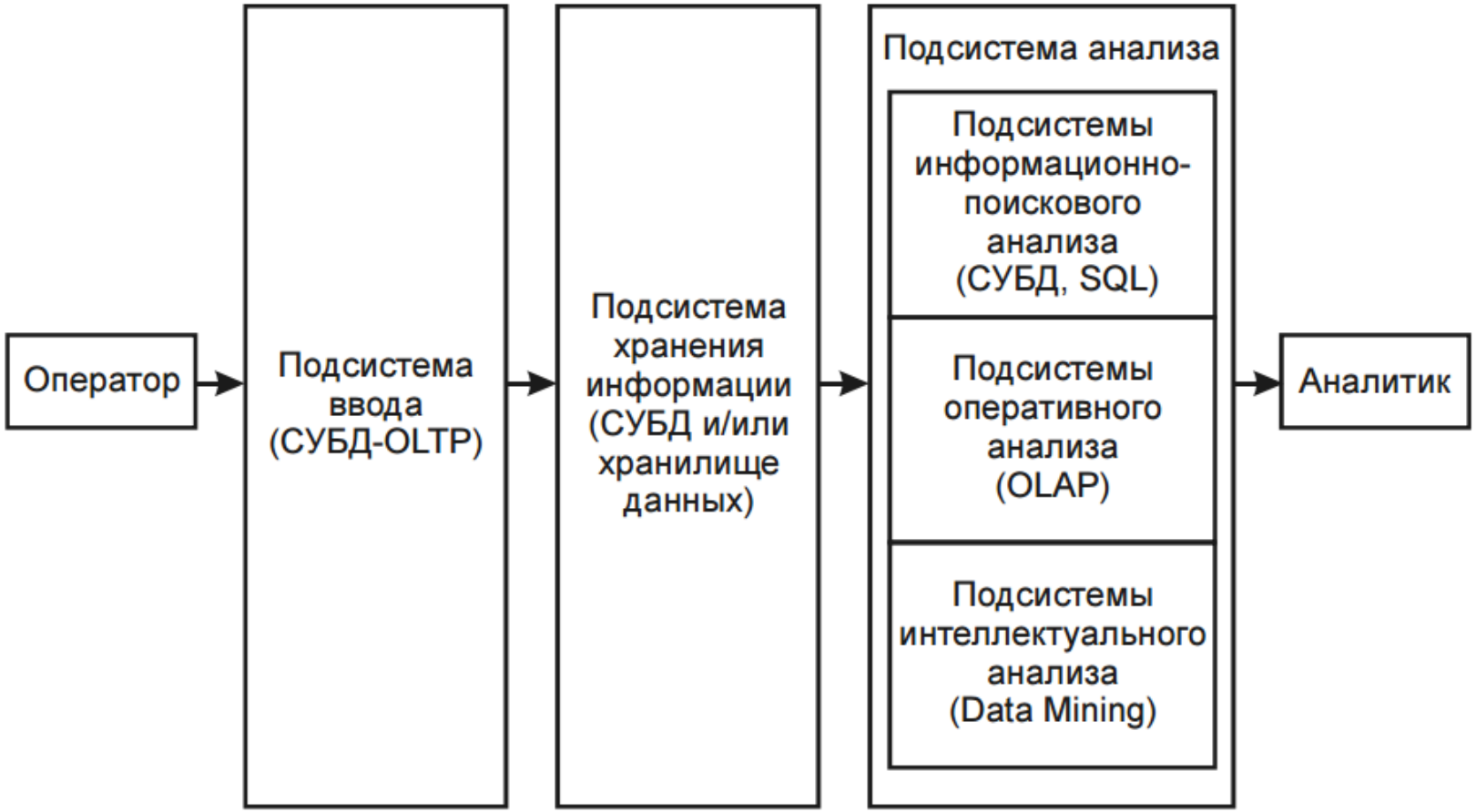
