Лекция 9, 10

Интеллектуальный анализ данных 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)

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Неочевидных – это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных – это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных – это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

Знания – совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д.

Использование знаний (knowledge deployment) означает действительное применение найденных знаний для достижения конкретных преимуществ (например, в конкурентной борьбе за рынок).

Интеллектуальный анализ данных - процесс обнаружения пригодных к использованию сведений в крупных наборах данных.

В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных.

Обычно такие закономерности нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных.

Эти закономерности и тренды можно собрать вместе и определить как *модель интеллектуального анализа данных*.

Модели интеллектуального анализа данных могут применяться к конкретным сценариям, а именно:

- **Прогноз**: оценка продаж, прогнозирование нагрузки сервера или времени простоя сервера

- **Риски и вероятности**: выбор наиболее подходящих заказчиков для целевой рассылки, определение точки равновесия для рискованных сценариев, назначение вероятностей диагнозам или другим результатам

- **Рекомендации**: определение продуктов, которые с высокой долей вероятности могут быть проданы вместе, создание рекомендаций

- **Определение последовательностей**: анализ выбора заказчиков во время совершения покупок, прогнозирование следующего возможного события

- **Группирование**: разделение заказчиков или событий на группы связанных элементов, анализ и прогнозирование общих черт

Модели интеллектуального анализа данных могут применяться к конкретным сценариям, а именно:

- **Прогноз**
- **Риски и вероятности**
- **Рекомендации**
- **Определение последовательностей**
- **Группирование**

К методам Data Mining иногда относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов).

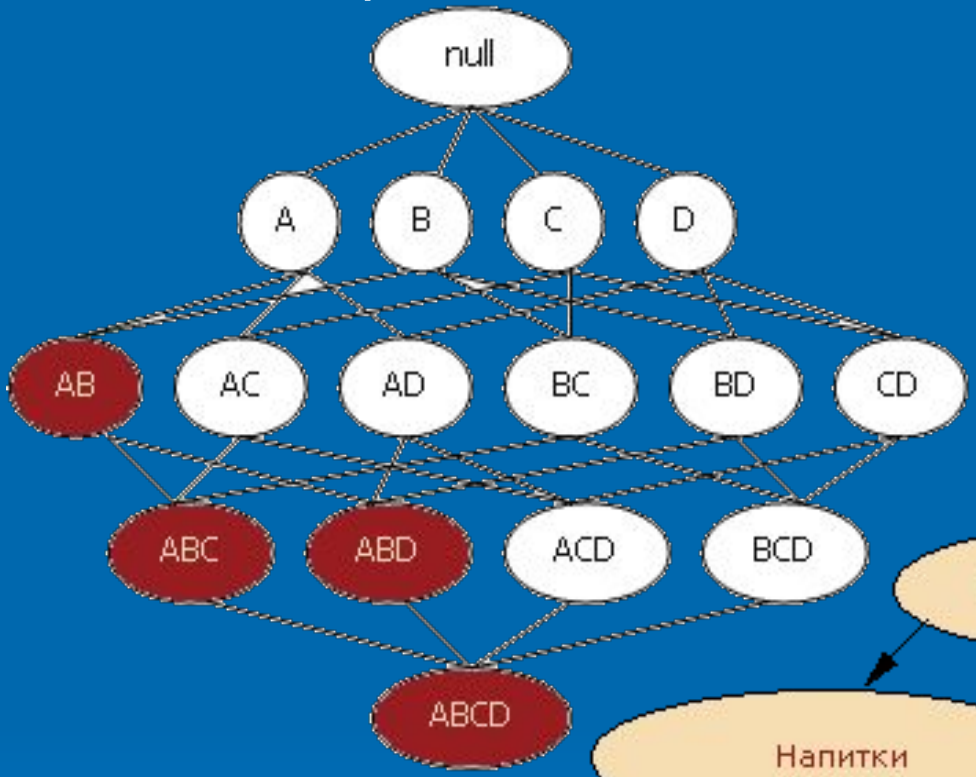
Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями Data Mining (обнаружение ранее неизвестных нетривиальных и практически полезных знаний).

Одно из важнейших назначений методов Data Mining состоит в наглядном представлении результатов вычислений, что позволяет использовать инструментарий Data Mining людьми, не имеющих специальной математической подготовки. В то же время, применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой.

