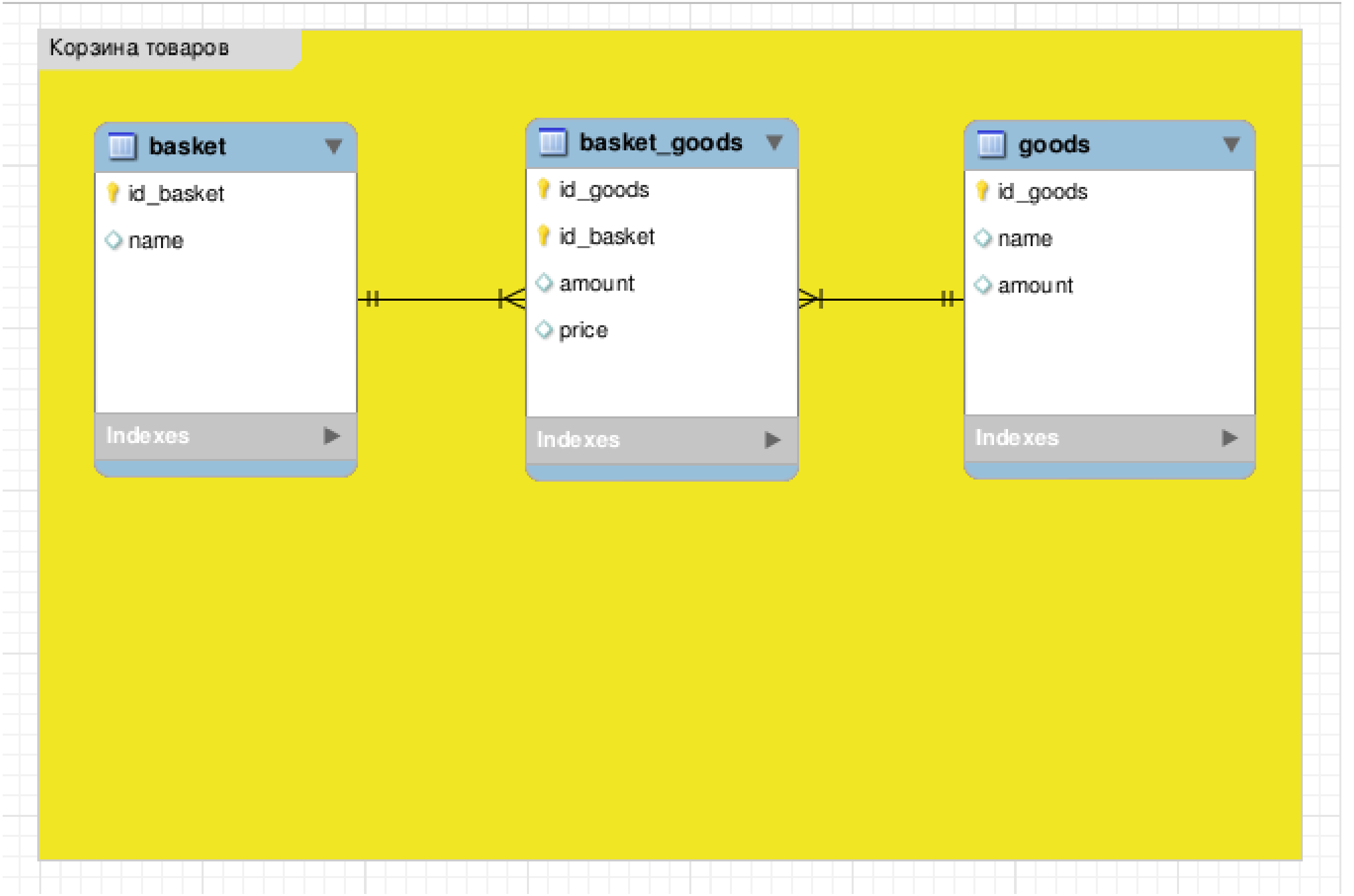
Запросы с агрегацией

Агрегатные функции

SUM() — суммирует значения столбца.
AVG() — среднее значение в столбце
MIN() — наименьшее значение в столбце.
MAX() — наибольшее значение в столбце.
COUNT() — количество записей в столбце.
DISTINCT — выводит значения без повторов.

