Факультет вычислительной
математики и кибернетики

Основы машинного обучения (МФК)

Анализ данных: основные понятия и задачи

А.В. Якушин, к.п.н., доцент, ВМК МГУ

Факультет ВМК МГУ имени М.В. Ломоносова

Основные задачи анализа данных

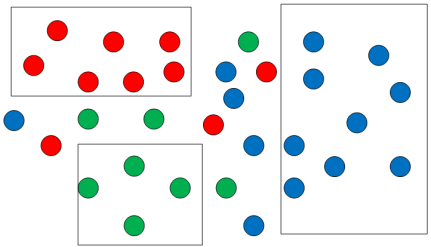
Основные задачи анализа данных

Если подойти с общей, абстрактной позиции, то в области анализа данных можно выделить следующие задачи:

- Извлечение знаний (структуризация)
- Классификация
- Идентификация
- Кластеризация
- Поиск
- Прогнозирование

Основные задачи анализа данных

Извлечение знаний (структуризация)

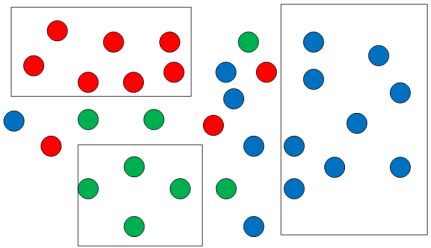


Исторически эта задача возникла при исследовании взаимозависимостей между косвенными показателями одного и того же явления. Требуется построить алгоритм, генерирующий набор объективных закономерностей между признаками,имеющих место в генеральной совокупности.
Стандартно в процессе извлечения знаний выделяют 5 основных задач:

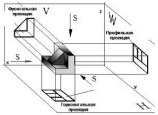
1. Извлечение именованных сущностей
2. Распознавание ссылок на релевантные сущности для устранения избыточности
3. Извлечение свойств сущностей
4. Извлечение отношений между именованными сущностями
5. Получение контекста ситуации

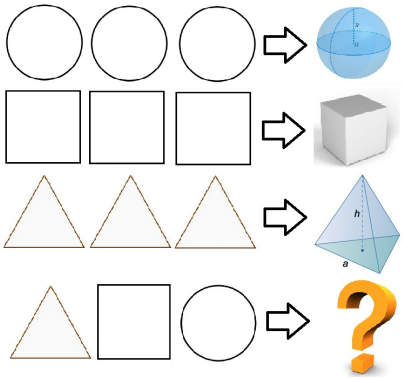
Основные задачи анализа данных

Примеры задач извлечения знаний



1. Медицина: поиск взаимосвязей (синдромов) между различными показателями при фиксированной болезни
2. Социология: определение факторов, влияющих на победу на выборах
3. Генная инженерия: выявление связанных участков генома
4. Научные исследования: получение новых знаний об исследуемом процессе
5. Биржевое дело: определение закономерностей между различными биржевыми показателями





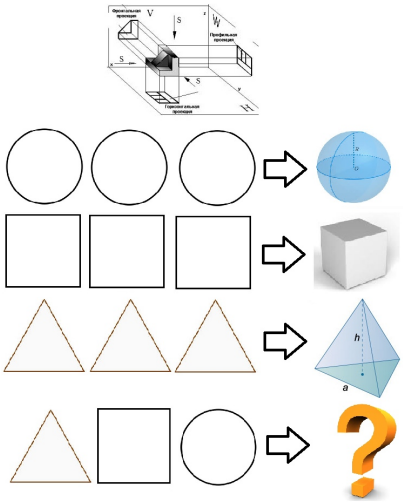
Классифицировать объект — значит, указать номер (или наименование класса), к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Классификация бывает **бинарная** (два класса, да или нет, черный или белый и пр.) и **многоклассовая**.

Основные задачи анализа данных

Примеры задач классификации



1. Медицинская диагностика: по набору медицинских характеристик требуется поставить диагноз
2. Геологоразведка: по данным зондирования почв определить наличие полезных ископаемых
3. Оптическое распознавание текстов: по отсканированному изображению текста определить цепочку символов, его формирующих
4. Кредитный скоринг: по анкете заемщика принять решение о выдаче/отказе кредита
5. Синтез химических соединений: по параметрам химических элементов спрогнозировать свойства получаемого соединения

Один раджа приехал в гости к падишаху из соседнего государства и привез множество подарков, и среди них — живого слона. А так как падишах никогда в жизни не видел это животное, гость решил загадать загадку. Заведя слона в темный зал, раджа попросил правителя: «Пусть твои советники опишут слона, который здесь находится. Хочу узнать, насколько они мудры. Первый советник в кромешной тьме наткнулся на ногу животного. «Этот зверь, как громадное дерево», — сказал он. «Слон — это огромная извивающаяся змея», — возразил ему второй мудрец, который нащупал хобот животного. Третьему удалось погладить туловище слона. «О правитель, они оба лгут, — закричал он. — Слон похож на большой лист бумаги — такой же широкий и шершавый».

Недоумевающий падишах обратился к радже: «Скажи, каков этот слон?» Когда гость вывел животное из зала, и мудрецы, и правитель сильно удивились. «Каждый из вас был прав по-своему. И все же все вы ошибались, — обратился раджа к советникам. — Ваши знания были неполными, ведь вы познали лишь часть целого. А потому слон оказался совсем не таким, каким вы себе его представляли».

