Data Mining

1. Докладчики
2. Введение в Data Mining
3. Деревья решений
4. Метод ближайшего соседа

Вопросы?

Докладчики

- Александра Симонова, Мат-Мех, 5 курс

История Data Mining

- **1960-е гг. –** первая промышленная СУБД система IMS фирмы IBM.
- **1970-е гг. –** Conference on Data System Languages (CODASYL)
- **1980-е гг. –** SQL
- **1990-е гг. –** Data Mining

Возникновение Data Mining. Способствующие факторы

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи данных;
- накопление большого количества ретроспективных данных;
- совершенствование алгоритмов обработки информации.

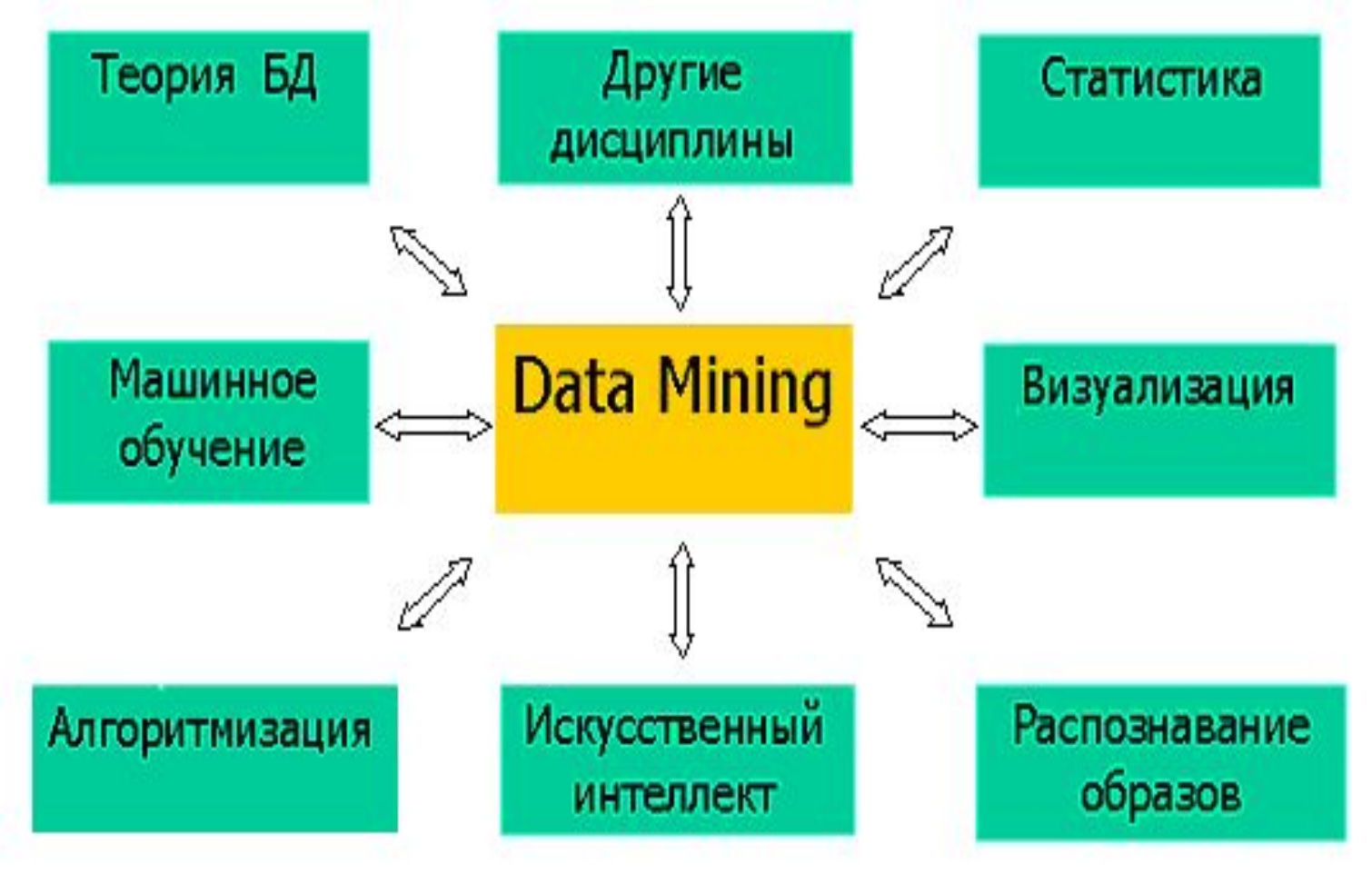
Понятие Data Mining

- ***Data Mining*** *- это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.*

