Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 6: Отбор признаков и уменьшение размерности

Дата: 29.05.2025

Квиз

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Дан датасет

какие переменные стоит исключить?

	age	sum_revenue	commission	city	type_client	count_credits	target
0	31.00	7977.81	NaN	Москва	1_cat	NaN	0
1	30.99	6663.96	0.23	Санкт-Петербург	2_cat	4.0	1
2	31.00	1693.85	0.57	Новосибирск	1_cat	NaN	0
3	30.98	3600.01	NaN	Екатеринбург	1_cat	NaN	0
4	31.00	15620.50	NaN	Казань	1_cat	NaN	0
5	31.01	5616.20	0.34	Нижний Новгород	1_cat	3.0	1
6	31.01	13946.07	NaN	Челябинск	1_cat	5.0	1
7	31.02	4307.93	NaN	Самара	2_cat	NaN	0
8	30.99	9859.48	NaN	Ростов-на-Дону	2_cat	NaN	0
9	31.01	8029.71	NaN	Уфа	2_cat	1.0	1

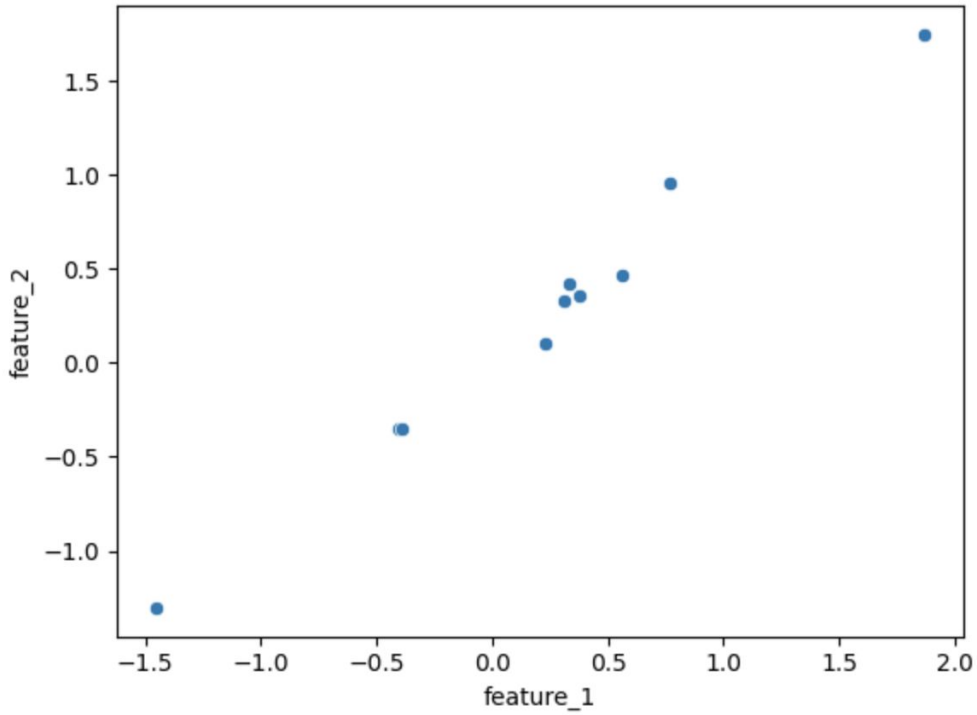
Квиз

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Дан датасет

какие переменные стоит исключить?

feature_1	feature_2
-0.406227	-0.353535
0.312785	0.327164
-1.456869	-1.302416
0.227018	0.099999
0.330188	0.419818
-0.390327	-0.353169
0.374345	0.359539
0.769362	0.956260
0.560357	0.464845
1.870892	1.751288



