

# Лекция 8

## Отбор признаков и методы снижения размерности

Юлия Конюшенко

ТГ: [@ko\\_iulia](https://t.me/ko_iulia)

[koniushenko.iun@phystech.edu](mailto:koniushenko.iun@phystech.edu)

## Лекция 1.

ML

→ обучение с учителем

обучение без учителя

- категоризация
- снижение размерн.
- визуализация

- классифр.
- регрессия
- ранжирование

## Лекция 2.

1) линейная регрессия

Обучение  $\equiv$  минимизация MSE

почему именно MSE? → есть вероятностная интерпретация

2) градиентного спуск

$$a(x) = (w, x)$$

$$MSE$$

$$w^{(k)} = w^{(k-1)} - \eta \triangleright Q(w^{(k-1)})$$

стochastic  $\rightarrow$  по 1 объекту  
mini-batch  $\rightarrow$  по батчу

## Лекция 3. Метрики качества и функционалы

$$MSE \rightarrow RMSE \rightarrow R^2$$

ошибки

$$MAE \rightarrow MSLE \rightarrow MAPE \rightarrow SMAPE$$

квантильная регрессия  
онлайн / офлайн / бизнес метрики

Признаки переобучения, регуляризация

- разница качества на train/test
- большие веса

+ оптимумы MSE и

$$MAE$$

среднее медиана

$$\begin{aligned} \cdot L_2: & + \sum w_i^2 \\ \cdot L_1: & + \sum |w_i| \end{aligned}$$

## Лекция 4.

### 1) Оценивание качества модели

- отложенная выборка
- кросс-валидация

K-fold

complete

leave-one-out

K=5

K=7

K=10

### 2) Способы кодирования категориальных признаков

- one-hot encoding

- стеммы
  - сжатие
  - подсет на отложенной выборке

### 3) Линейные модели классификации

$$a(x, w) = \text{sign} \left( \sum_{j=1}^n w_j x_j \right)$$

$$Q(a, X) = \frac{1}{n} \sum_{i=1}^n [M_i < 0] \rightarrow \min, \text{ где } M_i = y_i \cdot (w, x_i) - \text{отступ}$$

Для оптимизации используют верхние оценки эмпирического риска  
 Разные ф-ии потерь соответствуют различным типам моделей

Оптимизируются градиентным спуском

## Лекция 5

1. логистическая перспектива - это линейный классификатор!

$$a(x, w) = \tilde{\sigma}(w^T x), \quad \tilde{\sigma}(z) = \frac{1}{1 + e^{-z}}$$

лог. пот.

$$Q(w) = - \sum_{i=1}^e ([y_i = +1] \log(a(x_i, w)) + [y_i = -1] \log(1 - a(x_i, w)))$$

$$L(a, x) = \sum_{i=1}^e \log(1 + e^{-y_i(a(x_i, w))})$$

! логистическая функция потерь корректно предсказывает вероятности

## 2. Перцептрон

$$a(x, w) = [(w, x) > 0]$$

3. Метод опорных векторов

→ линейно разделимая выборка  
→ линейно неразделимая выборка  
2 задачи оптимизационные

безусловное задание оптимизационное:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^e \max(0, 1 - y_i((w, x_i) + w_0)) \rightarrow \min_{w, w_0}$$

коопротивное между линейной разделяющей поисковой и минимизирующей суммарной ошибкой

## Лекция 6

### 1) Метрики качества

классификации

- accuracy
- матрица ошибок

- precision
- recall

- f-мера

- roc -auc

- roc - кри-  
вая
- pr-rec - кри-  
зис,

### 2) Многоклассовая классификация

- One - vs - all
- all - vs - all

- multiclass vs multi label
- micro и macro усреднение метрик

### 3) Удробные методы

методы решения классификации задачи

## Лекция 7

### 1) Наивный байесовский классификатор

$$P(c|x) = \frac{P(x|c) \cdot P(c)}{P(x)}$$

\* предполагаем, что признаки независимы

### 2) Метод ближайших соседей (KNN)

когда - "схожие" объекты находятся "близко" друг к другу

3) Калибровка вероятностей  
однажды изог бер на отвечах  
тогда

кассировке

## 4) Работа с текстами

- токенизация
- bag of words
- n-gram bag of words
- tf-idf
- word2vec

## 5) Работа с выбросами

статистические  
методы

- правило трех сигм
- интерквартильный размах

→ ML-методы

- isolation forest
- one-class SVM
- с помощью KNN
- local outlier factor

# ПЛАН ЛЕКЦИИ

## 1. Методы отбора признаков

- VarianceThreshold
- Отбор по корреляции с целевой переменной
- Более сложные методы

*а зачем их отбирать?*

## 2. Линейные методы снижения размерности

- Метод главных компонент
- Линейный дискриминантный анализ

# VARIANCE THRESHOLD

- Можем удалить признаки, которые имеют очень маленькую дисперсию, т.е. практически константы.

# ОТБОР ПРИЗНАКОВ ПО КОРРЕЛЯЦИИ С ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- Для каждого признака вычислим его корреляцию с целевой переменной. Будем выкидывать признаки, имеющие маленькую корреляцию.
  - Что такое коррелирование и ковариацию?

# БОЛЕЕ СЛОЖНЫЕ МЕТОДЫ

- **Filtration methods** (фильтрационные методы)
- **Wrapping methods** (оберточные методы)
- **Model selection** (встроенный в модель отбор признаков)

# ФИЛЬТРАЦИОННЫЕ МЕТОДЫ

- Фильтрационные методы - это отбор признаков по различным статистическим тестам. Идея метода состоит в вычислении влияния каждого признака в отдельности на целевую переменную (с помощью вычисления некоторой статистики).

Очевидный плюс метода: скорость, так как мы вычисляем значения  $N$  статистик, где  $N$  - количество признаков.

# ФИЛЬТРАЦИОННЫЕ МЕТОДЫ

В `sklearn` есть сразу несколько методов, использующих отбор по статистическим критериям. Среди них выделим следующие:

- `SelectKBest` - оставляет  $k$  признаков с наибольшим значением выбранной статистики
- `SelectPercentile` - оставляет признаки со значениями выбранной статистики, попавшими в заданную пользователем квантиль
- и другие (см. `sklearn`)

*Это можно считать кроме корреляции?*

# СТАТИСТИЧЕСКИЕ ТЕСТЫ ДЛЯ ОТБОРА ПРИЗНАКОВ

- mutual information:

для векторов X и Y статистика вычисляется по формуле

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$

- хи-квадрат:

$$\chi^2(X; Y) = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

где  $O_{ij}$  - наблюдаемая частота,  $E_{ij}$  - ожидаемая частота.