Модель


Корзина товаров
basket
id_basket
name
Indexes
basket_goods
id_goods
id_basket
amount
price
Indexes
goods
id_goods
name
amount
Indexes

Данные

BASKET

ID_BASKET	NAME
1	Корзина1
2	Корзина2
3	Корзина3

GOODS

ID_GOODS	NAME	AMOUNT	PRICE
1	Шапка ушанка	10	**400**
2	Лапти	5	**300**
3	Самовар	4	**500**
4	Платок	45	**200**
5	Румяна	2	**700**

BASKET_GOODS

id_goods	id_basket	amount
1	1	**2**
5	1	**1**
3	1	**2**
2	1	**1**
4	1	**1**
1	2	**2**
3	2	**3**
4	2	**4**
2	3	**3**
2	3	**10**
3	3	**3**
5	3	**1**

COUTN()

```
// Кол-во товарных позиций на складе
select count(*) from goods;

// Кол-во уникальных товаров в корзине
select count(distinct id_goods) from
basket_goods;
```

MAX(), MIN()

```
// Максимальное  кол-во купленного товара
  select max(amount) from basket_goods;


 // Минимальное  кол-во купленного товара
      уникальных товаров в корзине
  select min(amount) from basket_goods;
```

Шаблон запроса

```
SELECT col1, summ(amount)
  FROM table_name
 WHERE expression
 GROUP BY col1
HAVING summ(col1) > value
 ORDER BY col2
```

Пример 1

```
// Получим сумму заказа по корзине

SELECT col1, summ(bg.amount*g.price)
  FROM basket_goods bg, goods g
 WHERE bg.id_goods = g.id_goods
   AND id_basket = 1
 GROUP BY id_basket
```

Пример 2

```
// Получим корзины где сумма заказа
   больше 1000 руб.

SELECT bg.id_basket,summ(bg.amount*g.price)
  FROM basket_goods bg, goods g
 WHERE bg.id_goods = g.id_goods
 GROUP BY bg.id_basket
HAVING summ(bg.amount*g.price) > 1000
 ORDER BY bg.id_basket
```

Задание

Для предложенной модели получить:

1. Товары которых нет в корзине

2. Средний чек по трем корзинам

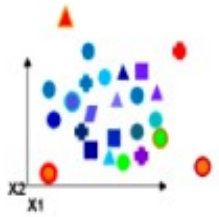
3. Максимальный, минимальный чек


DATA
MINING
PATTERNS
SET
ANALYSIS
LEARNING
EXAMPLE
CUSTOMERS
CUSTOMER
DECISION
METHODS
PROCESS
MACHINE
INTELLIGENCE
KNOWLEDGE
MODELING
DATABASE
APPLICATIONS
INFORMATION
ALGORITHMS
IDENTIFY
PEOPLE
USE
DISCOVERY
GROUPS
SOPHISTICATED
OFFERS
ALGORITHM
TASK
ARTIFICIAL
TARGET
PREDICT
NEEDED
COMMON
TERM
VISUALIZATION
MANAGEMENT
MIGHT
BASKET
SEND
SYSTEM
GENERAL
E-MAILS
MODELS
SENDING
NEW
STEP
CITATION
DETERMINE
ASSOCIATION
HELP
RULE
USING
BUSINESS
FINALLY
STRUCTURES
WITHIN
ANOTHER
SUPPORT
TEST
INVOLVES
TRAINED
KNOWN
PRESENT
NUMBER
FINAL
DISCOVERED
POTENTIAL
OFTEN
RECORDS
LARGER
LARGE
MODEL
MOST
WITHOUT
MAKING
MARKET
PRE-PROCESSING
PRODUCT
DETECTION
LIKELY
LEARNED
OFFER
E-MAIL
MARKETING
MANY
DESIRED

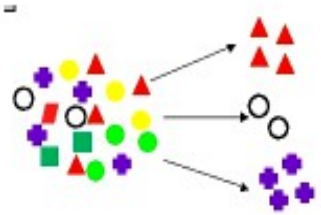
Добыча данных

DATA MINING (Интеллектуальный анализ данных)- это технология выявления скрытых взаимосвязей внутри больших баз данных