Коэффициент корреляции: 0.993

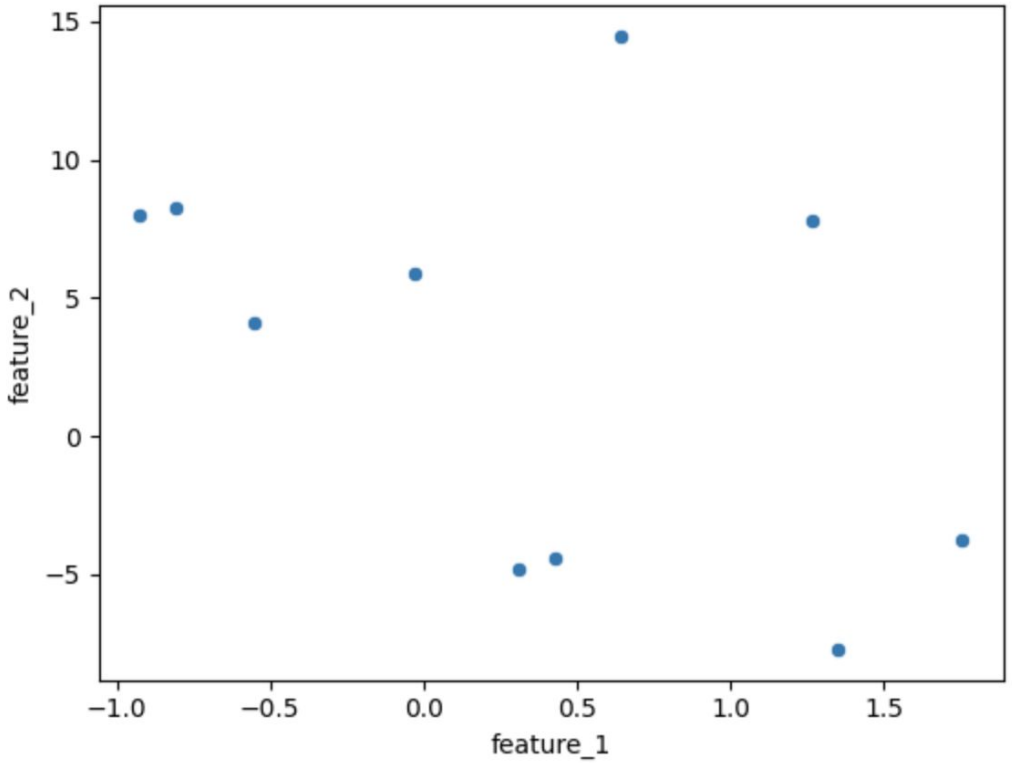
Квиз

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Дан датасет

какие переменные стоит исключить?

	feature_1	feature_2
0	0.308878	-4.815498
1	-0.924618	8.003207
2	1.755843	-3.762387
3	0.426720	-4.418161
4	-0.551693	4.072613
5	-0.808965	8.229064
6	1.266624	7.816509
7	1.350174	-7.696035
8	-0.027228	5.888969
9	0.642903	14.473554



Коэффициент корреляции: -0.438

Квиз

ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Регуляризация, для чего она нужна?

Что такое регуляризация?

Регуляризация L2

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Регуляризация Тихонова (или Ridge regularization, или гребневая регрессия) $R(w) = \|w\|_2 = \sum_{i=1}^{d} w_i^2$

$$Q(w, X) + \lambda||w||^2 \to \min_w.$$

$$||y - Xw||_2^2 + \lambda||w||_2^2 \to min$$

Веса чаще всего не становятся нулевыми - они будут стремиться к 0, но не станут 0

Регуляризация L1

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Регуляризация L1 или Lasso регуляризация

$$R(w) = \|w\|_1 = \sum_{i=1}^{d} |w_i|$$

$$||y - Xw||_2^2 + \lambda||w||_1 \to min$$

$$||w||_1 = |w_1| + |w_2| + \ldots + |w_n|$$

Веса могут стать нулевыми , что полезно для задачи отбора переменных или понижения размерности

Квиз

ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Какая регуляризация чаще всего зануляет коэффициенты?

Метод главных компонент. Простой пример

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Дан предмет в 3D — например, яблоко. Если направить фонарь на яблоко, то на стене будет тень, то есть в реальности у нас 3D-объект, а тень — это 2D-проекция

Мы теряем часть информации (глубину), но основную форму и структуру — сохранили

PCA делает то же самое с многомерными данными — проецирует их в меньшее измерение (2D, 3D и т.д.), чтобы сохранить максимальную суть

Метод главных компонент

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

PCA находит новые признаки, которые представляют собой комбинации старых и при этом несут максимум информации.