Какие шиканы у фильтрационных методов?

# ОБЕРТОЧНЫЕ МЕТОДЫ

Оберточные методы используют **жадный отбор признаков**, т.е. последовательно выкидывают наименее подходящие по мнению методов признаки.

В sklearn есть оберточный метод - Recursive Feature Elimination (RFE).

Параметры метода:

- a) алгоритм, используемый для отбора признаков (например, RandomForest)
- b) число признаков, которое мы хотим оставить.

# ЖАДНЫЙ ОТБОР ПРИЗНАКОВ

1 шаг: Перебираем все признаки и убираем тот, удаление которого сильнее всего уменьшает ошибку

2 шаг: Из оставшихся признаков убираем тот, удаление которого сильнее всего уменьшает ошибку

И т.д.

*Недостатки ?*

# ВСТРОЕННЫЕ В МОДЕЛЬ МЕТОДЫ

*Напоминание:*  $L_1$ -регуляризация умеет отбирать признаки.

$$Q(w) + \alpha \sum_{i=1}^d |w_j| \rightarrow \min_w$$

# ВСТРОЕННЫЕ В МОДЕЛЬ МЕТОДЫ

*Напоминание:*  $L_1$ -регуляризация умеет отбирать признаки.

$$Q(w) + \alpha \sum_{i=1}^d |w_j| \rightarrow \min_w$$

Рассмотрим другой вариант регуляризации, которая тоже умеет отбирать признаки ( $L_0$ -регуляризация):

$$Q(w) + \alpha \sum_{i=1}^d [w_j \neq 0] \rightarrow \min_w$$

# ИНФОРМАЦИОННЫЕ КРИТЕРИИ

- Рассмотрим вероятностную постановку задачи: на парах объект-ответ задано некоторое распределение  $p(x, y)$ .
- Функционал ошибки  $Q(a, X)$  соответствует методу максимального правдоподобия, т.е.

$$Q(a, X) = -\ln P = - \sum_{i=1}^l \ln p(x_i, a(x_i)) \rightarrow \min$$

# КРИТЕРИЙ AIC

## Критерий Акаике (AIC, Akaike Information Criterion)

- Дополнительно предполагаем, что модель  $a$  – линейная.

$$AIC(a, X) = Q(a, X) + \frac{2\hat{\sigma}^2}{l} n$$

$\hat{\sigma}^2$  - оценка дисперсии ошибки  $D(y_i - a(x_i))$

$n$  – количество используемых признаков

$l$  – число объектов

# КРИТЕРИЙ AIC

## Критерий Акаике (AIC, Akaike Information Criterion)

- Дополнительно предполагаем, что модель  $a$  – линейная.

$$AIC(a, X) = Q(a, X) + \frac{2\hat{\sigma}^2}{l} n$$

$\hat{\sigma}^2$  - оценка дисперсии ошибки  $D(y_i - a(x_i))$

$n$  – количество используемых признаков

$l$  – число объектов

- Если  $Q$  – среднеквадратичная ошибка для линейной регрессии, и шумы нормально распределены, то

$$AIC = -\ln P + n$$

# КРИТЕРИЙ AIC

## Критерий Акаике (AIC, Akaike Information Criterion)

- Дополнительно предполагаем, что модель  $a$  – линейная.

$$AIC(a, X) = Q(a, X) + \frac{2\hat{\sigma}^2}{l} n$$

$\hat{\sigma}^2$  - оценка дисперсии ошибки  $D(y_i - a(x_i))$

$n$  – количество используемых признаков

$l$  – число объектов

- Если  $Q$  – среднеквадратичная ошибка для линейной регрессии, и шумы нормально распределены, то

$$AIC = -\ln \Pi + n$$

- Часто переходят к задаче максимизации:

$$-\ln \Pi + n \rightarrow \min \Leftrightarrow \ln \Pi - n \rightarrow \max$$

# КРИТЕРИЙ ВІС

Критерий Шварца (BIC, Bayesian Information Criterion)

$$BIC(a, X) = \frac{l}{\hat{\sigma}^2} (Q(a, X) + \frac{\hat{\sigma}^2 l n l}{l} n)$$

$\hat{\sigma}^2$  - оценка дисперсии ошибки  $D(y_i - a(x_i))$

$n$  – количество используемых признаков

$l$  – число объектов

- Если  $Q$  – среднеквадратичная ошибка для линейной регрессии, и шумы нормально распределены, то

$$BIC = -\ln \Pi + \frac{n}{2} l n l$$

- Часто переходят к задаче максимизации:

$$-\ln \Pi + \frac{n}{2} l n l \rightarrow \min \Leftrightarrow \ln \Pi - \frac{n}{2} l n l \rightarrow \max$$

# ОТБОР ПРИЗНАКОВ С ПОМОЩЬЮ ИНФОРМАЦИОННЫХ КРИТЕРИЕВ

- Если в модели  $k$  признаков (регрессоров), то существует  $2^k$  всевозможных моделей
- В идеале необходимо построить все  $2^k$  моделей, для каждой посчитать значение критерия качества (AIC, BIC) и выбрать модель, лучшую по этому критерию
- При большом количестве регрессоров используют метод включений-исключений (жадный отбор признаков)

# ПРИМЕР

Задача предсказания уровня преступности в разных штатах по следующим признакам:

Регрессор
Нулевой коэффициент
Возраст
Южный штат(да/нет)
Образование
Расходы
Труд
Количество мужчин
Численность населения
Безработные (14-24)
Безработные (25-39)
Доход

# ПРИМЕР: ОТБОР ПРИЗНАКОВ ПО AIC

- Мы решаем задачу линейной регрессии с предположением, что ошибки нормально распределены, поэтому  $AIC = \ln P(a, X) - n \rightarrow \max$ .

В модели с полным набором регрессоров  $AIC = -310.37$ . В порядке убывания AIC при удалении каждой из переменных равен:

Численность населения ( $AIC = -308$ ), Труд ( $AIC = -309$ ), Южный штат ( $AIC = -309$ ), Доход ( $AIC = -309$ ), Количество мужчин ( $AIC = -310$ ), Безработные I ( $AIC = -310$ ), Образование ( $AIC = -312$ ), Безработные II ( $AIC = -314$ ), Возраст ( $AIC = -315$ ), Расходы ( $AIC = -324$ ).

