

Основы анализа текстовых данных

Кафедра управления и интеллектуальных технологий НИУ «МЭИ»
Весна 2023 г.

Что такое Data Mining?

Data Mining - совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Термин *Data Mining* («Добыча данных») введен Григорием Пятецким-Шапиро в 1989г: Имеется большая база данных, из которой хотим извлечь «Скрытые знания»:

- Ранее неизвестные
- Нетривиальные
- Полезные для практики
- Интерпретируемые

Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, а также статистические методы, из которых большую часть составляют методы машинного обучения (Machine Learning).

Что такое Machine Learning?



ML - обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Постановка задачи ML:

X – Множество объектов;

Y – Множество ответов

Имеется неизвестная целевая функция (target function) :

$$y: X \rightarrow Y$$

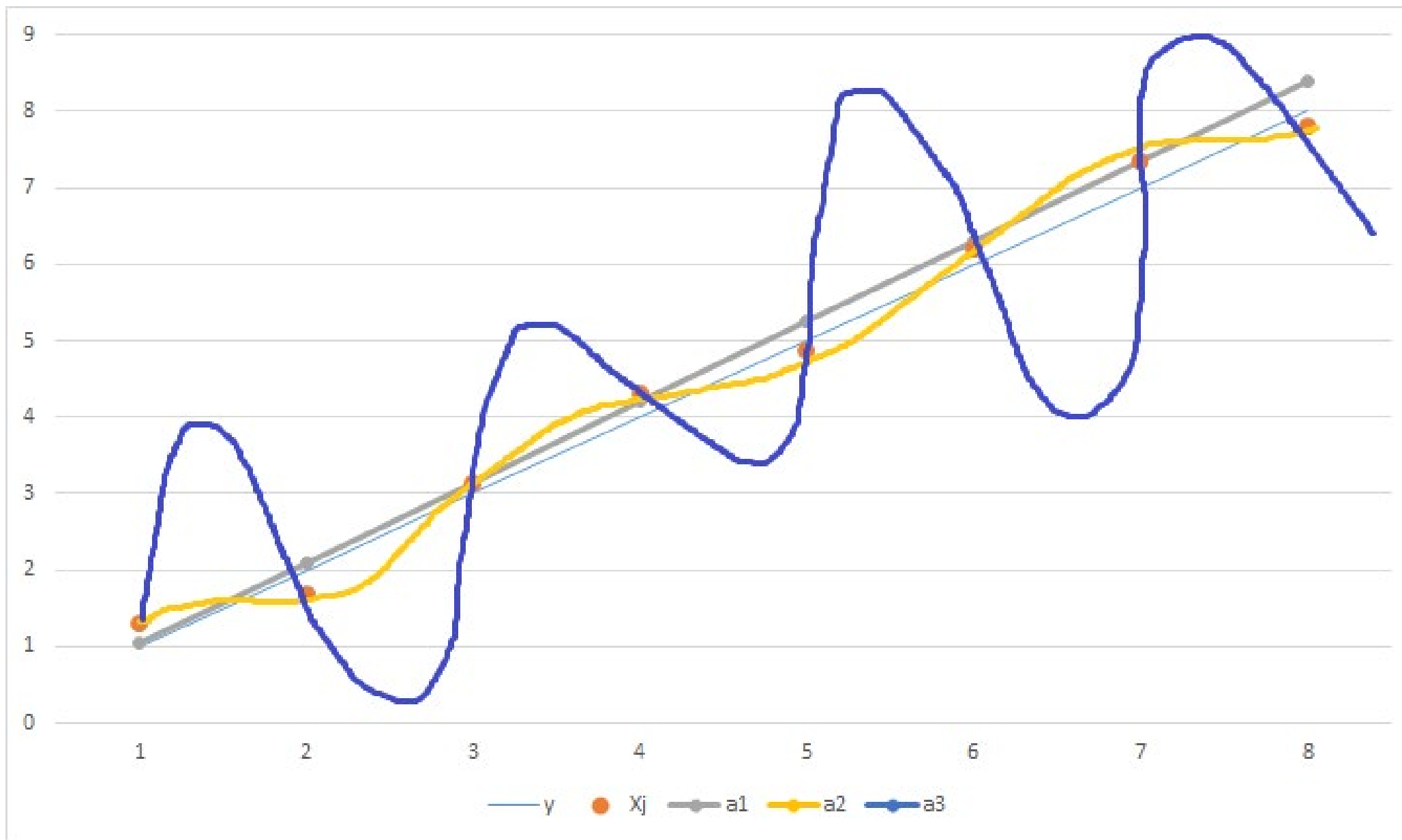
Дано:

$\{\vec{X}_1, \dots, \vec{X}_l\} \in X$ - Выборка объектов (dataset)

$y_j = y(\vec{X}_j); j = 1 \dots N$ - Известные ответы

Найти $a: X \rightarrow Y$ - алгоритм, решающая функция (decision function),
приближающийся к y на всем множестве X

Способы построения алгоритма а:



Объекты и признаки:

$x_j^{(i)}$ - Признаки/свойства (features)
 $i = 1 \dots M$

Выбор и построение (формирование) признаков – важный этап решения задачи ML

Виды признаков:

- Бинарный
- Номинальный
- Порядковый
- Количественный

Объект часто описывается в виде вектора:

$$\vec{X}_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_j^{(M)} \end{bmatrix}$$

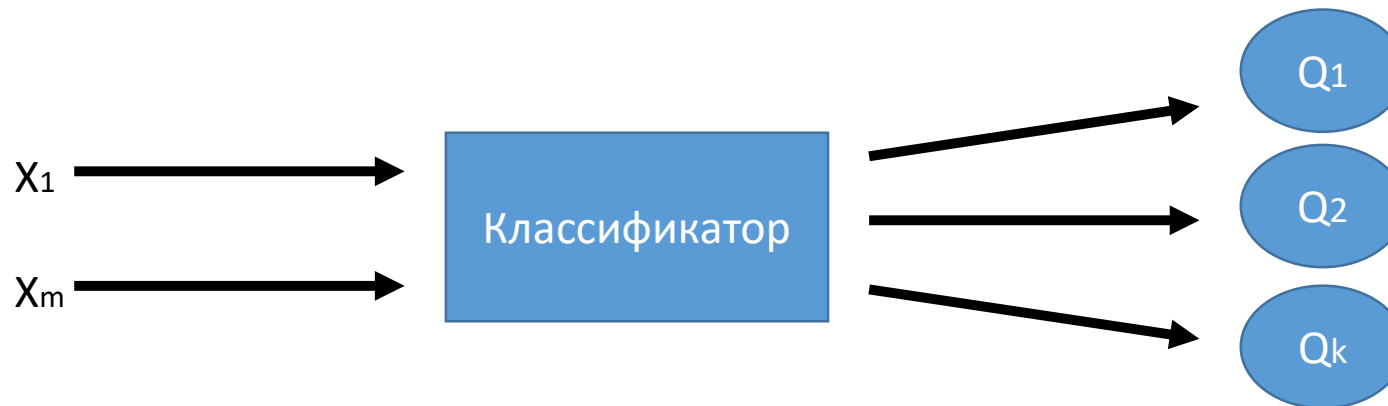
Выборка – в виде матрицы «объект-признак»:

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \dots & x_N^{(M)} \end{pmatrix}$$

Типы задач. Классификация

1. Задачи классификации (Classification):

- $Y = \{0;1\}$ – бинарная классификация (классификация на 2 класса)
- $Y = \{1,...K\}$ – На K непересекающихся классов
- $Y = \{0;1\}^K$ - На K классов, которые могут пересекаться
- $Y = \{0;1\}$ – одноклассовая классификация (есть один класс и «все остальное»)