Gregory Piatetsky-Shapiro

- Это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Мультидисциплинарность



Задачи Data Mining

- Классификация
- Кластеризация
- Прогнозирование
- Ассоциация
- Визуализация
- анализ и обнаружение отклонений
- Оценивание
- Анализ связей
- Подведение итогов

Стадии Data Mining

СВОБОДНЫЙ ПОИСК (в том числе ВАЛИДАЦИЯ)

ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

АНАЛИЗ ИСКЛЮЧЕНИЙ

Методы Data Mining. Технологические методы.

- Непосредственное использование данных, или сохранение данных:
  кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии (этот метод будет рассмотрен подробнее)
- Выявление и использование формализованных закономерностей, или дистилляция шаблонов:
  логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях

Методы Data Mining. Статистические методы.

- Дескриптивный анализ и описание исходных данных.
- Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
- Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
- Анализ временных рядов (динамические модели и прогнозирование).

Методы Data Mining. Кибернетические методы.

- Искусственные нейронные сети (распознавание, кластеризация, прогноз);
- Эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);
- Генетические алгоритмы (оптимизация);
- Ассоциативная память (поиск аналогов, прототипов);
- Нечеткая логика;
- Деревья решений; этот метод будет рассмотрен подробнее.
- Системы обработки экспертных знаний.

Визуализация инструментов Data Mining.

- Для деревьев решений - визуализатор дерева решений, список правил, таблица сопряженности.
- Для нейронных сетей - в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.
- Для карт Кохонена: карты входов, выходов, другие специфические карты.
- Для линейной регрессии - линия регрессии.
- Для кластеризации: дендрограммы, диаграммы рассеивания.

Проблемы и вопросы

- Data Mining не может заменить аналитика!
- Сложность разработки и эксплуатации приложения Data Mining. Основные аспекты:
  - Квалификация пользователя
  - Сложность подготовки данных
  - Большой процент ложных, недостоверных или бессмысленных результатов
  - Высокая стоимость
  - Наличие достаточного количества репрезентативных данных

Области применения Data mining

- ***Database marketers*** - Рыночная сегментация, идентификация целевых групп, построение профиля клиента
- ***Банковское дело*** - Анализ кредитных рисков, привлечение и удержание клиентов, управление ресурсами
- ***Кредитные компании*** - Детекция подлогов, формирование "типичного поведения" обладателя кредитки, анализ достоверности клиентских счетов , cross-selling программы
- ***Страховые компании*** - Привлечение и удержание клиентов, прогнозирование фингансовых показателей
- ***Розничная торговля*** - Анализ деятельности торговых точек, построение профиля покупателя, управление ресурсами
- ***Биржевые трейдеры*** - Выработка оптимальной торговой стратегии, контроль рисков

Области применения Data mining. Продолжение.

- ***Телекоммуникация и энергетика*** - Привлечение клиентов, ценовая политика, анализ отказов, предсказание пиковых нагрузок, прогнозирование поступления средств
- ***Налоговые службы и аудиторы*** - Детекция подлогов, прогнозирование поступлений в бюджет
- ***Фармацевтические компании*** - Предсказание результатов будущего тестирования препаратов, программы испытания
- ***Медицина*** - Диагностика, выбор лечебных воздействий, прогнозирование исхода хирургического вмешательства
- ***Управление производством*** - Контроль качества, материально-техническое обеспечение, оптимизация технологического процесса
- ***Ученые и инженеры*** - Построение эмпирических моделей, основанных на анализе данных, решение научно-технических задач

Перспективы технологии Data Mining.