Таким образом, имеет смысл удалить переменную “Население”.

## ПРИМЕР: ОТБОР ПРИЗНАКОВ ПО AIC

Южный штат (AIC = -308), Труд (AIC = -308), Доход (AIC = -308), Количество мужчин (AIC = -309), Безработные I (AIC = -309), Образование (AIC = -310), Безработные II (AIC = -313), Возраст (AIC = -313), Расходы (AIC = -329).

Удаляем переменные до тех пор, пока не удастся больше получить увеличения AIC.

Уровень преступности = 1.2 Возраст + 0.75 Образование + 0.87  
Расходы + 0.34 Количество мужчин – 0.86 Безработные I + 2.31  
Безработные II.

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS, PCA)

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Предыдущие методы отбирали из исходных признаков некоторое подмножество признаков.

Теперь мы *хотим придумать новые признаки, каким-то образом выражющиеся через старые, причем новых признаков хочется получить меньше, чем старых*. Сегодня будем рассматривать только случай, когда новые признаки линейно выражаются через старые.

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Постановка задачи:

$x_1, \dots, x_n$  - исходные числовые признаки,  $x_i = f_i(x)$

$z_1, \dots, z_d$  – новые числовые признаки,  $d \leq n$ ,  $z_j = g_j(x)$ .

Хотим:

1. чтобы новые числовые признаки  $z_j$  линейно выражались через исходные признаки  $x_i$
2. чтобы при переходе к новым признакам было потеряно наименьшее количество исходной информации

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Постановка задачи:

$x_1, \dots, x_n$  - исходные числовые признаки,  $x_i = f_i(x)$

$z_1, \dots, z_d$  – новые числовые признаки,  $d \leq n$ ,  $z_j = g_j(x)$ .

Хотим:

1. чтобы новые числовые признаки  $z_j$  линейно выражались через исходные признаки  $x_i$
2. чтобы при переходе к новым признакам было потеряно наименьшее количество исходной информации.

Дисперсия выборки, посчитанная в новых признаках, показывает, как много информации нам удалось сохранить после понижения размерности, поэтому дисперсия в новых признаках должна быть максимальной.

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Новые признаки линейно выражаются через исходные:

$$\begin{cases} z_1 = u_{11}x_1 + \dots + u_{1n}x_n \\ z_2 = u_{21}x_1 + \dots + u_{2n}x_n \\ \dots \\ z_d = u_{d1}x_1 + \dots + u_{dn}x_n \end{cases}$$

Будем искать такие векторы  $u_1, \dots, u_m$ , что они:

- ортогональны:  $(u_i, u_j) = 0$
- нормированы:  $\|u_i\| = 1$

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Новые признаки линейно выражаются через исходные:

$$\begin{cases} z_1 = u_{11}x_1 + \dots + u_{1n}x_n \\ z_2 = u_{21}x_1 + \dots + u_{2n}x_n \\ \dots \\ z_d = u_{d1}x_1 + \dots + u_{dn}x_n \end{cases}$$

Геометрическая интерпретация: новые признаки  $z_i$  – это проекции исходных признаков  $x_i$  на некоторые векторы (компоненты)  $u$ .

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Новые признаки линейно выражаются через исходные:

$$\begin{cases} z_1 = u_{11}x_1 + \dots + u_{1n}x_n \\ z_2 = u_{21}x_1 + \dots + u_{2n}x_n \\ \dots \\ z_d = u_{d1}x_1 + \dots + u_{dn}x_n \end{cases}$$

Геометрическая интерпретация: новые признаки  $z_i$  – это проекции исходных признаков  $x_i$  на некоторые векторы (компоненты)  $u$ .

- Проекция объекта  $x$  на компоненту  $u_i$ :  $z_i = (x, u_i) = u_{i1}x_1 + \dots + u_{in}x_n$
- Проекция всей выборки на компоненту  $u_i$ :  $Z_i = Xu_i$

# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Геометрическая интерпретация: новые признаки  $z_i$  – это проекции исходных признаков  $x_i$  на некоторые векторы (компоненты)  $u$ .

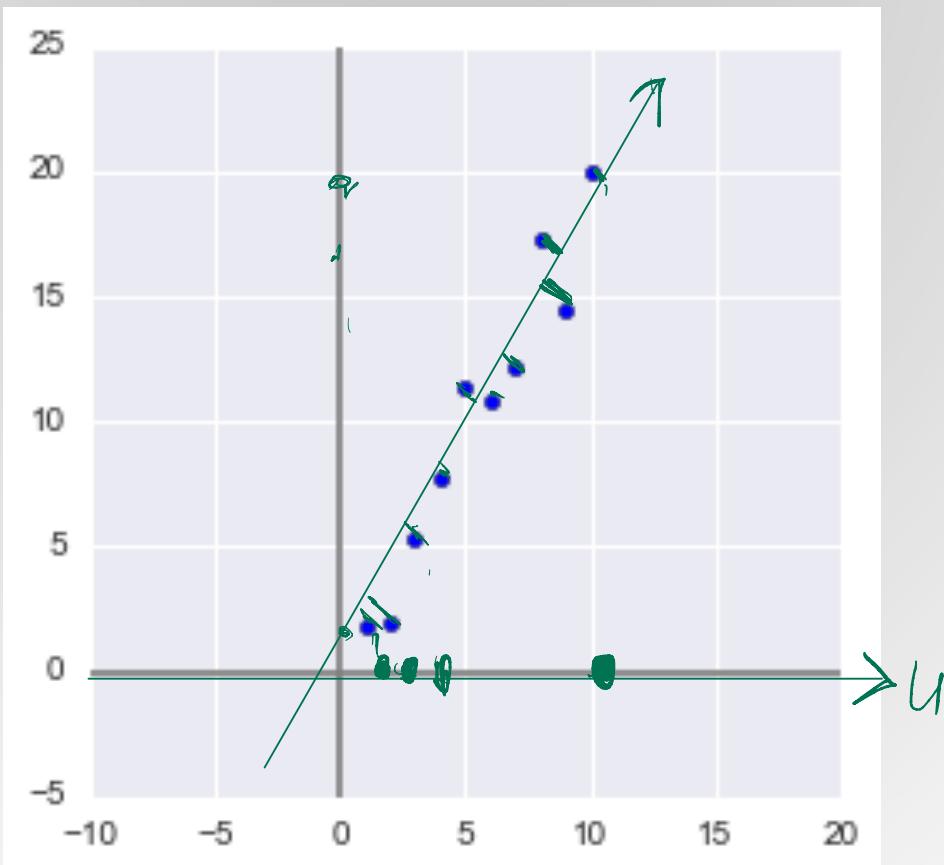
- Проекция объекта  $x$  на компоненту  $u_i$ :  $z_i = (x, u_i) = u_{i1}x_1 + \dots + u_{in}x_n$
- Проекция всей выборки на компоненту  $u_i$ :  $Z_i = Xu_i$