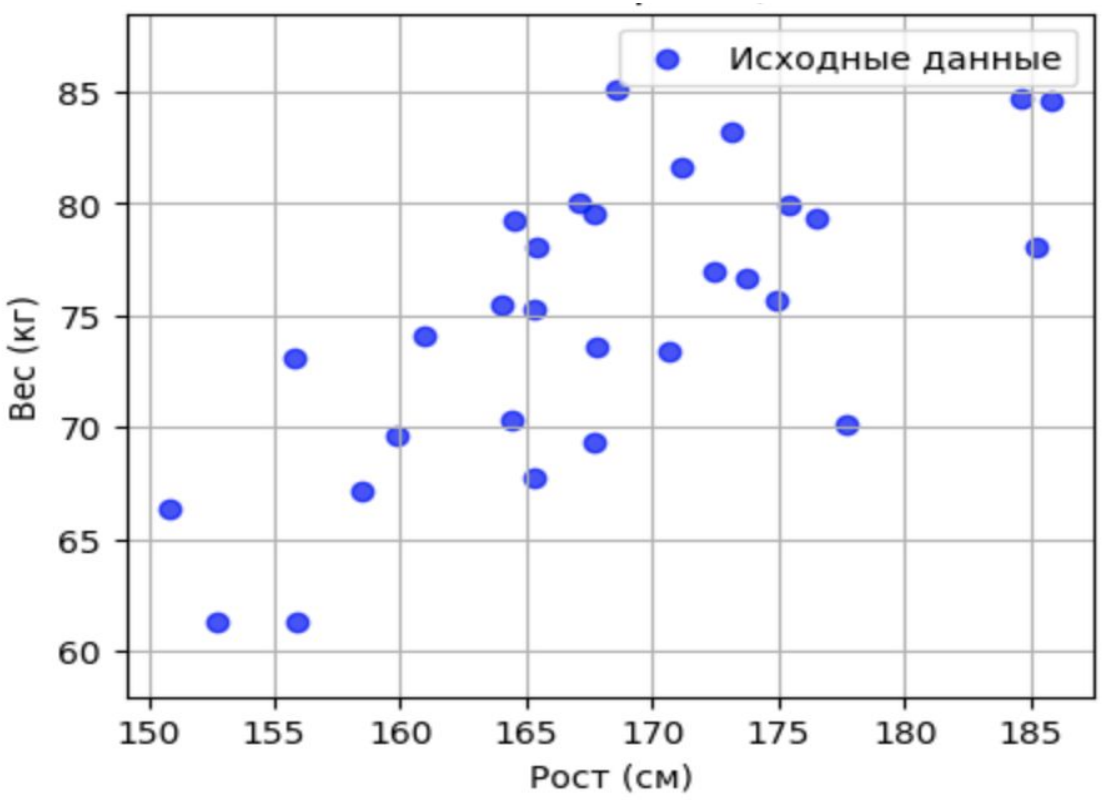
Эти новые признаки называются главными компонентами.

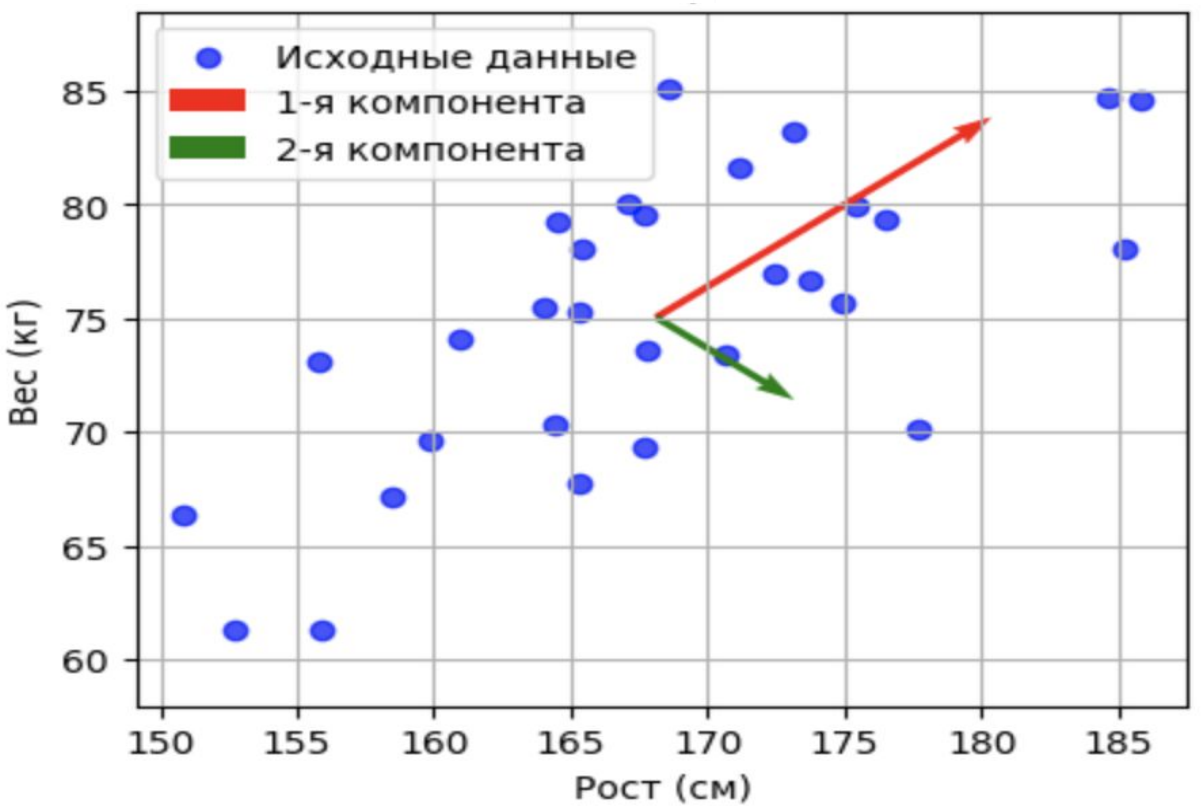
еще один простой пример: вы нашли главные маршруты в большом городе, которые помогают быстро добраться до любого места.