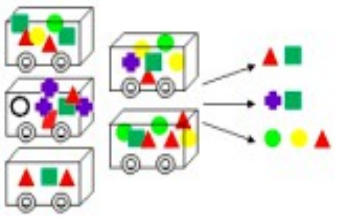
Разведочный анализ данных строится на алгоритмах



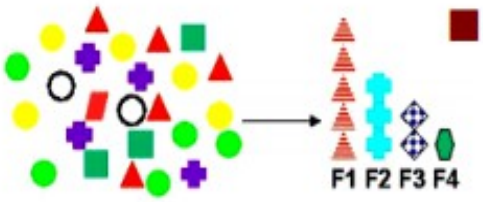
- Определение выбросов
  - SVM с одним классом



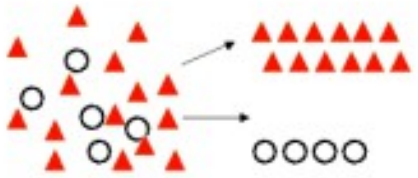
- Кластеризация
  - расширенный алгоритм k-средних
  - O-кластер



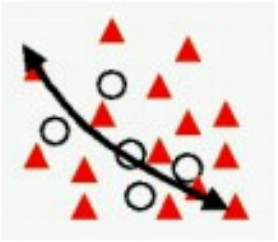
- Ассоциация
  - Apriori



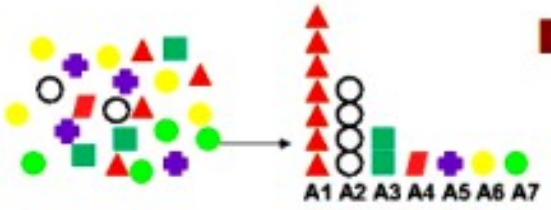
- Извлечение свойств



- Классификация
  - логистическая регрессия (GLM)
  - naïve Bayes
  - SVM
  - деревья решений