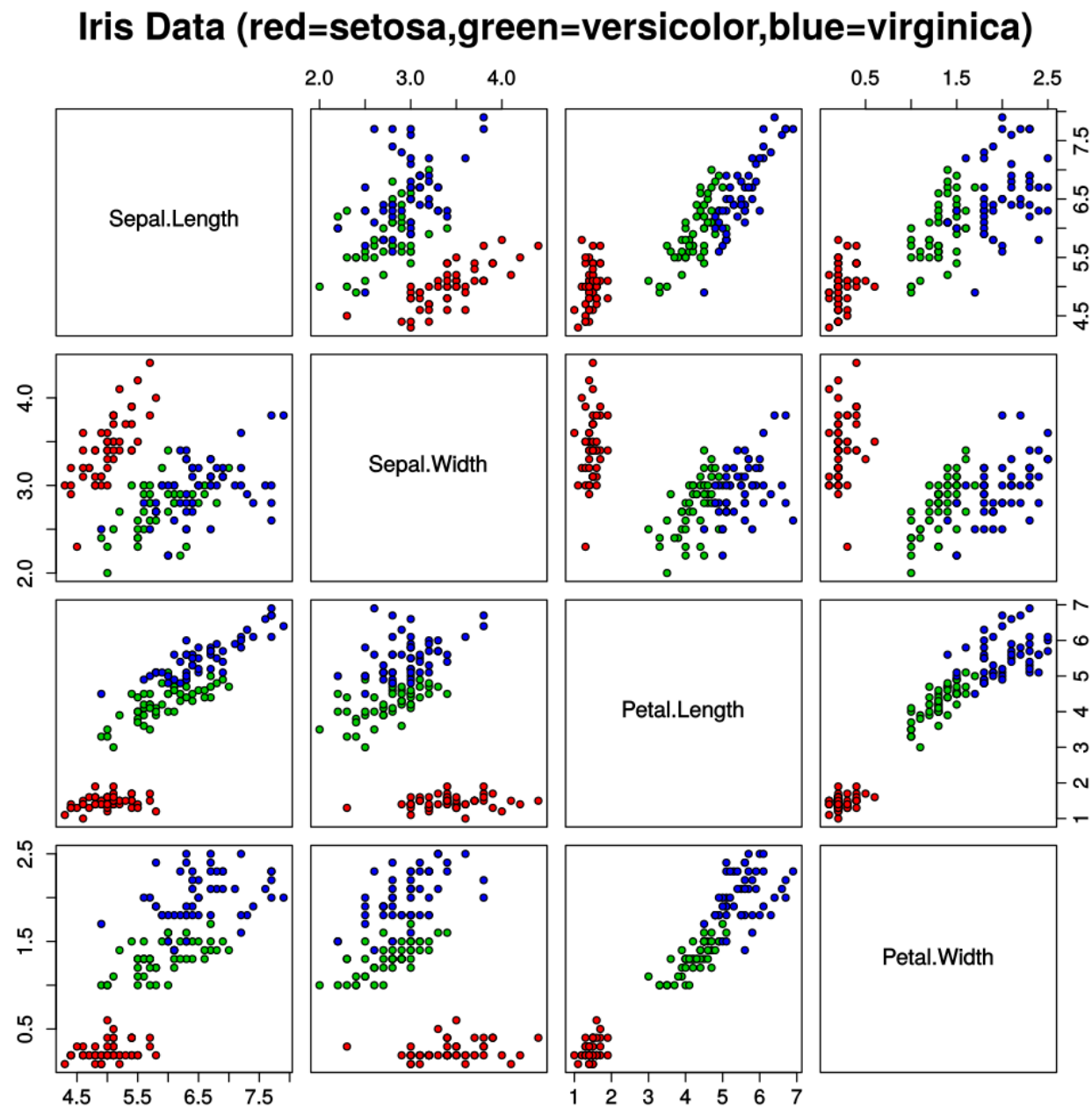


Пример: Задача
классификации цветков ириса
(Фишер, 1936г.)