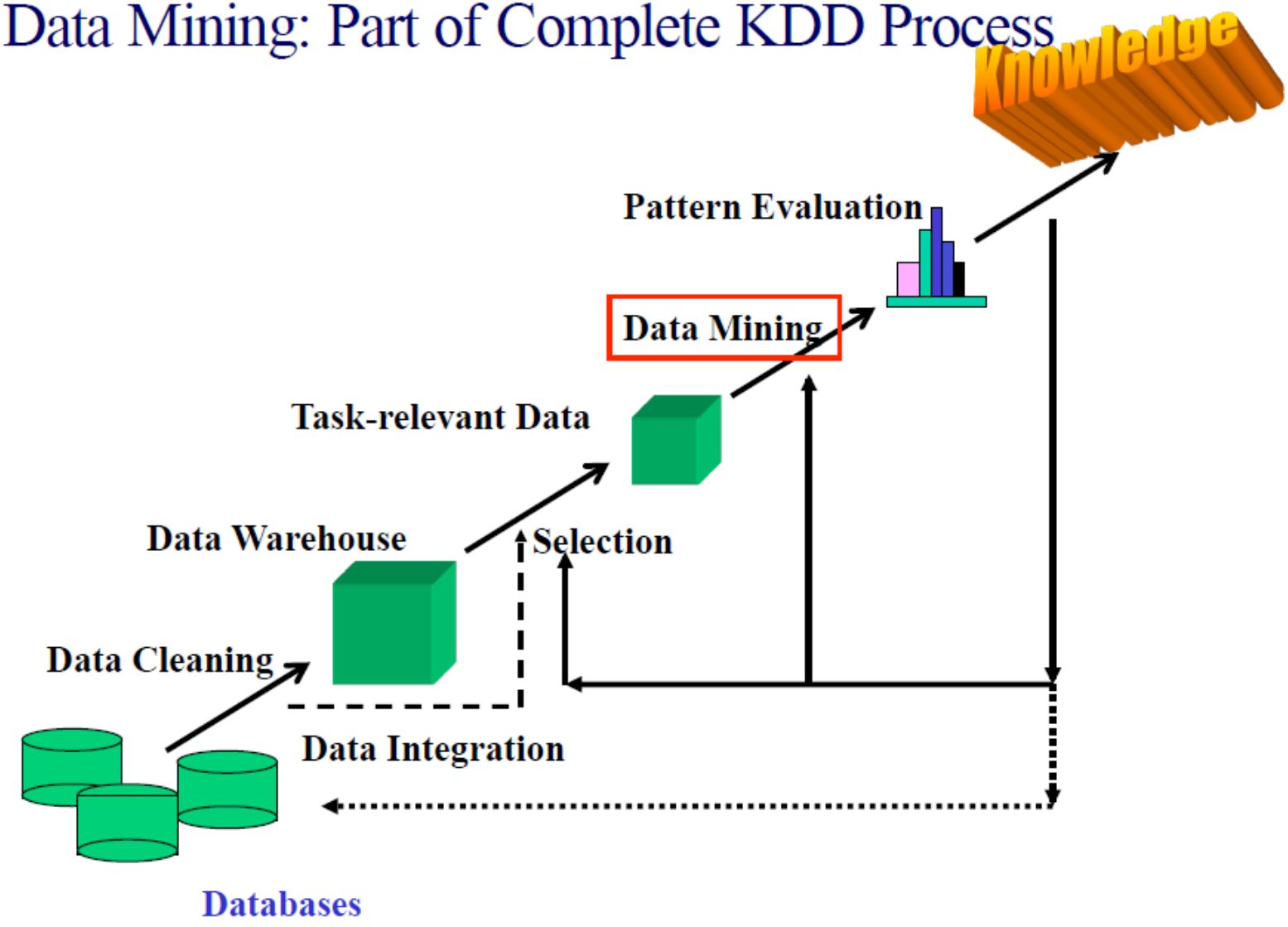
Введение в Data Mining

