Конспект лекции

А.Н. Шевляков

Машинное обучение

Введение в машинное обучение и основные понятия статистики. Первичная обработка данных

# Задачи и модели машинного обучения

**Где можно встретить модели машинного обучения?**

* Рекомендательные системы
* Работа с текстом
* Обработка изображений

**Другие задачи машинного обучения**

* Предсказание стоимости жилья
* Кредитный скоринг
* Прогнозирование спроса на товары
* Медицинская диагностика
* Ранжирование поисковой выдачи
* Поиски аномалий в данных

**Связь с математическими дисциплинами**

Методы машинного обучения используют результаты математических дисциплин: математическая статистика, методы оптимизации, идеи из геометрии и линейной алгебры.

**Основные задачи машинного обучения**

* Восстановление пропущенных или поврежденных данных.
* Поиск выбросов (outlier detection). Есть множество объектов М. Найти в нем все аномальные объекты.
* Поиск новизны (novelty detection). Есть множество объектов М. Определить, является ли объект А∉М похожим на объекты из М или нет?
* Кластеризация (clustering). Дано множество объектов. Их нужно разбить на несколько групп (кластеров), состоящих из похожих друг на друга объектов.
* Предсказание (prediction). Есть множество объектов М с известными значениями признака Y. Найти значение признака Y для нового объекта А∉М.

**Насколько эффективно машинное обучение?**

Проблемы:

- некомпетентность заказчиков;

МО как фетиш;

- эффект от внедрения МО можно заметить не сразу;

- кто будет нести ответственность за ошибки модели МО?

- если модель МО дает правильный ответ, то не совпадение ли это?

- сколько случаев совпадения с правильным ответом является признаком присутствия искусственного интеллекта?

**Что нужно знать при работе в области машинного обучения?**

- основные идеи алгоритмов МО; что они могут делать, а что они сделать не в состоянии;

- какие операции при работе с данными не поддаются автоматизации, насколько результат МО зависит от «человеческого фактора»;

- методы обмана заказчика (это должны знать и исполнители, и заказчики).

# Представление данных для машинного обучения

## Представление данных

Данные будем представлять в виде таблицы (Таблица 1).

Таблица 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Объекты** | **Признак 1** | **Признак 2** | **…** | **Признак m** |
| A1 | P11 | P12 | … | P1m |
| A2 | P21 | P22 | … | P2m |
| … | ... | … | … | … |
| An | Pn1 | Pn2 | … | Pnm |

Здесь n объектов, у каждого их которых имеется m признаков (фич).

С помощью P(A) будем обозначать значение признака Р для объекта А. Число n называется объёмом выборки.

## Типы признаков

Признак называется:

**Количественным (числовым) –** если область его значений – вещественные числа, и сам признак имеет числовую природу.

**Порядковым –** если признак задает порядок на объектах.

**Номинальным (категориальным)** – если признак не имеет числовой природы и, как правило, число его возможных значений конечно. В частности, бинарный признак – это номинальный признак с 2-мя возможными значениями.

Все объекты должны иметь одинаковое количество признаков, чтобы их можно было внести в одну таблицу.

# Характеристики признаков. Основные понятия математической статистики

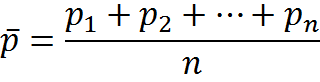
## Признаки

Пусть *Р* – столбец со значениями числового признака Р из нашей таблицы:



Что можно посчитать для *P*?

1. Минимальное и максимальное значение.
2. Среднеезначение



3. Медиану-  такое число *hP*, что ровно половина из элементов *pi* больше него, а другая половина меньше него.

Медиана – это не то же самое, что и среднее: для (3,5,5,9,11) среднее значение 4.6, а медиана 5.

## Вычисление медианы

Для выборок чётного объёма медиана не определена однозначно.

Где тут медиана: (1,3,4,5)?

Но на практике в качестве медианы берут среднее арифметическое между двумя «центральными» значениями.

То есть в примере выше медиана равна (3+4)/2=3.5

**Совет:** чтобы быстро вычислить медиану, нужно мысленно упорядочить массив по возрастанию и взять число из середины.

## Медиана vs среднее

Значение медианы не так сильно (как среднее) зависит от попадания в выборку аномально больших и аномально малых значений признака.

Теперь понимаете, почему официальная пропаганда всегда употребляет понятия:

* «Средняя зарплата по стране» (а не «Медианная зарплата по стране»)
* «Средняя продолжительность жизни»

## Симметричные выборки

Если медиана и среднее близки друг к другу, то выборка называется симметричной.

Важность симметричных выборок: для них проще искать аномалии.

## Мода

**Мода** – значение, которое встречается наиболее часто в выборке.

Например, модой здесь (2,0,1,1,3,2,3,2) будет 2.

Мода не всегда определена однозначно.

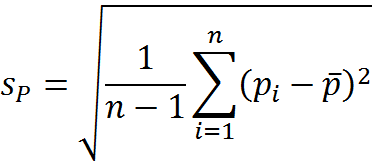
Мода (в отличие от среднего и от медианы) имеет смысл и для номинальных признаков.

## Отклонение

Среднее и медиана не достаточны для адекватного описания выборки.

Например, выборки (0,0,0,0,0) и (-2,-1,0,1,2) имеют одинаковые средние и медианы. Однако во второй выборке значения чаще отклоняются от среднего.

Отклонение (полное название: среднее квадратическое отклонение) считается по формуле:



## Симметричные выборки

Если медиана *hP* и среднее близки друг к другу, то выборка называется **симметричной**.

А что значит: близки друг к другу? На практике считают, что если выполнено неравенство



## Коэффициент корреляции

Нужна величина, которая показывает, как значения одного признака определяют значения другого признака. Эта величина должна иметь смысл и для признаков с разными единицами измерения.

В статистике для таких задач используют **коэффициент корреляции (КК).**

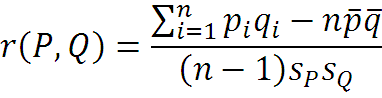
Пусть





признаки (столбцы из таблицы).

Тогда КК считается по формуле



**Свойства коэффициент корреляции**

*Коэффициент корреляции* (КК) – это число из отрезка [–1,1], которое имеет следующий смысл:

1. Если КК=0 (или близок к нему), то очевидной зависимости между признаками *P, Q* нет.
2. Если КК>0, то бОльшим значениям признака P, как правило, соответствуют бОльшие значения признака Q.
3. Если КК<0, то бОльшим значениям признака P, как правило, соответствуют меньшие значения признака Q.
4. Чем ближе значение КК к единице, тем сильнее зависимость между признаками P, Q.
5. Если модуль КК равен 1, то между признаками P, Q существует линейная зависимость.