Наша цель: найти такие компоненты  $u_i$ , чтобы дисперсия проекции выборки на них была максимальной:

$$D(Xu_i) \rightarrow \max_{u_i}, \quad i = 1, \dots, d$$

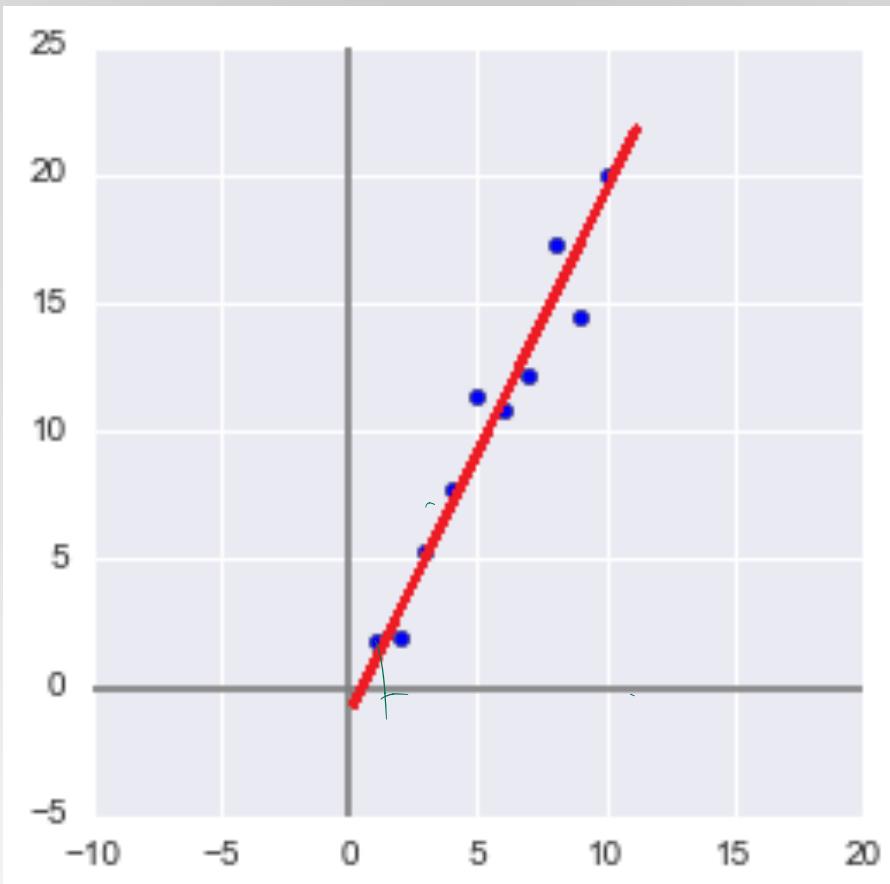
# ПРИМЕР

Хотим спроектировать двумерные данные  $X$  на одномерный вектор  $u$  так, чтобы дисперсия проекции  $Xu$  была максимальной:



# ПРИМЕР

Хотим спроектировать двумерные данные  $X$  на одномерный вектор  $u$  так, чтобы дисперсия проекции  $Xu$  была максимальной:



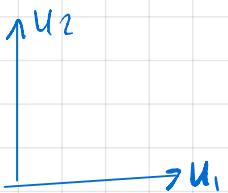
# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Центрируем исходные данные, то есть вычтем из каждого признака его среднее значение.

# ПРОЕКЦИИ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

- Пусть  $X$  – матрица объект-признак для исходных признаков.
- Метод главных компонент делает проекцию исходных объектов на гиперплоскость некоторой размерности  $d$ .

**Теорема (док-во на доске).** Базисные векторы этой гиперплоскости – это собственные векторы матрицы  $X^T X$  (матрица ковариаций), соответствующие  $d$  её наибольшим собственным значениям.



Скалярное произведение  $u_1$ :

$$\begin{cases} \|Xu_1\|^2 \rightarrow \max_{u_1} \\ \|u_1\| = 1 \end{cases}$$

$$\begin{aligned} \|Xu_1\|^2 &= (Xu_1, Xu_1) = (Xu_1)^T \\ &\sim (Xu_1) = u_1^T X^T X u_1 \\ L &= \|Xu_1\|^2 + \lambda (\|u_1\|^2 - 1) \end{aligned}$$

↑  
для ранжирования

$$\frac{\partial L}{\partial u} = 2X^T Xu + 2\lambda u = 0$$

$$X^T Xu = -\lambda u \quad \rightarrow u - \text{собственный}\\ \text{вектор } X^T X$$

$\lambda - ?$

$$\|Xu\|^2 \rightarrow \max$$

$$\|Xu\|^2 = u^T \underbrace{X^T X u}_{\lambda u} = \lambda \|u\|^2 = \lambda \rightarrow \max_u$$

