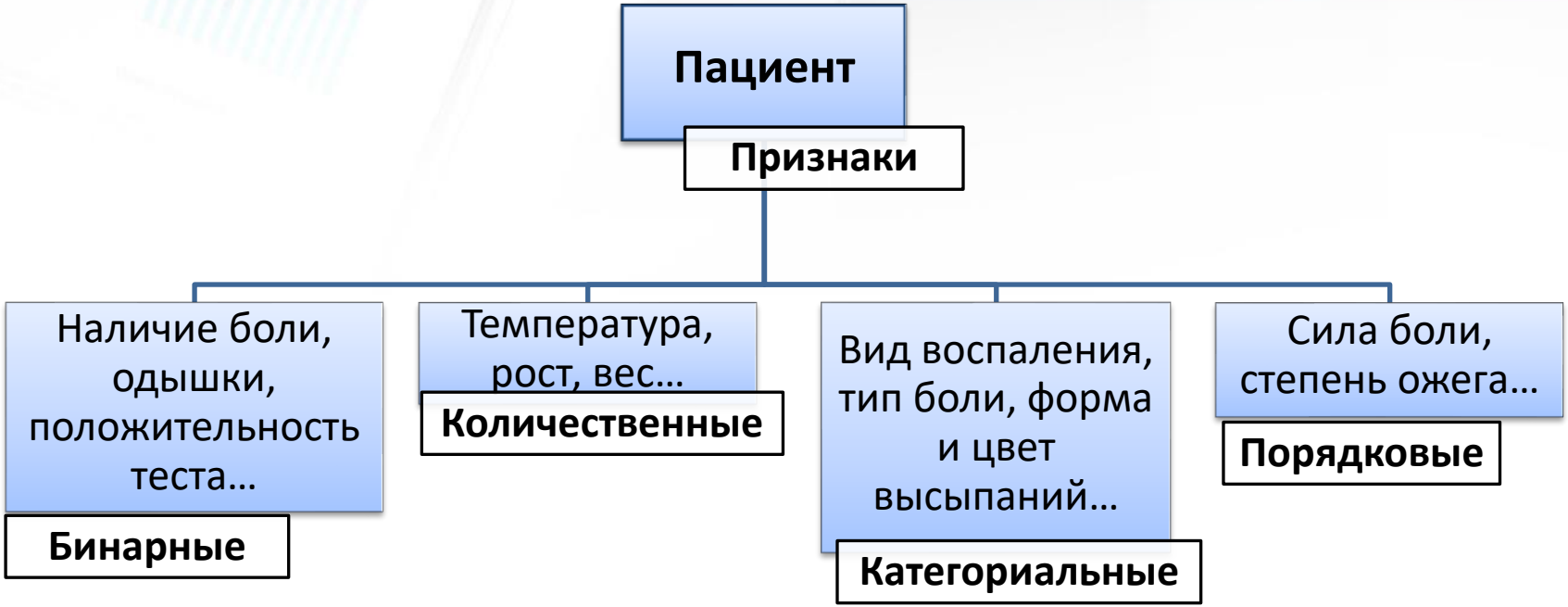


Типы входных данных

Входными данными являются некоторые наборы **признаков** объекта $x \in X$.
Признаком называется числовая характеристика объекта. Формально признак это отображение из множества объектов X в множество D_f числовых значений признака - $f: X \to D_f$.

Типы признаков:
Бинарный признак - D_f={0,1}.
Количественный признак - D_f=$\mathbb{R}$
Категориальный признак - D_f={1,2,…,K} (например цвет - красный, синий)
Порядковый признак - D_f - конечное упорядоченное множество (например рост – низкий, средний, высокий)

Пример входных данных

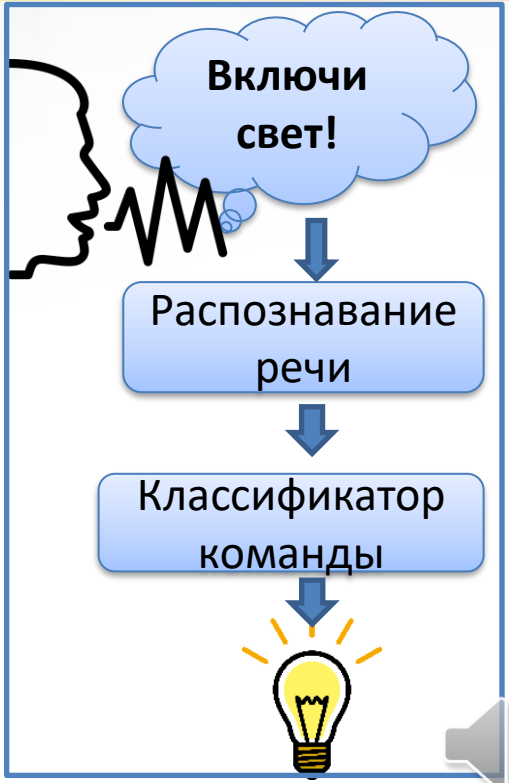


Сферы приложения

Биоинформатика;
Медицина(Медицинская диагностика);
Геология и геофизика;
Социология;
Экономика(оценка кредитных рисков, обнаружение мошенничества, биржевой технический анализ…);
Техника (техническая диагностика, робототехника, компьютерное зрение, распознавание речи…);
Офисная автоматизация (Распознавание текста, Обнаружение спама, Категоризация документов …).