$i = 4$ признака

$|Y| = 3$ класса

Длина выборки $N = 150$



Этап обучения и применения

- Обучение. Строим алгоритм a по выборке:

$$\begin{pmatrix} x_1^{(1)} & \dots & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \dots & x_N^{(M)} \end{pmatrix} \rightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \Rightarrow a$$

- Применение. Алгоритм a для новых объектов выдает ответы:

$$\begin{pmatrix} \tilde{x}_1^{(1)} & \dots & \tilde{x}_1^{(M)} \\ \vdots & \ddots & \vdots \\ \tilde{x}_T^{(1)} & \dots & \tilde{x}_T^{(M)} \end{pmatrix} \xrightarrow{a} \begin{pmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_T \end{pmatrix}$$

Типы задач. Регрессия и ранжирование

2. Задачи восстановления регрессии (Regression)

- $Y = R$

3. Задачи ранжирования (ranking):

- Y - конечное упорядоченное множество

Типы задач. Кластеризация

4. Кластеризация (clustering) – обучение без учителя (unsupervised learning)

- Y – не задается. В этом случае требуется искать зависимости между объектами

Кредитный скоринг

- **Объект** – заявка на получение кредита
- **Классы:** good, bad
- **Примеры признаков:**
 - Бинарные: пол, наличие телефона,...
 - Номинальные: место работы, профессия, место жительства,...
 - Порядковые: должность, образование,...
 - Количественные: возраст, зарплата, стаж работы, сумма кредита,...
- **Особенности задачи:**
 - Вероятны пропуски данных
 - Возможна недостоверность данных
 - Нужно оценить вероятность дефолта $P(bad)$

Предсказание оттока клиентов

- **Объект** – абонент в определенный момент времени
- **Классы:** уйдет или не уйдет в следующем месяце
- **Примеры признаков:**
 - Бинарные: включенные услуги, корпоративный клиент...
 - Номинальные: тарифный план, регион проживания,...
 - Количественные: длительность разговоров (входящих, исходящих, СМС, трафик), сумма оплаты, частота оплаты,...
- **Особенности задачи:**
 - Сверхбольшие выборки
 - Непонятно, какие признаки вычислять по «сырым данным»
 - Нужно оценить вероятность ухода

Задача ранжирования поисковой выдачи

- **Объект** – пара <запрос, документ>
- **Классы:** релевантен или не релевантен
- **Примеры признаков:**
 - Количественные: частота слов запроса в документе, число ссылок на документ, число кликов на документ,...
- **Особенности задачи:**
 - Оптимизируется не число ошибок, а качество ранжирования
 - Сверхбольшие выборки

Категоризация текстовых документов

- **Объект** – текстовый документ
- **Классы:** Рубрики тематического каталога
- **Примеры признаков:**
 - Номинальные: автор, год, издание,...
 - Количественные: Частота появления терминов в документе, в названии, в ключевых словах,...
- **Особенности задачи:**
 - Очень большое количество признаков (слов, словоформ)
 - Документ написан на естественном языке (ЕЯ)
 - Документ может относиться к нескольким рубрикам

Text Mining – интеллектуальный анализ текстов

- **Категоризация текстов (*classification*)** –
 - отнесении документов из коллекции к одной или нескольким группам (классам, кластерам) схожих между собой текстов
- **Извлечение информации (*information extraction*)** –
 - это задача автоматического извлечения (построения) структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов (распознавание имен людей, названий организаций, поиск ключевых слов для текста, автореферирование)
- **Информационный поиск (*information retrieval*)** –
 - процесс поиска *неструктурированной* документальной информации, удовлетворяющей информационные потребности (процесс выявления в некотором множестве документов всех тех, которые посвящены указанной теме)