# КОНСТРУКТИВНОЕ ПОСТРОЕНИЕ БАЗИСА В РСА

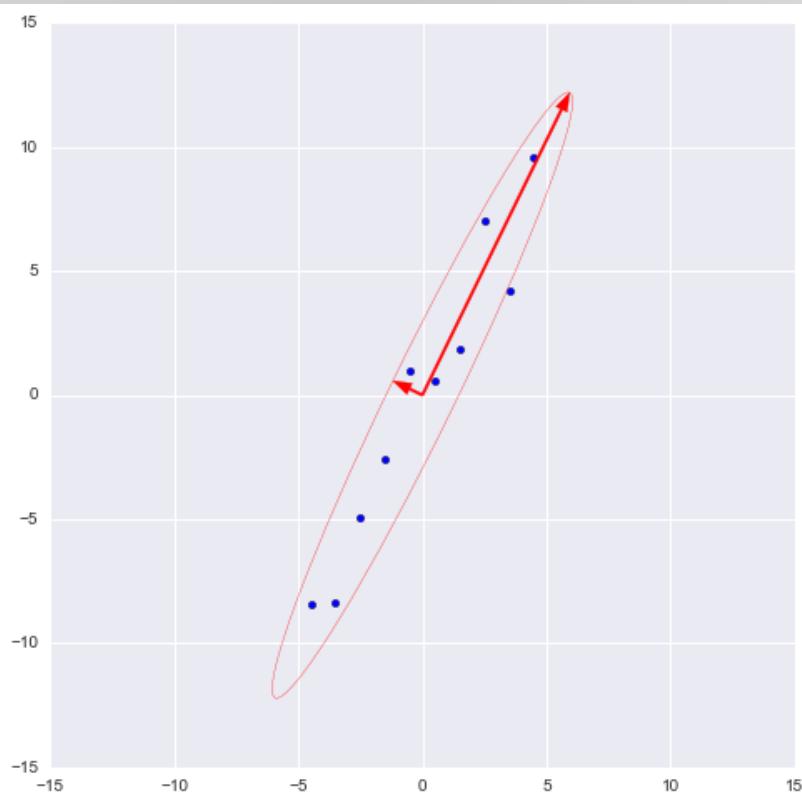
- Находим вектор  $u_1 = \operatorname{argmax}_u(D(Xu))$  и нормируем его:  $u_1 \rightarrow \frac{u_1}{\|u_1\|}$
- Находим вектор  $u_2 = \operatorname{argmax}_u(D(Xu))$  такой, что  $(u_1, u_2) = 0$  и нормируем его:  $u_2 \rightarrow \frac{u_2}{\|u_2\|}$
- Находим вектор  $u_3 = \operatorname{argmax}_u(D(Xu))$  такой, что  $(u_1, u_3) = (u_2, u_3) = 0$  и нормируем его:  $u_3 \rightarrow \frac{u_3}{\|u_3\|}$ .

И т.д.

Получаем ортонормированный базис  $\{u_1, u_2, \dots, u_d\}$ .

# ГЕОМЕТРИЧЕСКИЙ СМЫСЛ РСА

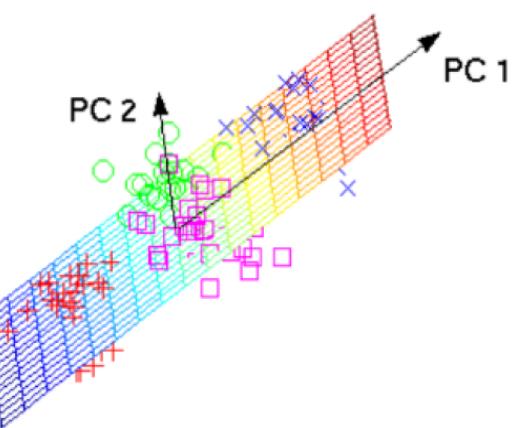
- Нахождение собственных векторов матрицы  $X^T X$  позволяет нам аппроксимировать исходные данные эллипсоидом, натянутым на эти векторы



- Затем мы делаем проекцию на подпространство, натянутое на собственные векторы с наибольшими собственными значениями

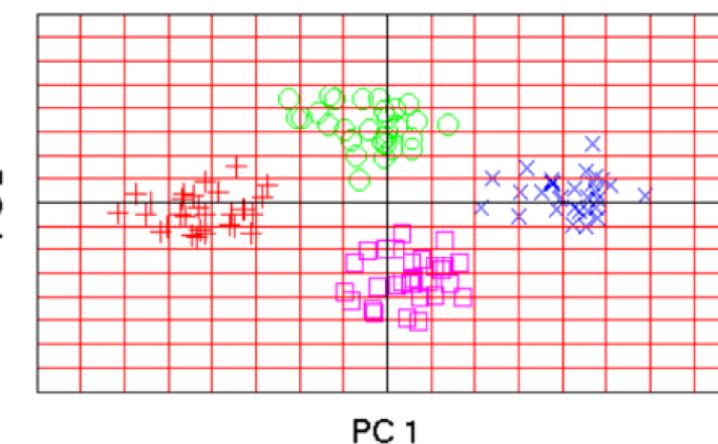
# ПРОЕКЦИЯ НА ГИПЕРПЛОСКОСТЬ

original data space



PCA

component space

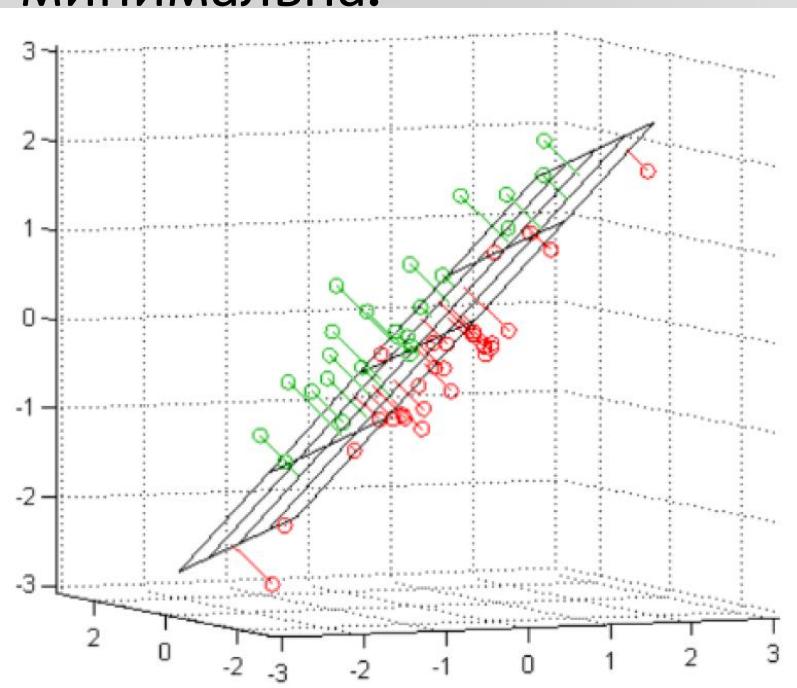


## АЛЬТЕРНАТИВНАЯ ПОСТАНОВКА ЗАДАЧИ

Найти новые признаки  $Z$  и матрицу проецирования  $U$ ,  
наилучшим образом восстанавливающие исходные