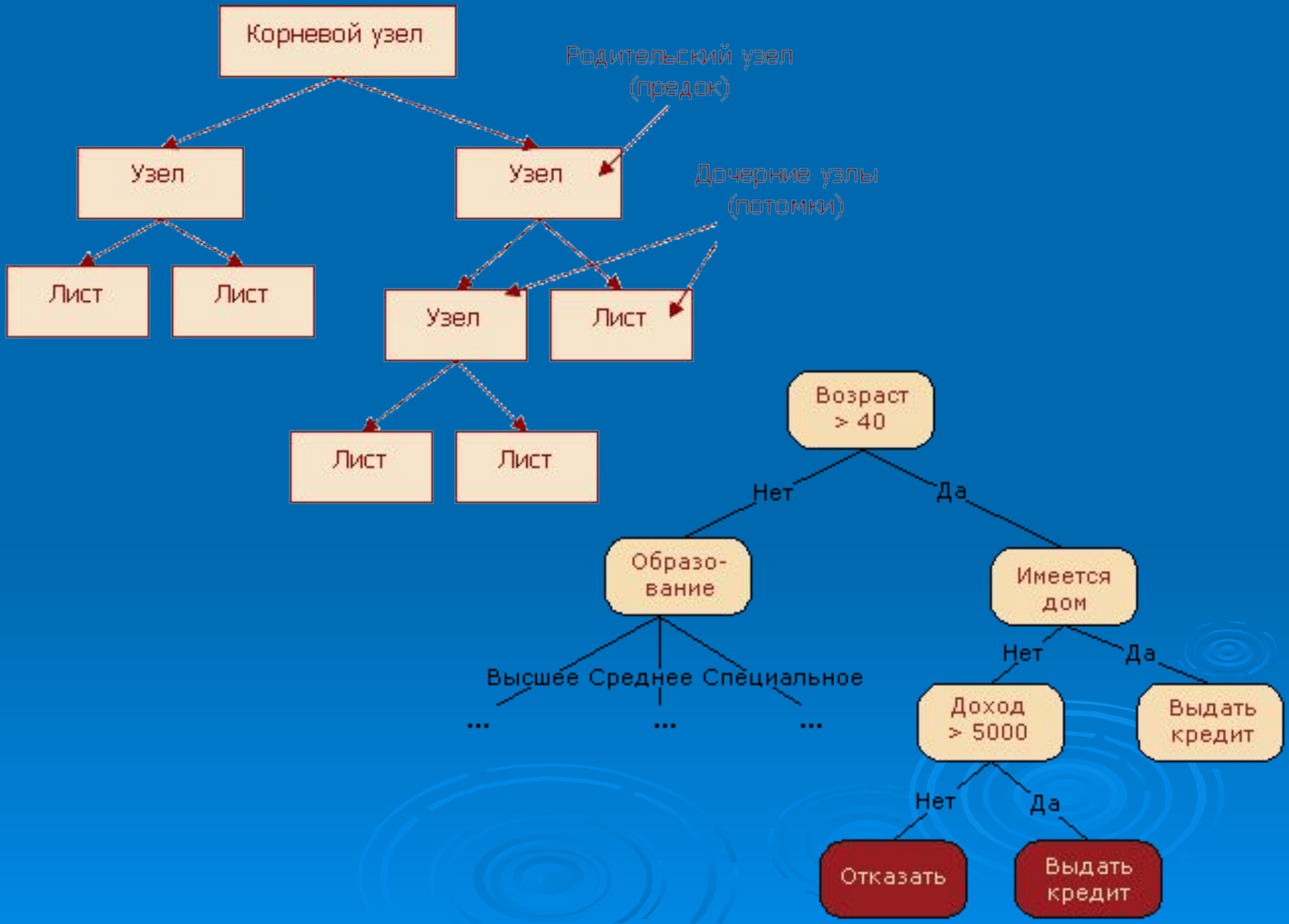
Знания, добываемые методами Data mining, принято представлять в виде моделей:

- Ассоциативные правила – логические закономерности;
- Деревья решений – средства ППР для прогнозных моделей;
- Кластеры – объединение в группы схожих объектов;
- Математические функции.

