# Глава 0. Введение. Общие сведения об интеллектуальном анализе данных

к.ф.-м.н. А.И. Майсурадзе
кафедра математических методов прогнозирования

ноябрь 2008

# Оглавление лекции

# Три направления решения интеллектуальных задач

**Интеллектуальная задача**

Задача обработки информации, возникающая в плохо формализованной прикладной области. Адекватные математические модели реальных объектов отсутствуют.

- Строгое математическое моделирование прикладной области
  Отдельные успехи: Нобелевская премия акад. Канторовича.
  Обычно создание адекватной модели невозможно.
- Моделирование процесса мышления
  Человек справляется — пытаемся моделировать его способ решения. Персептроны. Экспертные системы.
- Реализуем процесс преобразования информации
  Несмотря на отсутствие модели того, как решает человек, несмотря на отсутствие адекватной математической модели реальной ситуации, опираясь на обычный здравый смысл.

# Понятие об интеллектуальном анализе данных

## Основная идея

Моделирование данных, а не явлений.

- Физика — моделирование явления (пример классического моделирования)
  Основной задачей физики является построение математической модели некоторой системы или явления. Модель позволяет прогнозировать развитие явления или управлять системой.
- ИАД — моделирование данных (неклассическое моделирование)
  Если мы не можем создать физическую модель явления, можно попытаться моделировать данные.

# Эвристические информационные модели

**Определение**

*Эвристическая информационная модель — это модель, описывающая данные. Она не опирается на законы природы, хотя и может базироваться на некоторых разумных предпосылках.*

Формализация — параметрические семейства алгоритмов.

# Эвристические информационные модели

**Определение**

*Эвристическая информационная модель — это модель, описывающая данные. Она не опирается на законы природы, хотя и может базироваться на некоторых разумных предпосылках.*

Формализация — параметрические семейства алгоритмов.

- линейный классификатор
- байесовские модели
- нейронный сети
- метрические модели
  - ближайшие соседи
  - парзеновские окна
  - потенциальные функции
- логические модели
  - решающие списки
  - решающие деревья
  - тестовый подход
- ABO
- ... многое другое ...

# Фундаментальные задачи ИАД

При решении прикладных задач методами ИАД принято делить общую задачу на несколько подзадач (провести декомпозицию), каждая из которых уже известна и изучена.

Исторически сложился некоторый набор таких подзадач, на которые удобно проводить декомпозицю. Именно такие задачи называют фундаментальными задачами ИАД.

# Задачи обучения с учителем

**Supervised learning**

- Задачи классификации
  Надо получить алгоритм (классификатор), который каждый объект распознавания относит к некоторым классам из конечного, заранее заданного набора классов
- Задачи восстановления регрессии
  Надо получить алгоритм (регрессию), который каждому объекту распознавания сопоставляет некоторое значение из бесконечного (непрерывного) множества
- Задачи обучения по прецедентам
  Классификатор или регрессия настраиваются по заданному конечному набору прецедентов - объектов с заранее известными правильными ответами.

# Задачи обучения с учителем

- Задачи прогнозирования
  Обычно прогнозирование сводится к классификации или восстановлению регрессии, когда один из признаков определяет время.
- Задачи последовательного обучения
  Прецеденты приходят последовательно во времени (один за другим). Алгоритм постоянно донастраивается.
- Архивирование, настройка модели источника
  Символы на архивацию приходят последовательно один за другим.

# Задачи обучения без учителя

**Unsupervised learning**

- Кластеризация (Сегментация)
  Надо разбить все множество объектов на непересекающиеся подмножества (кластеры, сегменты), в которых объекты в каком-либо смысле похожи друг на друга.
- Нечеткая кластеризация, бикластеризация и т.д.
- Иерархическая кластеризация (Таксономия)
  Надо построить дерево подмножеств, в котором каждый последующий слой является измельчением предыдущего.