Гагарина Д.А., 2016

Обобщенная архитектура СППР



Data Mining: Part of Complete KDD Process



Data Mining

- Совокупность методов обнаружения в данных ранее неизвестных и нетривиальных знаний для принятия решений
- Термин введен в 1989 Пятецким-Шапиро
- В основе методы статистики, искусственного интеллекта и базы данных

Требования к Data Mining

- **Новые**, ранее неизвестные знания.
- **Нетривиальные** знания (неочевидные, неожиданные, скрытые).
- **Практически полезные** знания, применимые на новых данных с высокой степенью достоверности.
- Знания, доступные для понимания и **интерпретации**.

Классификация задач Data Mining по назначению

- **Классификация**: определение **класса** объекта.
- **Регрессия**: определение значения **параметра**.
- **Поиск ассоциативных правил**: нахождение зависимостей и **связей**.
- **Кластеризация**: **группировка** объектов.

Классификация задач Data Mining по назначению:

- **Описательные** (descriptive): понимание анализируемых данных:
  - Кластеризация
  - Поиск ассоциативных правил.
- **Предсказательные** (predictive): моделирование и предсказание результатов:
  - Классификация
  - Регрессия
  - Поиск ассоциативных правил, если результаты используются для предсказания.

Классификация задач Data Mining по способу решения:

- **Supervised learning** (обучение с учителем):
  - Этапы:
    - Построение классификатора
    - Обучение классификатора
  - Задачи:
    - Классификация
    - Регрессия
- **Unsupervised learning** (обучение без учителя):
  - Можно использовать без предварительных знаний о данных
  - Задачи:
    - Кластеризация
    - Поиск ассоциативных правил.

Классификация и регрессия

- **Классификация** – определение значения одного из параметров анализируемого объекта на основании значений других параметров.

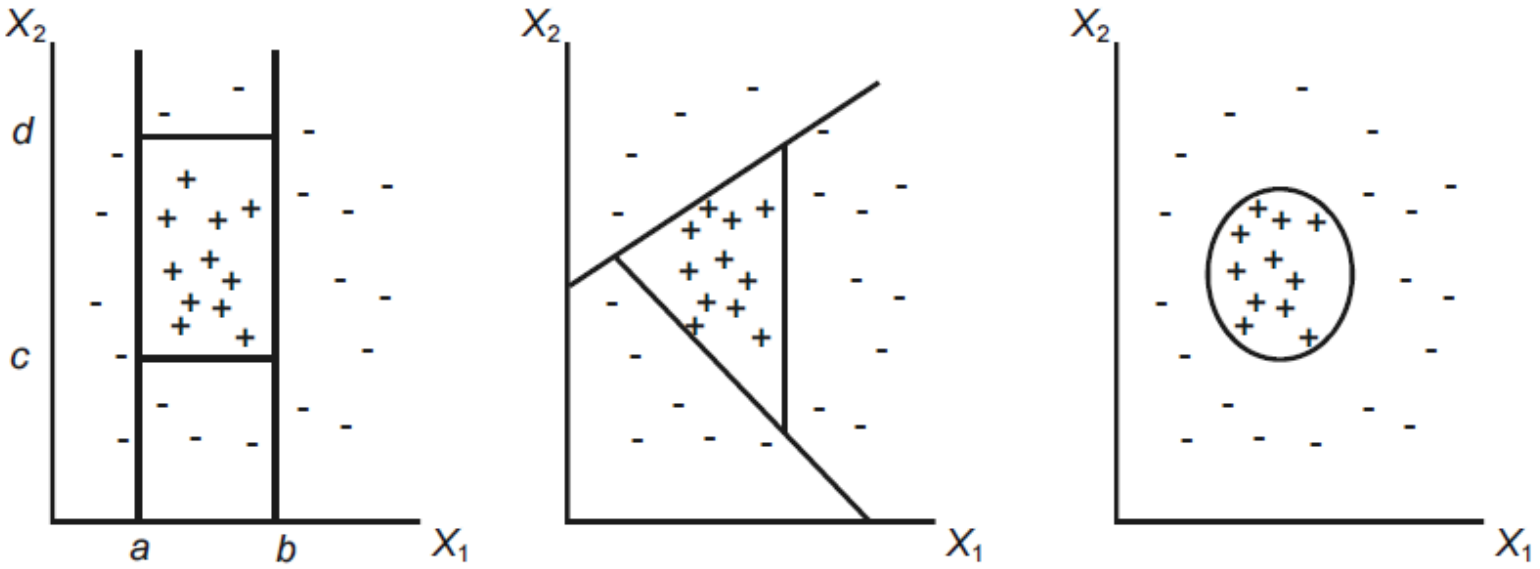
- **Регрессия** – то же самое, но значение параметра – действительное число.