Ассоциативные правила – логические закономерности;

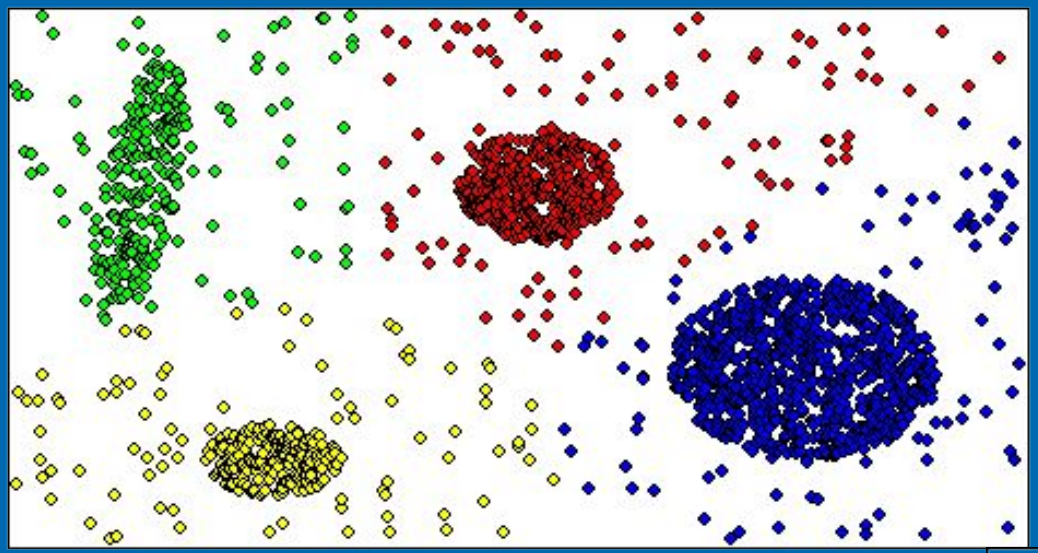


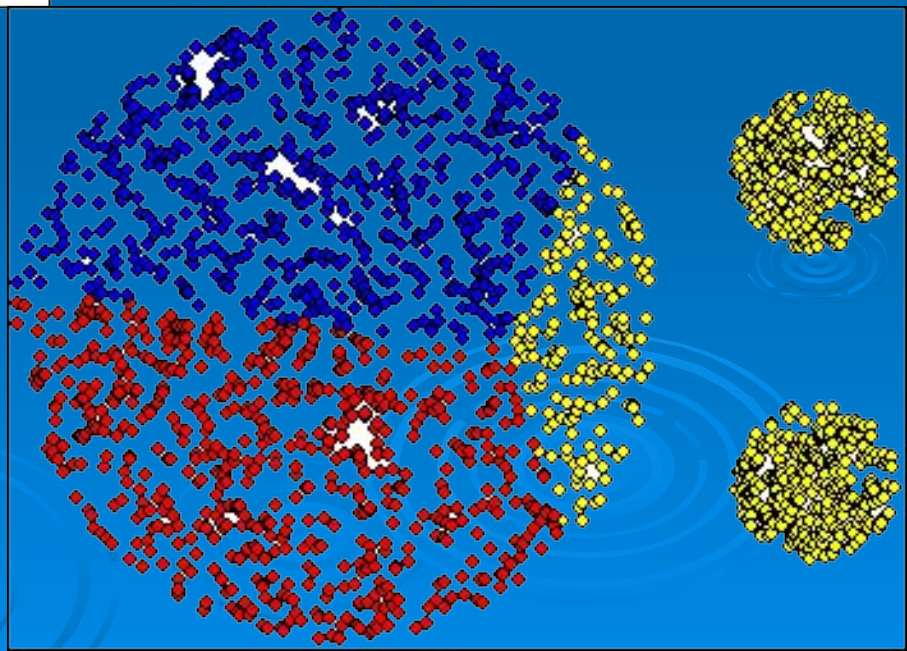
Деревья решений;



Деревья решений - средство поддержки принятия решений - средство поддержки принятия решений, использующееся в статистике - средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей ;

Кластеры - объединение в группы схожих объектов ;





К методам и алгоритмам Data Mining относятся:

1. Искусственные нейронные сети
2. Деревья решений, символьные правила
3. Методы ближайшего соседа и k-ближайшего соседа
4. Метод опорных векторов
5. Байесовские сети
6. Линейная регрессия
7. Корреляционно-регрессионный анализ
8. Иерархические методы кластерного анализа
9. Неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы
10. Методы поиска ассоциативных правил, в том числе алгоритм Apriori
11. Метод ограниченного перебора
12. Эволюционное программирование и генетические алгоритмы

А также методы визуализации данных и др. методы.