# Задачи с частичным обучением

**Semisupervised learning**

- Задачи с частичным обучением
  Кроме прецедентой информации имеется информация о том, что некоторый набор объектов действительно существует и будет использован в ходе решения прикладной задачи. То есть для настройки алгоритма можно использовать прецедентную информацию и информацию о существовании данных объектов. Пример: база данных фотографий, ищем фотографии с лицами.

# Выявление отклонений, детектирование

- Выявление ошибок в данных: так не может быть
  Поступающая информация может содержать ошибки.
  Источником возникновения ошибок может служить, например, неисправность измерительного прибора.
- Выявление нетипичного поведения: так раньше не было
  Наша атомная электростанция раньше никогда не взрывалась.
  Мы не знаем, как выглядит станция, собирающаяся взорваться, но систему мониторинга создать должны.
- Устранение отклонений из обучения (фильтрация)
  Необходимо выявить те прецедентные данные, которые мешают качественно настроить модель.

# Задачи восстановления пропусков

- Заполнение пропусков в прецедентах
  Выбранный метод обучения модели требует, чтобы присутствовали все данные без пропусков.
- Заполнение пропусков в описаниях распознаваемых объектов
  Настроенная модель требует, чтобы во вновь приходящих на обработку описаниях не было пропусков.

Решение обычно сводится к задачам классификации или восстановления регрессии.

# Различные виды представления исходных данных

- **Признаковое описание** объектов
  Фиксированный набор признаков. Каждому объекту ставится в соответствие вектор значений признаков.
- Плоские таблицы, кросс-таблицы
- **Описание эталонами**
  Одна или несколько функций расстояния. Каждому объекту ставится в соответствие набор всех расстояний до всех заданных эталонов.
- **Метрическое описание** пар объектов
  Фиксированный набор функций расстояния. Каждой паре объектов ставится в соответствие вектор значений расстояний.
- **Транзакционные данные, Формальные контексты**
  Фиксировано множество элементов. Транзакция — конечное подмножество элементов.
- Наличие элемента в транзакции — бинарный признак

# Фундаментальные задачи ИАД (продолжение)

**Анализ наборов** (не учитываем время транзакции)
Термин: анализ рыночной корзины

- Поиск популярных наборов
  Популярный набор: чай и мёд часто покупают вместе
- Поиск ассоциативных правил
  Ассоциативное правило: те, кто купил мёд, часто покупают чай

**Анализ последовательностей** (учитываем время)

- Поиск последовательных правил
  Последовательное правило: Если сегодня купил принтер, то через месяц купит картридж

**Анализ формальных понятий** — формализация описания понятия в виде пары (объём, содержание)

- Поиск формальных понятий
- Построение и анализ решёток понятий

# Исторический обзор развития ИАД

- Решение отдельных задач, методы под конкретные задачи. Первый этап начался в конце 50-х годов. Для конкретных прикладных задач инженерами разрабатывались отдельные алгоритмы распознавания.
- Появление моделей (обобщение методов, отбор успешных методов). Оформление (параметрических) семейств алгоритмов. Место моделей прикладных областей заняли семейства алгоритмов, которые можно считать моделями процессов преобразования.
- Коллективы алгоритмов. Начались попытки совместно использовать несколько моделей для решения одной задачи.
- Операции над алгоритмами, алгебры алгоритмов. Предпосылкой для возникновения алгебраического подхода послужило внутреннее противоречие, присущее самой идее использования заранее фиксированных пареметрических семейств алгоритмов.

# Технологические спутники ИАД

- Базы данных (Databases)
- Консолидация данных
- Хранилища данных (Warehouses)

---

- Аналитическая обработка (OLAP, on-line analytical processing)
- Системы отчетности (Reporting systems)

---

- Интеллектуальный анализ данных (Data Mining, Knowledge Discovery, Machine Learning)

# Технологические спутники ИАД

- Сбор данных
- Проверка гипотез
- Генерация гипотез

# Заключение

Ваши вопросы?