Основные задачи анализа данных

Идентификация

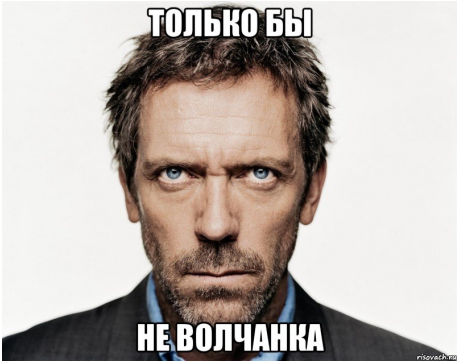
Из группы Изоизоляция
Фото: Дарья Селенская

Исторически возникла из классификации, необходимости отделить объекты, обладающие определенным свойством, от «всего остального». Идентификация –это процесс распознавания элемента системы, обычно с помощью заранее определенного идентификатора или другой уникальной информации.

Особенностью задачи является то, что все объекты принадлежат одному классу, причем не существует возможности сделать репрезентативную выборку из класса «все остальное». Требуется построить алгоритм (идентификатор), который по вектору признаков x определил бы наличие свойства A у объекта x, либо вернул оценку степени его выраженности.

Основные задачи анализа данных

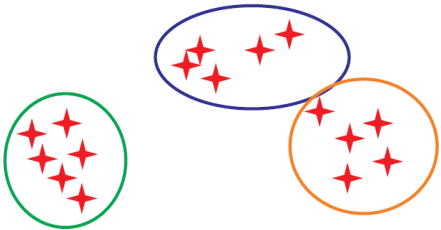
Примеры задач идентификации



1. Медицинская диагностика: по набору медицинских характеристик требуется установить наличие/отсутствие конкретного заболевания
2. Системы безопасности: по камерам наблюдения в подъезде идентифицировать жильца дома
3. Банковское дело: определить подлинность подписи на чеке
4. Обработка изображений: выделить участки с изображениями лиц на фотографии
5. Искусствоведение: по характеристикам произведения (картины, музыки, текста) определить, является ли его автором тот или иной автор

Основные задачи анализа данных

Кластеризация

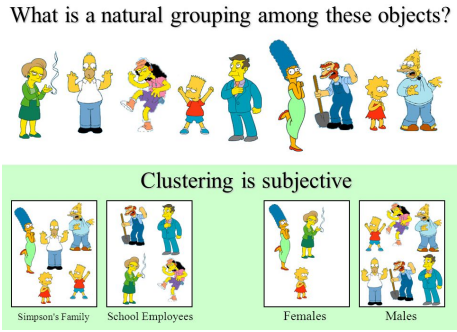


Кластерный анализ (Data clustering) — задача разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Задача кластеризации (то же самое, что кластерный анализ) решена, если получено разбиение множества объектов на подмножества, называемых кластерами. Основное условие решения: объекты, принадлежащие одному кластеру должны быть больше похожи друг на друга, чем объекты из других кластеров. Критерий «похожести» может быть один, но их может быть и несколько.

Основные задачи анализа данных

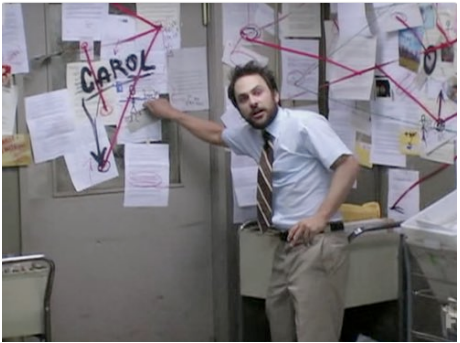
Примеры задач кластеризации



1. Экономическая география: по физико-географическим и экономическим показателям разбить страны мира на группы схожих по экономическому положению государств
2. Финансовая сфера: по сводкам банковских операций выявить группы «подозрительных», нетипичных банков, сгуппировать остальные по степени близости проводимой стратегии
3. Маркетинг: по результатам маркетинговых исследований среди множества потребителей выделить характерные группы по степени интереса к продвигаемому продукту
4. Социология: по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества, а также характерные фокус-группы населения

Основные задачи анализа данных

Поиск



Поиск информации представляет собой процесс выявления в некотором множестве документов (текстов) всех тех, которые посвящены указанной теме (предмету), удовлетворяют заранее определенному условию поиска (запросу) или содержат необходимые (соответствующие информационной потребности) факты, сведения, данные.

Процесс поиска включает последовательность операций, направленных на сбор, обработку и предоставление информации.

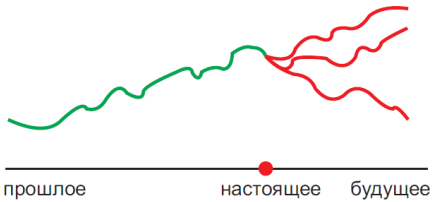
Основные задачи анализа данных

Примеры задач поиска

1. Документоведение: выборка из массива документов информации соответствующей определенным требованиям
2. Финансовая сфера: подбор инструментов для инвестиций на основе системы критериев
3. Медицина: поиск аналогичных случаев в медицинских базах данных
4. Социология: выборка из социальных сетей постов с определенными характеристиками

Основные задачи анализа данных

Прогнозирование



Исторически возникла при исследовании временных рядов и попытке предсказания их значений через какой-то промежуток времени.

Прогнозирование (от греческого Prognosis), в широком понимании этого слова, определяется как опережающее отражение будущего. Целью прогнозирования является предсказание будущих событий. Прогнозирование (forecasting) является одним из ключевых моментов при принятии решений.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных.

Основные задачи анализа данных

Примеры задач прогнозирования

1. Биржевое дело: прогнозирование биржевых индексов и котировок
2. Системы управления: прогноз показателей работы реактора по данным телеметрии
3. Экономика: прогноз цен на недвижимость
4. Демография: прогноз изменения численности различных социальных групп в конкретном ареале
5. Гидрометеорология: прогноз геомагнитной активности