Большинство аналитических методов, используемые в технологии Data Mining – это известные математические алгоритмы и методы.

Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств.

Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта.

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Свойства методов Data Mining

Различные методы Data Mining характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных.

Основные свойства и характеристики методов Data Mining:

- точность,
- масштабируемость,
- интерпретируемость,
- проверяемость,
- трудоемкость,
- гибкость,
- быстрота
- популярность.

Свойства методов Data Mining

Масштабируемость – свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и пр., при добавлении к ней вычислительных ресурсов.

В таблице приведена сравнительная характеристика некоторых распространенных методов.

Оценка каждой из характеристик проведена следующими категориями, в порядке возрастания:

чрезвычайно низкая,
очень низкая,
низкая/нейтральная,
нейтральная/низкая,
нейтральная,
нейтральная/высокая,
высокая,
очень высокая.

Свойства методов Data Mining

АЛГОРИТМ	ТОЧНОСТЬ	МАСШТАБИРУЕМОСТЬ	ИНТЕРПРЕТИРУЕМОСТЬ	ПРИГОДНОСТЬ К КИСП
линейная регрессия	нейтральная	высокая	высокая / нейтральная	высокая
нейронные сети	высокая	низкая	низкая	низкая
методы визуализации	высокая	очень низкая	высокая	высокая
деревья решений	низкая	высокая	высокая	высокая / нейтральная
нейронные сети	высокая	нейтральная	низкая	высокая / нейтральная
k-ближайшего соседа	низкая	очень низкая	высокая / нейтральная	нейтральная

АЛГОРИТМ	ТРУДОЕМКОСТЬ	РАЗНОСТОРОННОСТЬ	БЫСТРОТА	ПОПУЛЯРНОСТЬ
линейная регрессия	нейтральная	нейтральная	высокая	низкая
нейронные сети	нейтральная	низкая	очень низкая	низкая
методы визуализации	очень высокая	низкая	чрезвычайно низкая	высокая / нейтральная
деревья решений	высокая	высокая	высокая / нейтральная	высокая / нейтральная
нейронные сети	низкая / нейтральная	нейтральная	низкая / нейтральная	нейтральная
k-ближайшего соседа	нейтральная низкая	низкая	высокая	низкая

Таблица 1 Сравнительная характеристика методов Data Mining

Свойства методов Data Mining

	Точность	Масштаби-руемость	Интерпре-тируемость	Трудоем-кость	Быстрота	Популяр-ность
Линейная регрессия	3	5	4	3	5	2
Нейронные сети	5	2	2	3	1	2
Деревья решений	2	5	5	5	4	4
Полином. нейронные сети	5	3	2	2	2	3
К-ый ближайший сосед	2	1	5	2	5	2

Как видно из рассмотренной таблицы, каждый из методов имеет свои сильные и слабые стороны. Но ни один метод, какой бы не была его оценка с точки зрения присущих ему характеристик, не может обеспечить решение всего спектра задач Data Mining.

Методы Data Mining.

1. Технологические методы.
2. Статистические методы.
3. Кибернетические методы.

Методы Data Mining.

1.Технологические методы.

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

Методы Data Mining.

2. 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.