признаки:  $\|X - ZU^T\|^2 \rightarrow \min_{Z, U}$

Геометрически это эквивалентно нахождению  
гиперплоскости, сумма квадратов расстояний от которой до  
точек выборки минимальна:



# ДОЛЯ ОБЪЯСНЕННОЙ ДИСПЕРСИИ

- Упорядочим собственные значения матрицы  $X^T X$  по убыванию:  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ .
- Доля дисперсии, объяснённой  $j$ -й компонентой (explained variance ratio):

$$\delta_j = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$$

- Доля дисперсии, объясняемой первыми  $k$  компонентами:

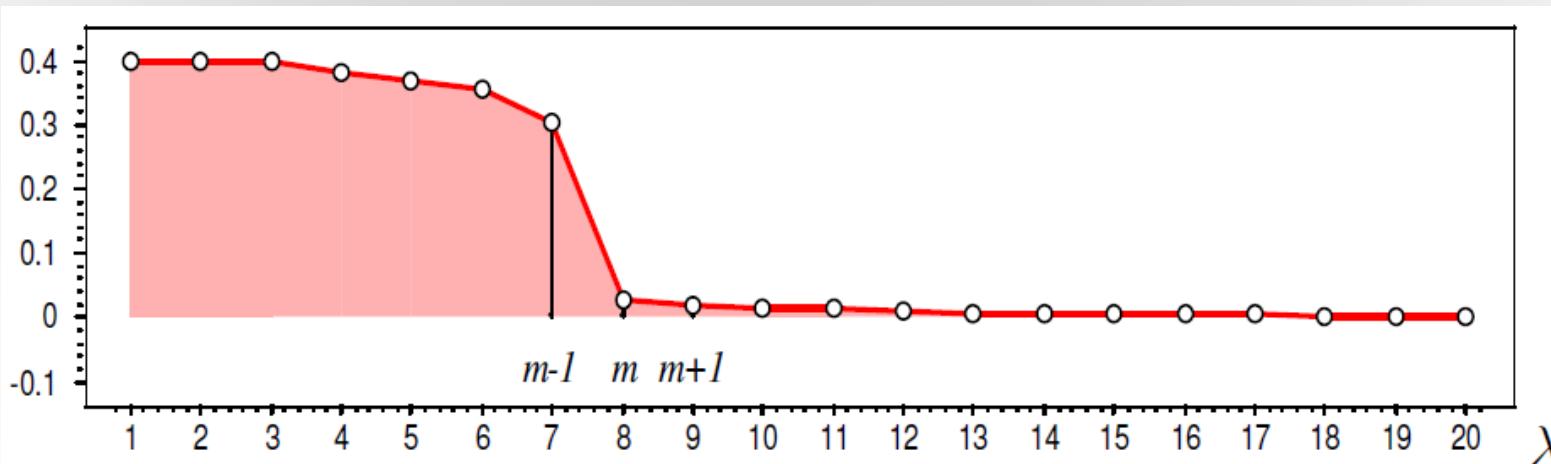
$$\delta = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_n} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}$$

# ВЫБОР ЧИСЛА ГЛАВНЫХ КОМПОНЕНТ

- Эффективная размерность выборки – это наименьшее целое  $m$ , при котором *доля необъясненной дисперсии*

$$E_m = \frac{\|ZU^T - X\|^2}{\|X\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\sum_{i=1}^n \lambda_i} \leq \varepsilon$$

Критерий крутого склона:



# ПРИМЕР: FACES DATASET



# FACES DATASET (ГЛАВНЫЕ КОМПОНЕНТЫ)



# ВОССТАНОВЛЕННОЕ ИЗОБРАЖЕНИЕ

#efaces=1, res=57.804

#efaces=2, res=57.611

#efaces=5, res=54.054

#efaces=10, res=52.01

#efaces=20, res=45.897



#efaces=40, res=35.868

#efaces=60, res=29.624

#efaces=80, res=24.103

#efaces=100, res=20.317

#efaces=150, res=16.154



#efaces=200, res=13.257

#efaces=300, res=9.581

#efaces=400, res=6.908

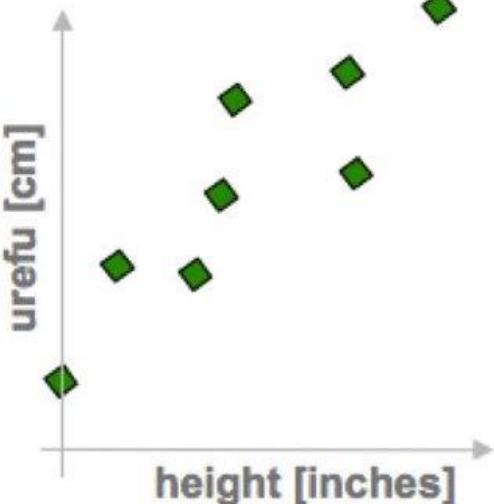
#efaces=1000, res=0.924

#efaces=1071, res=0.653

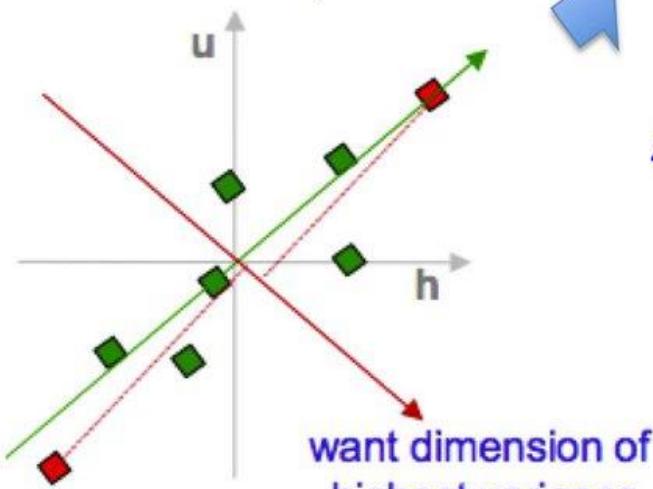


# PCA in a nutshell

1. correlated hi-d data  
("urefu" means "height" in Swahili)



2. center the points



3. compute covariance matrix

$$\begin{matrix} h & u \\ \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \end{matrix} \rightarrow \text{cov}(h, u) = \frac{1}{n} \sum_{i=1}^n h_i u_i$$