- Регрессия
  - множественная регрессия
  - SVM



- Значимые атрибуты
  - принцип минимальной длины

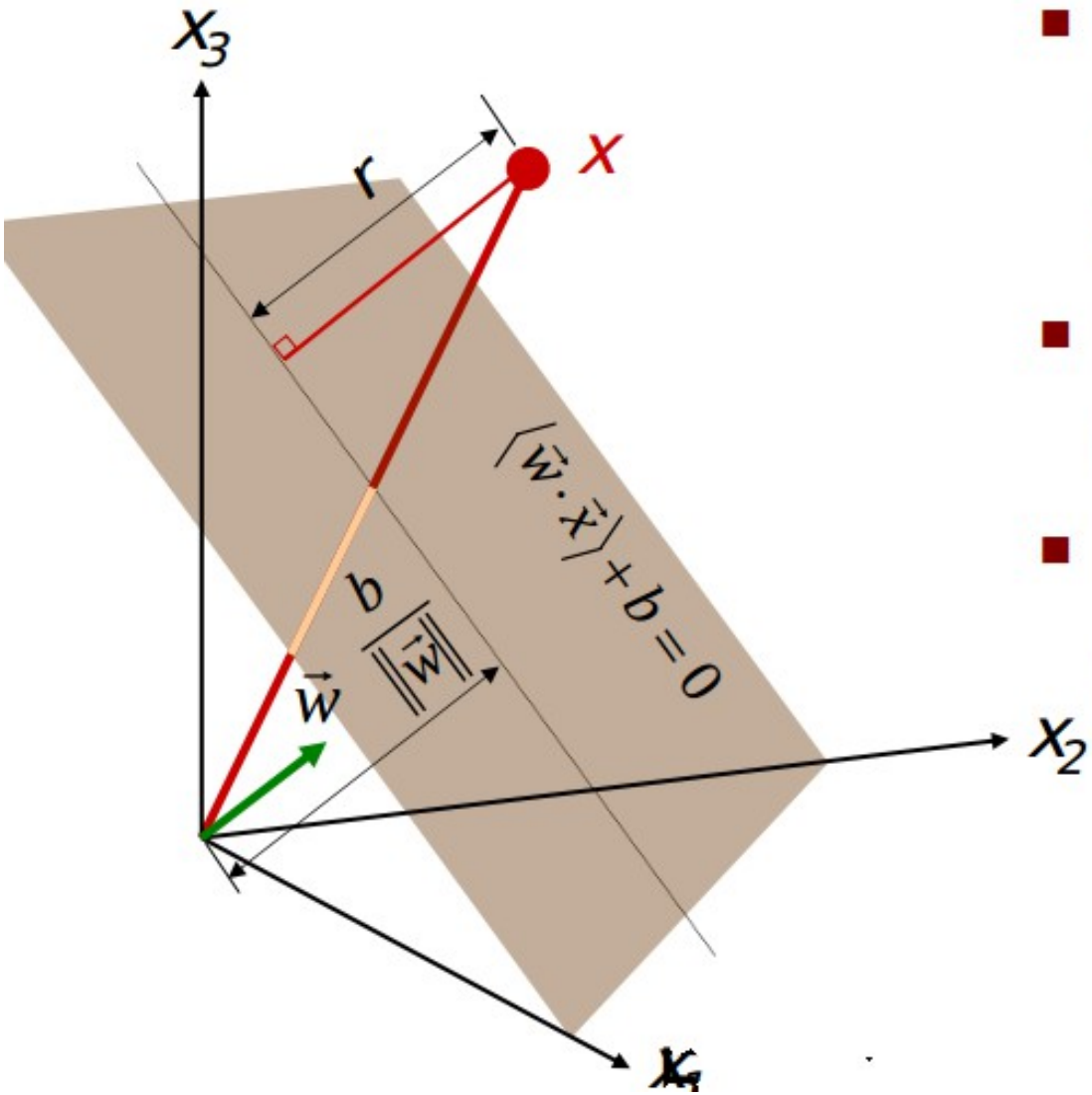
Кластерный анализ

- Используется в маркетинге (группы населения с одними и теми же характеристиками), медицина (пациенты с тем же беспокойством), управлении персоналом и т.д.
- Отличается с классификацией, поскольку не используется обучение

Ассоциативный анализ

- Используется в маркетинге (группы населения с одними и теми же характеристиками), медицина (пациенты с тем же беспокойством), управлении персоналом и т.д.
- Разделение БД на подмножества, так что внутри подгруппы различия между отдельными объектами меньше, чем между разными подгруппами

SVM



- Определяется гиперплоскость в пространстве параметров
- Коэффициенты $\vec{w}$ и смещение b
- Прогнозирование:
  $f = sign(\langle \vec{w} \cdot \vec{x} \rangle + b = 0)$

Функции БД Oracle

- Ранжирование
  - rank, dense_rank, cume_dist, percent_rank, ntile
- Агрегирование
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value
- Корреляция и регрессия
  - Correlation, linear regression family, covariance
- Линейная регрессия
  - МНК.
  - COVAR_POP, COVAR_SAMP, and CORR functions.
- Соответствие распределениям
  - тесты Колмогорова-Смирнова, Андерсона-Дарлинга, хи-квадрат, Гаусса, Вейбула, экспоненциальный
- Описательная статистика
  - среднее, стд. отклонение, дисперсия, min, max, медиана, мода
  - DBMS_STAT_FUNCS: описательная статистика по числовым колонкам
- Корреляции
  - Пирсона, Спирмана, Кендалла
- Кросс-табуляции
  - χ^2, φ, V Крамера, коэффициента сопряженности, λ Кохена
- Hypothesis Testing
  - тест Стьюдента, Фишера, биноминальный, Уилкоксона, χ^2, Манна-Уитни, Колмогорова-Смирнова, дисперсионный анализ

Примеры

- Космос
  - Проект SKYCAT. За 6 лет в Second Palomar Observatory собрали 3 ТБ изображений примерно о 2 млн. объектов в небе.
  - Используя кластеризацию и деревья решений объекты были систематизированы. Результаты помогли астрономам открыть 16 новых квазаров, определение которых связано с большими сложностями.