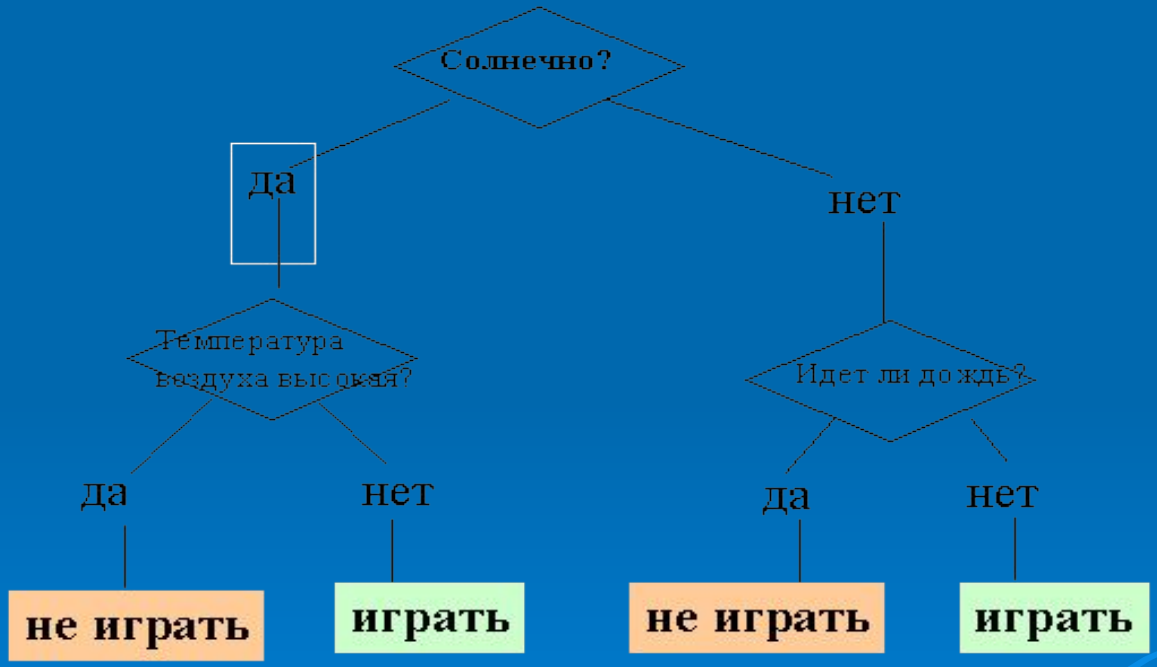
3. Кибернетические методы.

- 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;
- Системы обработки экспертных знаний.