- выделение типов предметных областей с соответствующими им эвристиками
- создание формальных языков и логических средств, с помощью которых будет формализованы рассуждения
- создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории, опирающиеся на эмпирические данные;
- преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

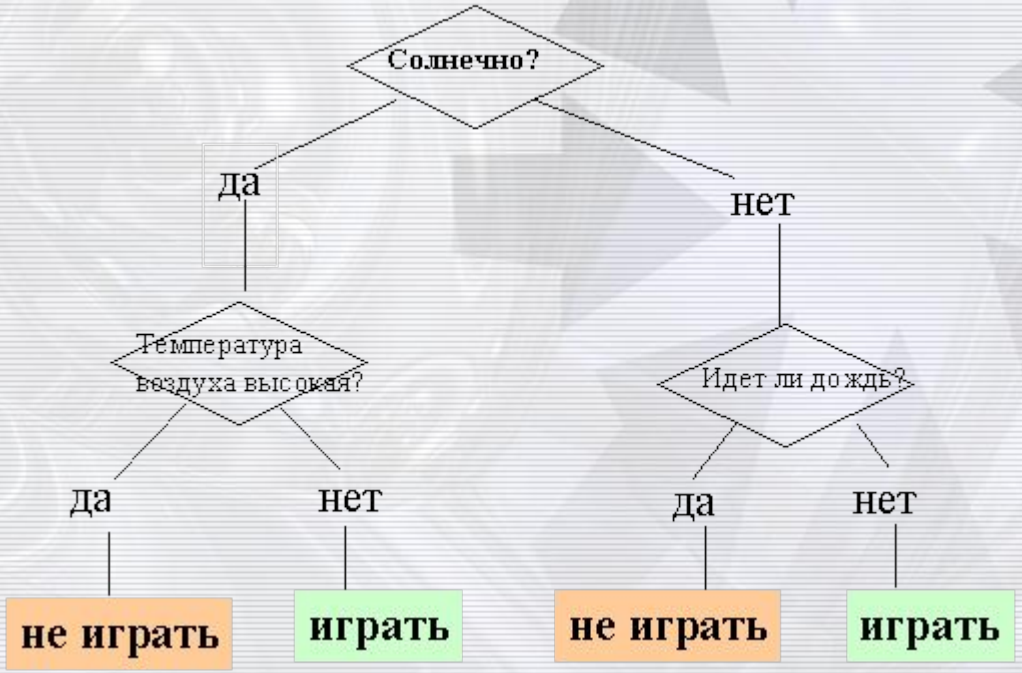
Литература по Data Mining

- "Wikipedia about Data Mining" (http://en.wikipedia.org/wiki/Data_mining)
- "Data Mining Tutorials" (http://www.eruditionhome.com/datamining/tut.html)
- "Thearling intro paper" (http://www.thearling.com/text/dmwhite/dmwhite.htm)
- "Что такое Data mining?“ (http://www.megaputer.ru/doc.php?classroom/whatis_dm/whatis_dm.html)
- "INTUIT.ru: Учебный курс - Data Mining“ (http://www.intuit.ru/department/database/datamining/)
- "Data Mining - подготовка исходных данных“ (http://www.basegroup.ru/tasks/datamining_prepare.htm)

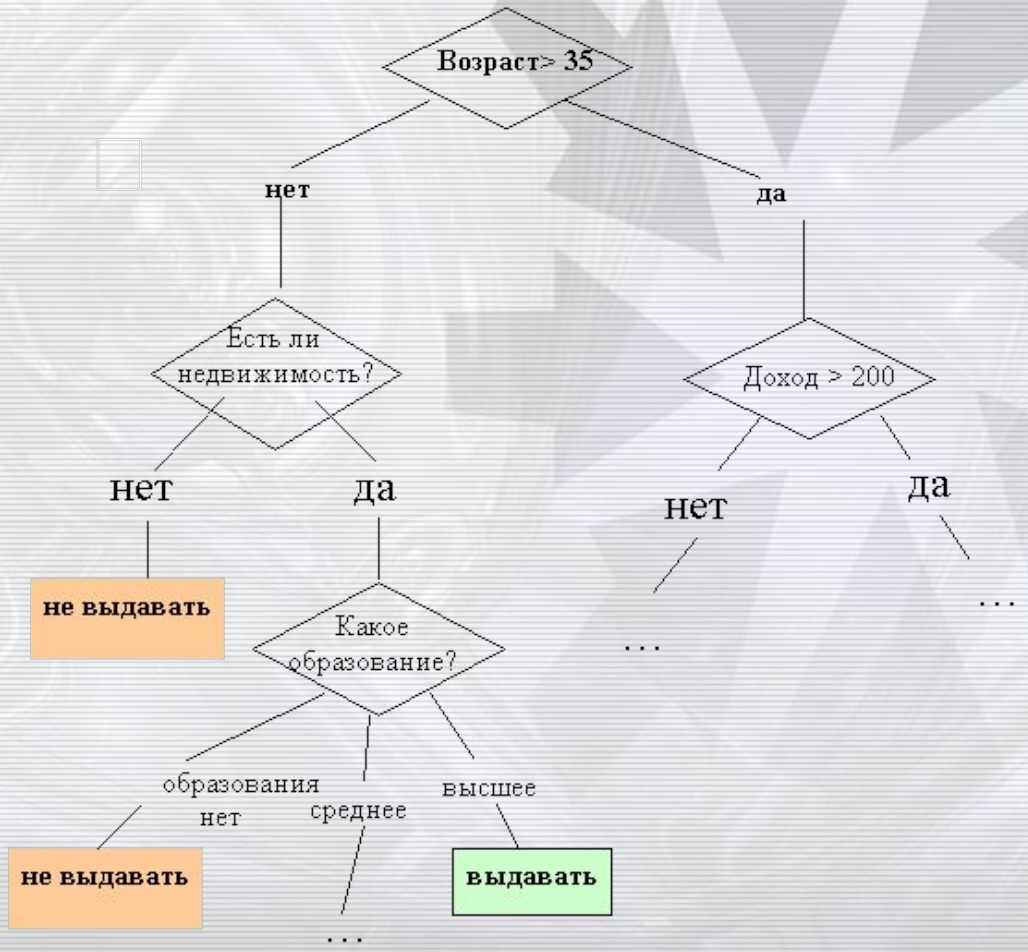
Деревья решений. История и основные понятия.

