



САНКТ-ПЕТЕРБУРГСКИЙ  
ГОСУДАРСТВЕННЫЙ  
ЭКОНОМИЧЕСКИЙ  
УНИВЕРСИТЕТ

# Отбор признаков

Занятие 3

Глазунова Е.В.

# Зачем отбирать признаки?

## Зачем отбирать признаки?

- устранение шума из данных
- эффективность алгоритмов зависит от количества признаков
- проклятие размерности
- упрощение визуального анализа данных
- количество данных и количество признаков зависимы

## Почему мы можем уменьшить размерность?

- Шум или информация не имеет отношения к решению.
- Избыточная информация. Несколько факторов рассказывают примерно про одно и то же, но несколько по-разному.

# Johnson–Lindenstrauss lemma

Для  $0 < \epsilon < 1$  и множества  $X$  из  $m \in \mathbb{Z}_{\geq 1}$  точек в  $\mathbb{R}^N$  ( $N \in \mathbb{Z}_{\geq 0}$ ) и целого  $n > \frac{8(\ln m)}{\epsilon^2}$  существует линейная функция  $f: \mathbb{R}^N \rightarrow \mathbb{R}^n$ , такая что:

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

Для любых  $u, v \in X$ .

# Корреляционный анализ

Коэффициент корреляции Пирсона можно использовать, если данные имеют нормальное распределение.

Для коэффициента корреляции Спирмена не требуется нормальность данных, так как при их расчете анализируются не сами значения параметров, а их взаимное расположение – ранги.

Корреляцию можно считать только для числовых признаков!

Значимость коэффициента корреляции вычисляется на основе **t-критерия Стьюдента** по формуле:

$$t_{\text{набл}} = \sqrt{\frac{r_{x,y}^2}{1-r_{x,y}^2}} (n - 2),$$

где  $n$  – объем выборки.

Гипотеза о значимости коэффициента принимается, если  $t_{\text{набл}} < t_{\text{табл}}$ , и отклоняется в обратном случае.

Для категориальных признаков используются кросс-таблицы и статистический критерий Хи-квадрат.

# Значения Шепли

Значения Шепли объясняют как «справедливо» оценить вклад каждого признака в прогноз модели.

Значение Шепли признака  $i$  для предсказания  $f$  определяется формулой:

$$\phi_i = \sum_{S \subset N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (f(S \cup \{i\}) - f(S))$$

где  $N$  – полный набор признаков;

$S$  – подмножество признаков без  $i$ ;


$f(S \cup \{i\})$  - предсказание модели с признаком  $i$ ;

$f(S)$  - предсказание модели без признака  $i$ .

# Пермутационная важность

Идея алгоритма проста: нужно в наборе данных (валидационном наборе) перетасовать значения признака, влияние которого изучается на данной итерации, оставив остальные признаки (столбцы) и целевой вектор без изменения. Признак считается «важным», если метрики качества модели падают, и соответственно – «неважным», если перестановка не влияет на значения метрик. Перестановочная важность вычисляется после того как модель будет обучена.

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24



# Пермутационная важность

Алгоритм:

1. Обучить модель (в нашем случае линейной регрессии).
2. Для каждого признака (на тестовых данных) делать:
  - 1) перемешать этот признак (столбец);
  - 2) получить прогноз с использованием перемешанного признака;
  - 3) оценить качество прогноза с не перемешанной целевой переменной.
3. Отсортировать признаки по убыванию качества прогноза.