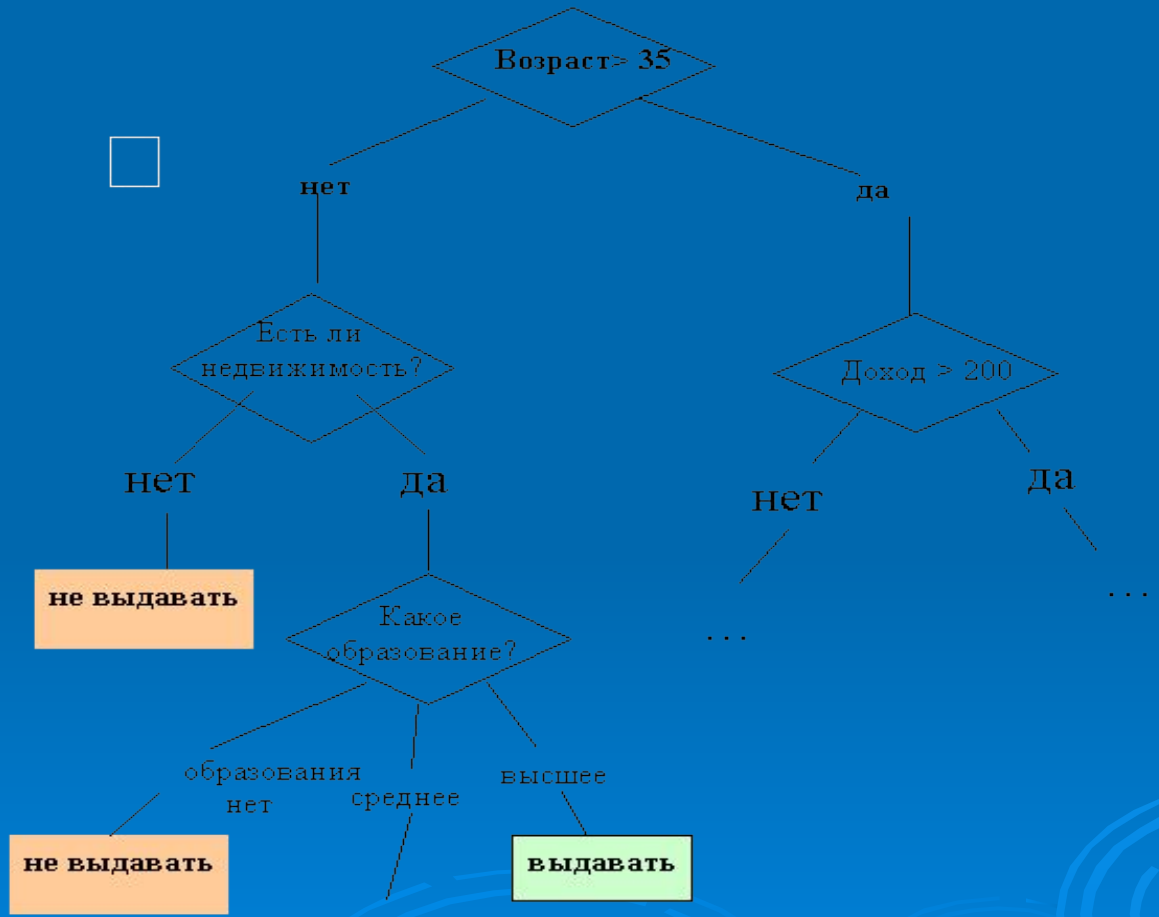
Деревья решений

- 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

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.

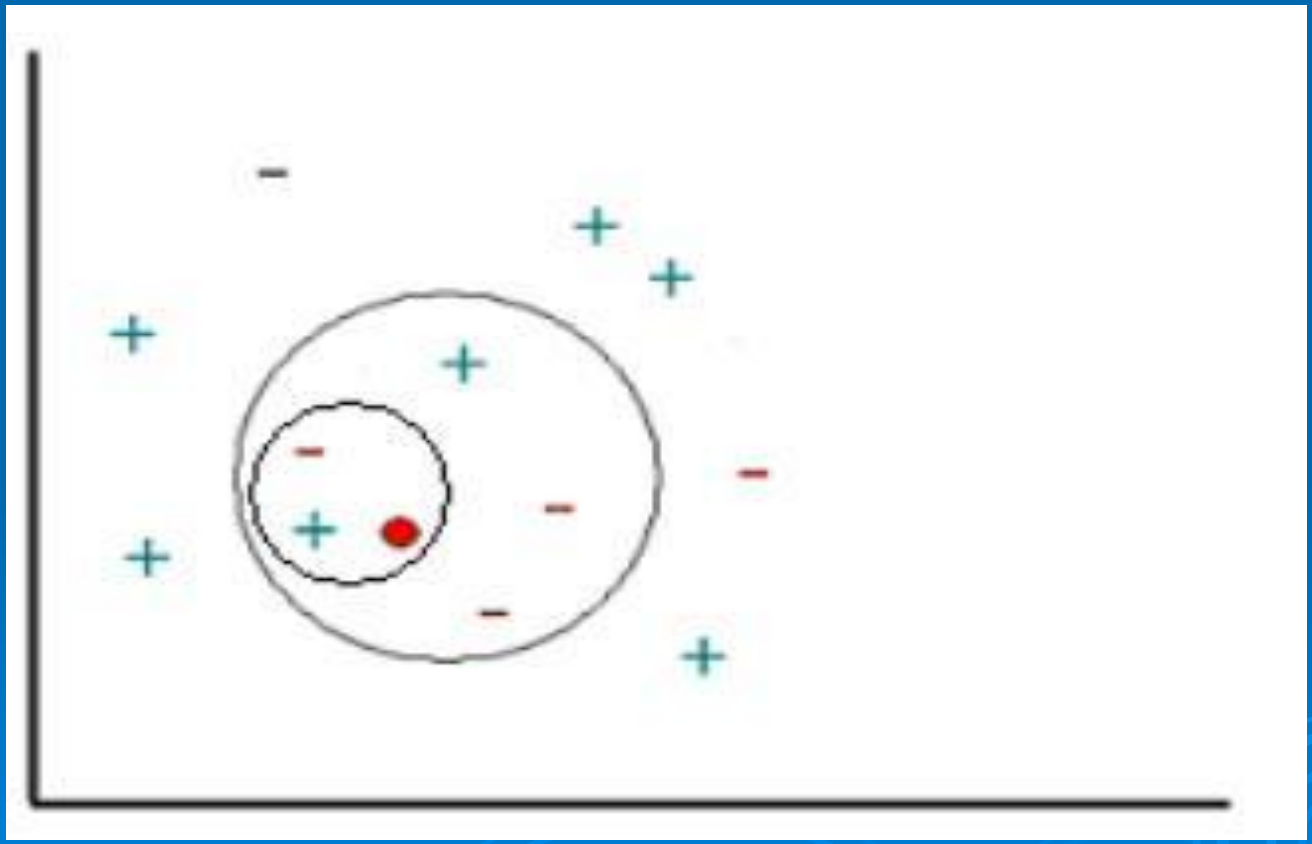
Метод "ближайшего соседа". Преимущества.