Алгоритм распознавания

Лена Сёдерберг



Алгоритмы

Обучение с учителем:

- Регрессионный анализ
- Деревья решений
- Метод опорных векторов
- Байесовский классификатор
- Метод k ближайших соседей
- Нейронная сеть

Обучение без учителя:

- Метод k-средних
- Дискриминантный анализ
- EM-алгоритм
- Нейронная сеть
- Иерархическая кластеризация

Этапы машинного обучения

1. **Получение данных** (с устройств, измерений, баз данных и т.д.)
2. **Предобработка -** очистка, нормализация данных, фильтрация шумов и выбросов
3. **Понижение размерности** - выявление, отбор подмножества значимых, независимых признаков
4. **Отбор и разделение данных –** формирование выборки и разделение ее на обучающую и тестовую выборки
5. **Обучение** – классификация, регрессия, кластеризация и т.д.
6. **Тестирование**
7. **Анализ результатов**

Оценки качества обучения

Все метрики для оценки качества модели рассчитываются по тестовым (приемочным) выборкам. Тестовая выборка — это набор объектов и ответов, заранее известный и неиспользуемый при обучении.

Перекрестная проверка (кроссвалидация - crossvalidation):

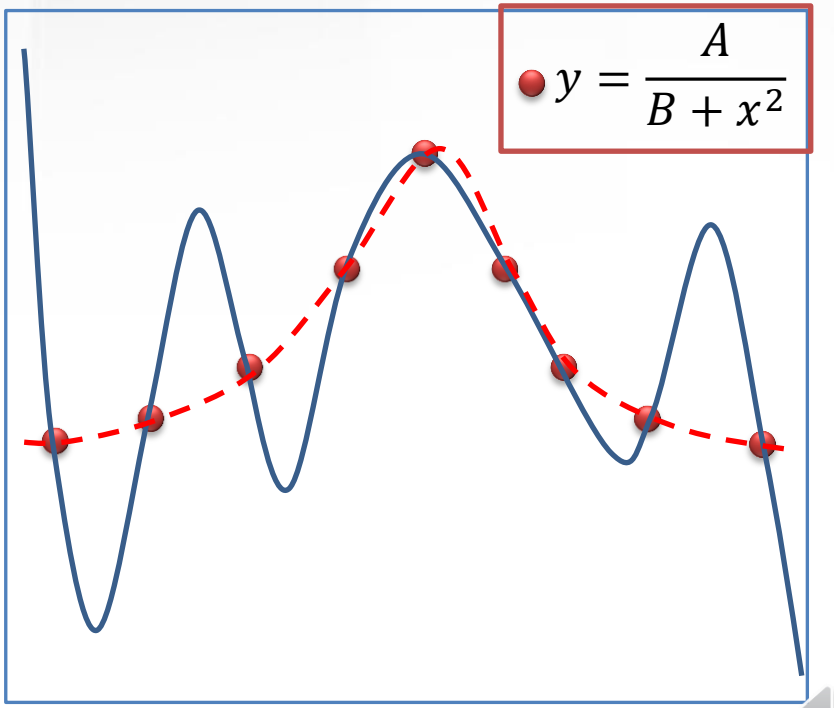
- исходная обучающая выборка разделяется на N частей и производится обучение по N - 1 части (без повторов) и оценка по оставшейся одной части;
- оценки усредняются и рассчитывается стандартное отклонение по выбранной метрике.

Наличие большого стандартного отклонения говорит о том, что данный набор факторов или модель плохо подходят для решения задачи. При малых значениях стандартного отклонения используются средние значения метрик.

Проблемы машинного обучения

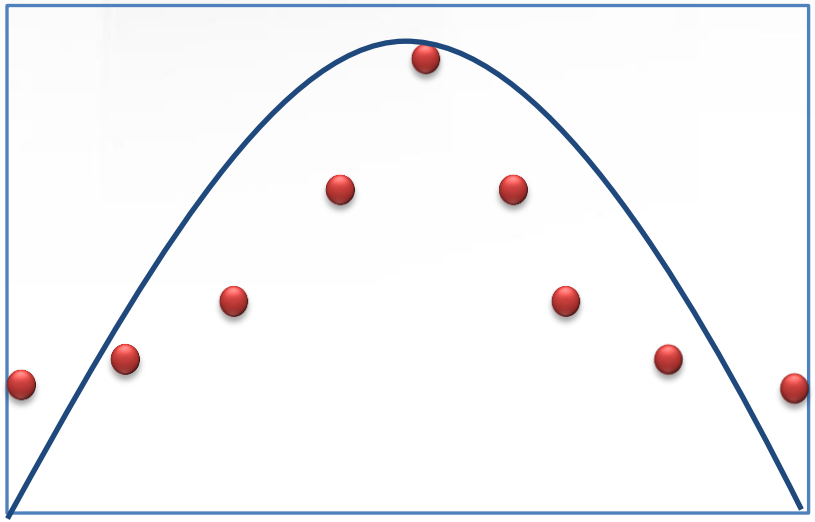
Модель обладает **обобщающей способностью**, когда вероятность появления ошибки на тестовой выборке мала и не сильно отличается от вероятности ошибки на обучающей выборке

Переобучение - явление, возникающее при обучении модели, когда вероятность появления ошибки на тестовой выборке существенно выше, чем на обучающей.



Проблемы машинного обучения

Недообучение — явление, возникающее в процессе обучения модели, когда вероятность ошибки на обучающей выборке достаточно высока.



Контрольные вопросы и задания

1. Чем отличаются друг от друга постановки следующих задач машинного обучения: обучение по прецедентам, обучение с учителем, обучение без учителя?
2. Приведите примеры ответов (откликов), которые можно отнести к булевым, номинальным, порядковым и количественным. Свяжите тип отклика с видом функции потерь.
3. Придумайте другие примеры функции потерь.
4. Имеется заданное множество людей с известными значениями их роста и охвата груди (или веса). Ставится задача разработать небольшое количество типовых размеров одежды для них. К какому типу задач машинного обучения относится сформулированная задача? Опишете формальную (математическую) постановку этой задачи.