- Задача – построить функцию

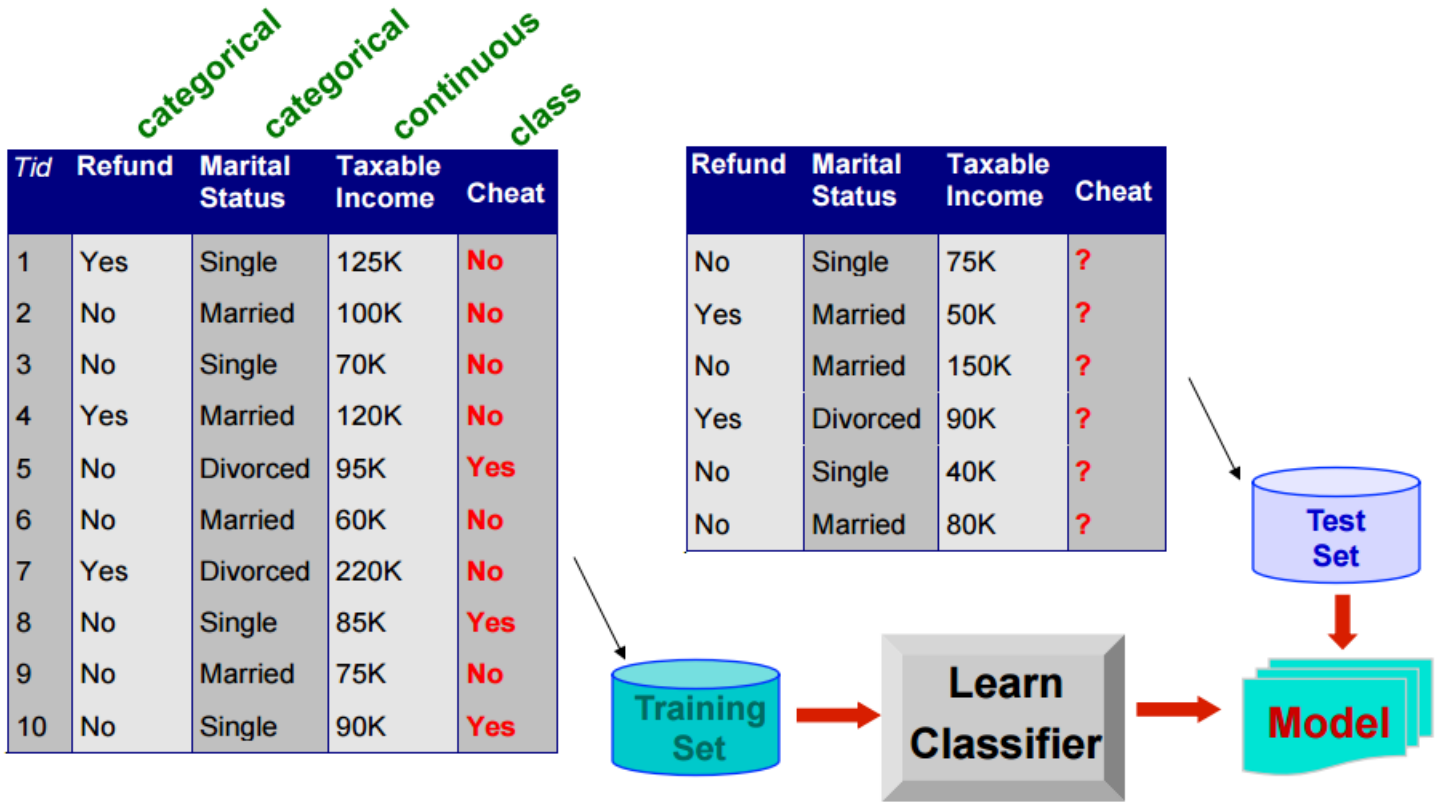
Классификация и регрессия

Зависимый параметр	Возможные значения	Независимые параметры	Обучающая выборка
Кредитоспособность клиента	Да / нет	зарплата, возраст, количество детей…	информация о клиентах, которым выдавались кредиты
Сумма кредита	Число		
Тип сообщения	Spam / mail	частота появления определенных слов	сообщения, классифицированные вручную
Цифра образа	0, 1,..., 9	значения цвета пикселей матрицы	распознанные ранее матрицы образов цифр

Классификация при двух независимых переменных



Пример классификации



Трудности классификации и регрессии

- **Плохое качество** исходных данных: ошибки, пропуски.
- Различные **типы** атрибутов.
- Разная **значимость** атрибутов.
- **Overfitting:** функция "слишком хорошо" адаптируется к данным, ошибки интерпретирует как часть структуры
- **Underfitting:** большое количество ошибок на обучающем множестве, закономерности не обнаруживаются.

Поиск ассоциативных правил

- Суть задачи: определении часто встречающихся **наборов объектов** в большом множестве таких наборов.

- **Сиквенциальный анализ**: обнаружение последовательности событий (объектов).

Поиск ассоциативных правил, примеры

- Basket Analysis:
  - какие товары покупаются вместе,