- 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.

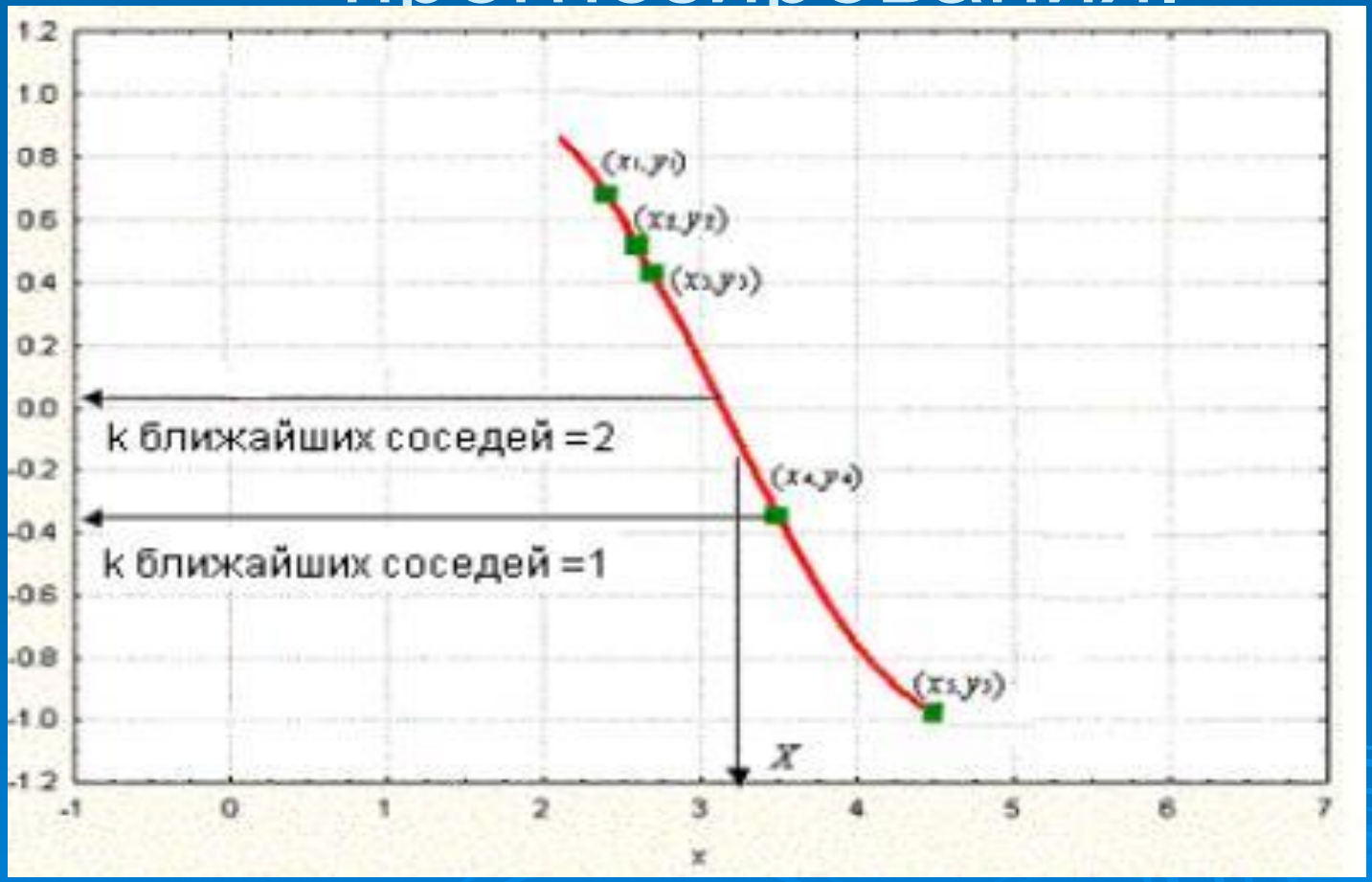
Метод "ближайшего соседа". Недостатки.

- 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.