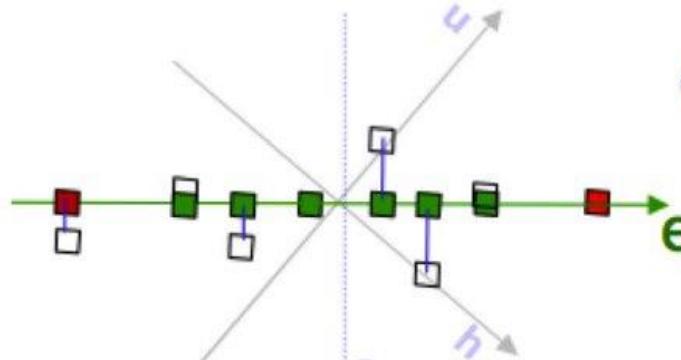
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

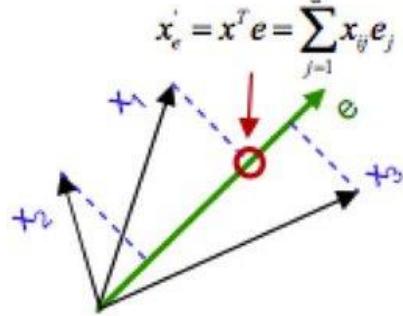
$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

`eig(cov(data))`

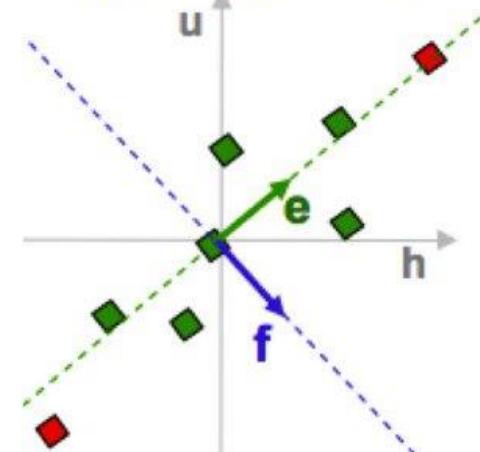
7. uncorrelated low-d data



6. project data points to those eigenvectors



5. pick  $m < d$  eigenvectors w. highest eigenvalues



# СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ (SINGULAR VALUE DECOMPOSITION, SVD)

**Теорема.** Матрицу  $A \in \mathbb{R}^{m \times n}$  можно представить в виде

$$A = U\Sigma V^T,$$

- где  $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$  - ортогональные матрицы,
- $\Sigma \in \mathbb{R}^{m \times n}$  - диагональная матрица с ненулевыми элементами  $\sigma_i = \sqrt{\lambda_i}$ , где  $\lambda_i$  - собственные значения матрицы  $A^T A$ .

# СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ (SVD)

**Теорема.** Матрицу  $A \in \mathbb{R}^{m \times n}$  можно представить в виде

$$A = U\Sigma V^T,$$

- где  $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$  - ортогональные матрицы,
- $\Sigma \in \mathbb{R}^{m \times n}$  - диагональная матрица с ненулевыми элементами  $\sigma_i = \sqrt{\lambda_i}$ , где  $\lambda_i$  - собственные значения матрицы  $A^T A$ .

При этом

- Столбцы матрицы  $U$  являются собственными векторами матрицы  $AA^T$
- Столбцы матрицы  $V$  являются собственными векторами матрицы  $A^T A$ .

# SINGULAR VALUE DECOMPOSITION

- При  $m \leq n$ :

$$\begin{matrix} m \times n \\ A \end{matrix} = \begin{matrix} m \times m \\ U \end{matrix} \cdot \begin{matrix} m \times n \\ \Sigma \end{matrix} \cdot \begin{matrix} n \times n \\ V^T \end{matrix}$$

$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$

- При  $m > n$ :

$$\begin{matrix} m \times n \\ A \end{matrix} = \begin{matrix} m \times m \\ U \end{matrix} \cdot \begin{matrix} m \times n \\ \Sigma \end{matrix} \cdot \begin{matrix} n \times n \\ V^T \end{matrix}$$

# СВЯЗЬ SVD И РСА

Пусть  $X$  – матрица объект-признак, для которой мы хотим снизить размерность и  $X = U\Sigma V^T$  её SVD-разложение.

Тогда:

- Столбцы матрицы  $V$  – это собственные векторы матрицы  $X^T X$ , т.е. векторы  $v_1, \dots, v_n$  – главные компоненты.

- Столбцы матрицы  $U\Sigma$  – это новые признаки, то есть, проекции исходных признаков на главные компоненты  
 $Z = Xv$

$$(X = U\Sigma V^T \Leftrightarrow U\Sigma = XV).$$

- Сингулярные числа матрицы  $\Sigma$  – это корни из собственных чисел матрицы  $X^T X$ .

# СВЯЗЬ SVD И РСА

- Столбцы матрицы  $V$  – это собственные векторы матрицы  $X^T X$ , т.е. векторы  $v_1, \dots, v_n$  – главные компоненты.
- Столбцы матрицы  $U\Sigma$  – это новые признаки  $z = Xv$  ( $X = U\Sigma V^T \Leftrightarrow U\Sigma = XV$ ).
- Сингулярные числа матрицы  $\Sigma$  – это корни из собственных чисел матрицы  $X^T X$ .

Для снижения размерности берем первые  $k$  столбцов матрицы  $U$  и верхний  $k \times k$ -квадрат матрицы  $\Sigma$ , тогда матрица  $U_k \Sigma_k$  содержит  $k$  новых признаков, соответствующих первым  $k$  главным компонентам.

# ПОСТРОЕНИЕ СИНГУЛЯРНОГО РАЗЛОЖЕНИЯ

Ищем сингулярное разложение:  $X = U\Sigma V^T$

- Сингулярные числа матрицы  $\Sigma$  – это корни из собственных чисел матрицы  $X^T X$

$\Rightarrow$  находим  $\lambda_1 \geq \dots \geq \lambda_k$  собственные числа матрицы  $X^T X$  и получаем матрицу  $\Sigma$  – у которой на диагонали стоят  $\sigma_1 = \sqrt{\lambda_1}, \dots, \sigma_k = \sqrt{\lambda_k}$ .