Метод главных компонент. Простой пример

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Чем выше человек, тем он чаще тяжелее, то есть, признаки связаны.
PCA находит такое направление на этом графике, вдоль которого разброс точек максимальный. Это направление и есть **новый признак**, который примерно описывает и **рост, и вес вместе**.





Метод главных компонент

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Цель PCA найти такие направления в пространстве признаков, вдоль которых:

- дисперсия (разброс данных) максимальна
- направления ортогональны друг другу (чтобы не было зависимости между новыми признаками)

Что такое собственные векторы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Если умножать матрицу на вектор, то обычно вектор "поворачивается".

Но иногда — не поворачивается, а просто растягивается или сжимается.

Вот это особенный случай:

$$Av = \lambda v$$

Здесь:

- A — квадратная матрица (например, ковариационная матрица)
- v — **собственный вектор**
- λ — **собственное значение** (на сколько растянули)

Собственный вектор — это направление, которое не меняет направления при действии матрицы, только масштаб.

Что делает PCA и при чём тут собственные векторы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- Строим ковариационную матрицу. Матрица показывает, как связаны признаки между собой:

$$\Sigma = \frac{1}{n-1} X^\top X$$

- Далее ищем собственные векторы этой матрицы

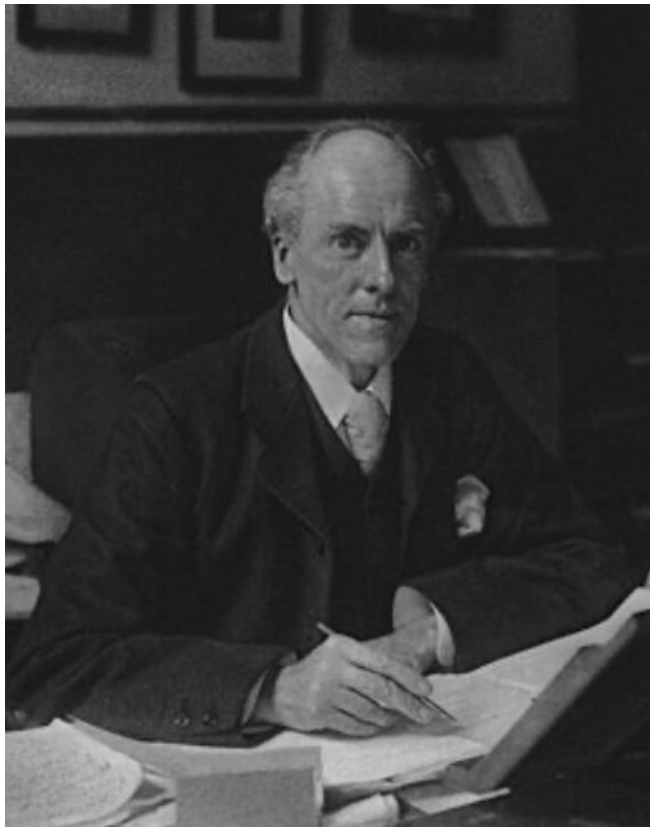
Вот тут и происходит вся магия:

- Собственные векторы v1,v2,… — это новые оси (направления) PCA
- Собственные значения λ1,λ2,… — разброс вдоль этих осей

PCA = выбор собственных векторов ковариационной матрицы как новых осей пространства признаков.

Методы понижения размерности

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС



PCA - один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.

Изобретён Карлом Пирсоном в 1901 году.
Применяется во многих областях, в том числе в эконометрике, биоинформатике, обработке изображений, для сжатия данных, в общественных науках

Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Спасибо за внимание

Метод главных компонент

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Метод главных компонент (Principal Component Analysis или же PCA) — алгоритм **обучения без учителя**, используемый для понижения размерности и выявления наиболее информативных признаков в данных.

Суть метода заключается в предположении о **линейной зависимости данных и их проекции на подпространство ортогональных векторов**, в которых дисперсия будет максимальной.