- Возникновение - 50-е годы (Ховиленд и Хант (Hoveland, Hunt))
- Метод также называют деревьями решающих правил, деревьями классификации и регрессии
- *Это способ представления правил в иерархической, последовательной структуре*

Деревья решений. Пример 1.



Деревья решений. Пример 2.



Деревья решений. Преимущества метода.

- **Интуитивность деревьев решений**
- Возможность извлекать правила из базы данных на **естественном языке**
- **Не требует от пользователя выбора входных атрибутов**
- **Точность** моделей
- Разработан ряд **масштабируемых алгоритмов**
- **Быстрый процесс обучения**
- Обработка **пропущенных значений**
- Работа и с числовыми, и с **категориальными** типами данных

Деревья решений. Процесс конструирования.

Основные этапы алгоритмов конструирования деревьев:

- "построение" или "создание" дерева (tree building)
- "сокращение" дерева (tree pruning).

Деревья решений. Критерии расщепления.

- "мера информационного выигрыша" (information gain measure)
- индекс Gini, т.е. gini(T), определяется по формуле:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

- Большое дерево не означает, что оно "подходящее"

Деревья решений. Остановка построения дерева.

Остановка - такой момент в процессе построения дерева, когда следует прекратить дальнейшие ветвления.

Варианты остановки:

- "ранняя остановка" (prepruning)
- ограничение глубины дерева
- задание минимального количества примеров

Деревья решений. Сокращение дерева или отсечение ветвей.

Критерии:

- Точность распознавания
- Ошибка

Деревья решений. Алгоритмы. CART .

- CART (Classification and Regression Tree)
- разработан в 1974-1984 годах четырьмя профессорами статистики - Leo Breiman (Berkeley), Jerry Friedman (Stanford), Charles Stone (Berkeley) и Richard Olshen (Stanford)
- CART предназначен для построения бинарного дерева решений.

Особенности:

- функция оценки качества разбиения;
- механизм отсечения дерева;
- алгоритм обработки пропущенных значений;
- построение деревьев регрессии.

Деревья решений. Алгоритмы. C4.5 .

- Строит дерево решений с неограниченным количеством ветвей у узла
- Дискретные значения => только классификация
- Каждая запись набора данных ассоциирована с одним из предопределенных классов => один из атрибутов набора данных должен являться меткой класса.
- Количество классов должно быть значительно меньше количества записей в исследуемом наборе данных.

Деревья решений. Перспективы метода и выводы.

- Разработка новых масштабируемых алгоритмов (Sprint, предложенный Джоном Шафером)
- Метод деревьев - иерархическое, гибкое средство предсказания принадлежности объектов к определенному классу или прогнозирования значений числовых переменных.
- Качество работы зависит как от выбора алгоритма, так и от набора исследуемых данных.
- Чтобы построить качественную модель, необходимо понимать природу взаимосвязи между зависимыми и независимыми переменными и подготовить достаточный набор данных.