- Столбцы матрицы  $V$  – это собственные векторы матрицы  $X^T X \Rightarrow$  находим собственные векторы  $v_i$ :  $(X^T X - \lambda_i I)v_i = 0$ .
- Столбцы матрицы  $U\Sigma$  – это векторы  $Xv_1, Xv_2, \dots$ , т.е.

$$\sigma_i u_i = X v_i \Rightarrow u_i = \frac{1}{\sigma_i} X v_i$$

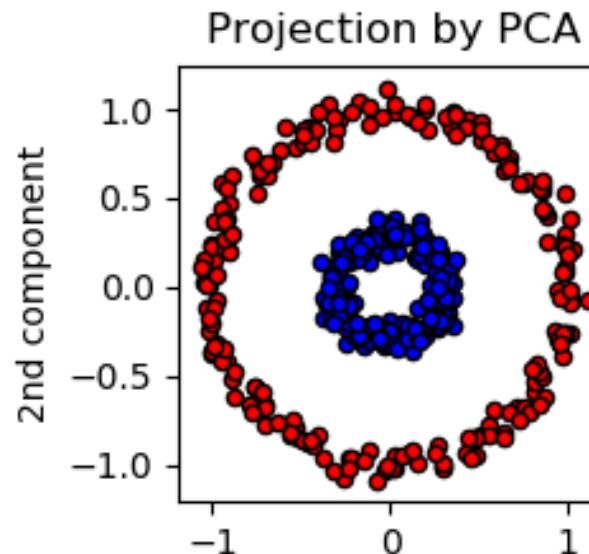
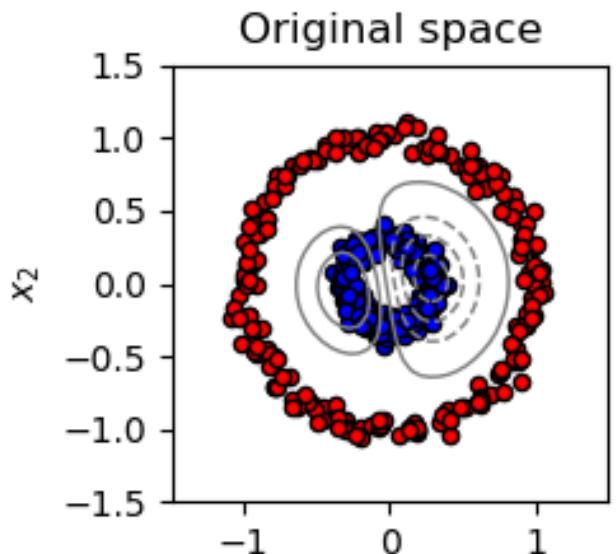
(либо находим  $u_i$ :  $(X X^T - \lambda_i I)u_i = 0$ )

# ЯДРОВЫЙ МЕТОД ГЛАВНЫХ КОМПОНЕНТ

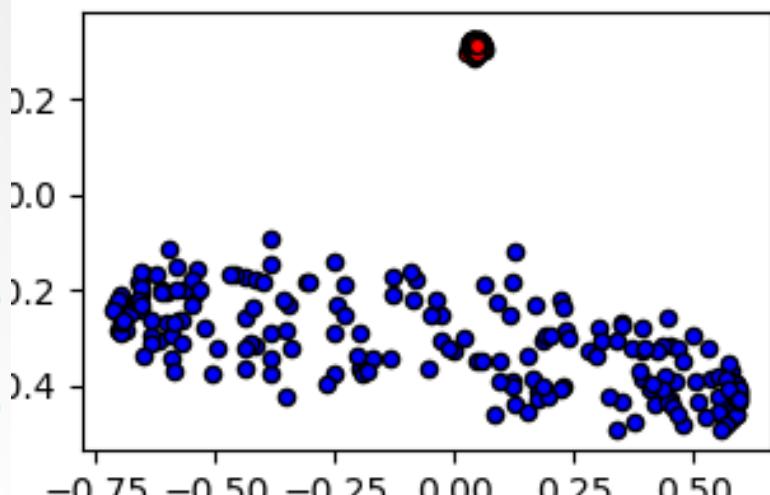
Перейдем к новым признакам  $x \rightarrow \varphi(x)$  ( $X \rightarrow \Phi$ )

- В методе главных компонент мы получили, что новые признаки:  $z_i = (x, u_i)$ , где  $u_1, \dots, u_n$  - собственные векторы матрицы  $X^T X$  (главные компоненты).
- В пространстве новых признаков:  $z_i = (\varphi(x), u_i)$ , где  $u_1, \dots, u_n$  - собственные векторы матрицы  $\Phi^T \Phi$ .

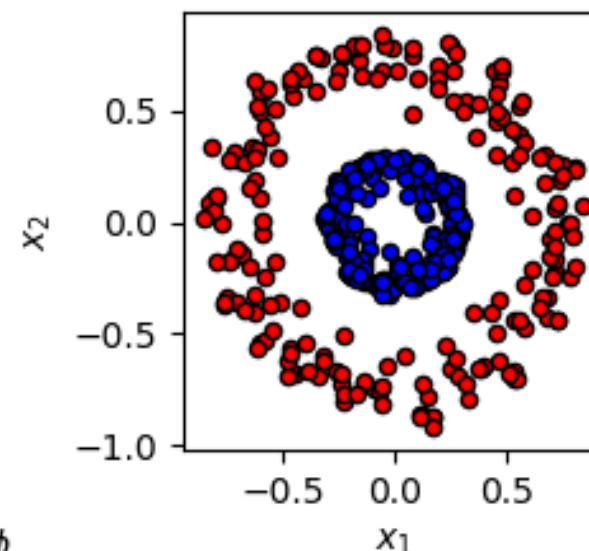
# ПРИМЕР KERNEL PCA (SCIKIT-LEARN)



Projection by KPCA



Original space after inverse transform



# ЯДРОВЫЙ МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Перейдем к новым признакам  $x \rightarrow \varphi(x)$  ( $X \rightarrow \Phi$ )

- В методе главных компонент мы получили, что главные компоненты:  $z_i = (x, u_i)$ , где  $u_1, \dots, u_n$  - собственные векторы матрицы  $X^T X$  (главные компоненты).
- В пространстве новых признаков:  $z_i = (\varphi(x), u_i)$ , где  $u_1, \dots, u_n$  - собственные векторы матрицы  $\Phi^T \Phi$ .

**Утверждение.**  $z_j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^l v_{ji} K(x_i, x)$ , где  $v_j$  - собственный вектор матрицы  $K = \Phi \Phi^T$ .

Т.е. мы можем вычислять проекции, не используя напрямую признаковое описание объектов в новом пространстве.