  - в какой последовательности,
  - какие категории потребителей какие товары предпочитают,
  - в какие периоды времени
- Сфера обслуживания:
  - какие услуги используются в совокупности
- Медицина:
  - сочетания болезней и симптомов

Пример поиска правил

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Кластеризация

- Задача: разделение исследуемого множества на группы "похожих" объектов – кластеры.

- Чем кластеризация отличается от классификации?

Отличия кластеризации:

- Нет выделенной зависимой переменной,
- Можно использовать на начальных этапах, когда данных мало
- Unsupervised learning
- Описательная задача, позволяет понять данные.

Кластеризация, примеры

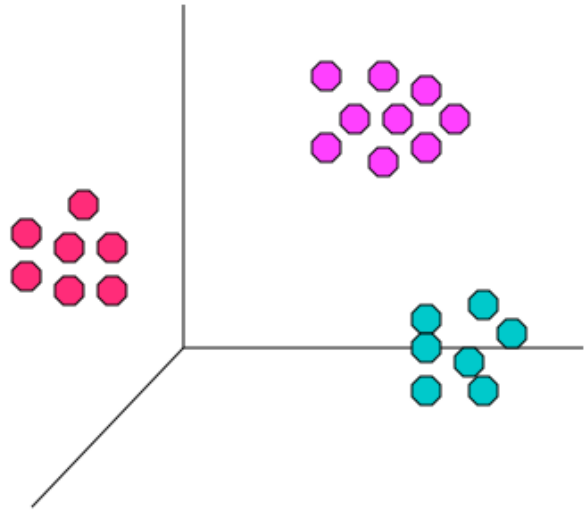
- Сегментация в маркетинге:
  - Кластеризация потребителей по географии, социальны, половозрастным характеристикам, мотивас совершения покупки и т. п.
- Таблица Менделеева

Иллюстрация кластеризации

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Применение Data Mining

- Страхование
- Торговля
- Интернет
- Коммуникации
- Телекоммуникации
- Промышленность
- Банковское дело

Методы Data Mining

- Перебор
- Статистические методы (корреляционный, регрессионный анализ и др.)
- Нечеткая логика
- Генетические алгоритмы
- Нейронные сети

Этапы анализа данных