Метод "ближайшего соседа" или системы рассуждений на основе аналогичных случаев.

Прецедент - это описание ситуации в сочетании с подробным указанием действий, предпринимаемых в данной ситуации.
Этапы:

- сбор подробной информации о поставленной задаче;
- сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;
- выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов;
- адаптация выбранного решения к текущей проблеме, если это необходимо;
- проверка корректности каждого вновь полученного решения;
- занесение детальной информации о новом прецеденте в базу прецедентов.

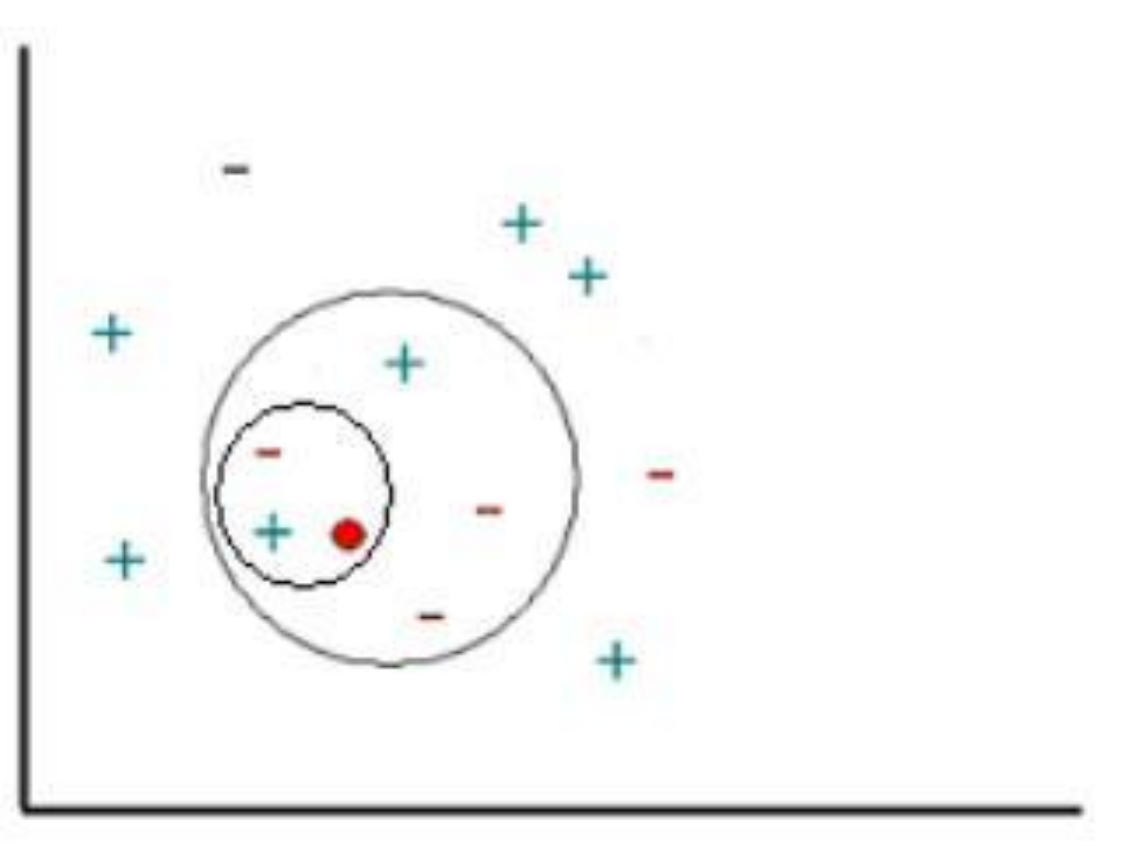
Метод "ближайшего соседа". Преимущества.

- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных.

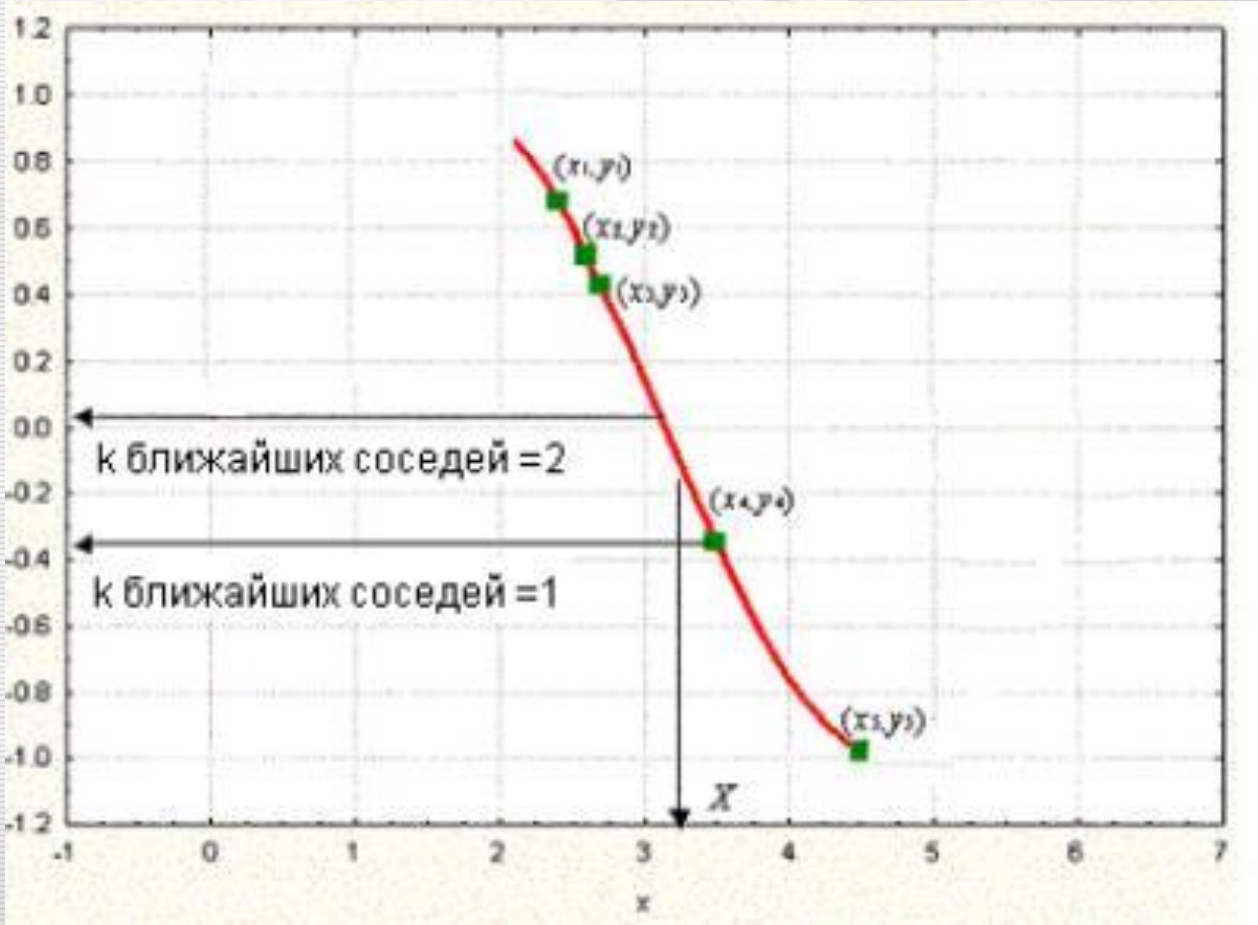
Метод "ближайшего соседа". Недостатки.

- Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт
- Сложность выбора меры "близости" (метрики).
- Высокая зависимость результатов классификации от выбранной метрики.
- Необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Метод "ближайшего соседа". Решение задачи классификации новых объектов.



Метод "ближайшего соседа". Решение задачи прогнозирования.



Метод "ближайшего соседа". Оценка параметра k методом кросс-проверки.

- *Кросс-проверка* - известный метод получения оценок неизвестных параметров модели.
- Основная идея - разделение выборки данных на v "складок". V "складки" здесь суть случайным образом выделенные изолированные подвыборки.

Метод "ближайшего соседа". Примеры использования и реализации.

- Использование - программное обеспечение центра технической поддержки компании Dell, разработанное компанией Inference.
- Реализация - CBR Express и Case Point (Inference Corp.), Apriori (Answer Systems), DP Umbrella (VYCOR Corp.), KATE tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США), а также некоторые статистические пакеты, например, Statistica.

?

- Вопросы??