Занятие 10
Методы снижения размерности и поиск аномалий

Блуменау М.И.

На основе материалов Кантонистовой Е.О.

ВШЭ, 2025

Метод главных компонент

Предыдущие методы отбирали из исходных признаков некоторое подмножество признаков. Теперь мы *хотим придумать новые признаки, каким-то образом выражающиеся 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, $x_i = f_i(\boldsymbol{x})$

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

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:

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

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:

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

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

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Новые признаки линейно выражаются через исходные:

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

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

- Проекция объекта $\boldsymbol{x}$ на компоненту u_i: $z_i = (x, u_i) = u_{i1}x_1 + \cdots + u_{in}x_n$
- 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.

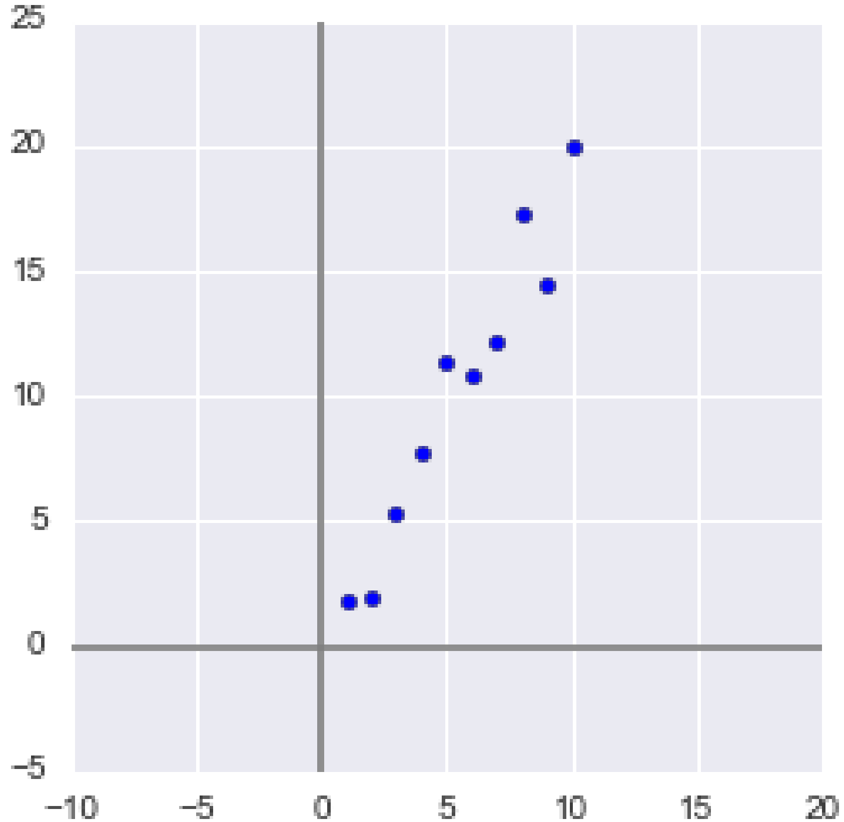
- Проекция объекта $\boldsymbol{x}$ на компоненту u_i: $z_i = (x, u_i) = u_{i1}x_1 + \cdots + u_{in}x_n$
- Проекция всей выборки на компоненту u_i: $Z_i = Xu_i$

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

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

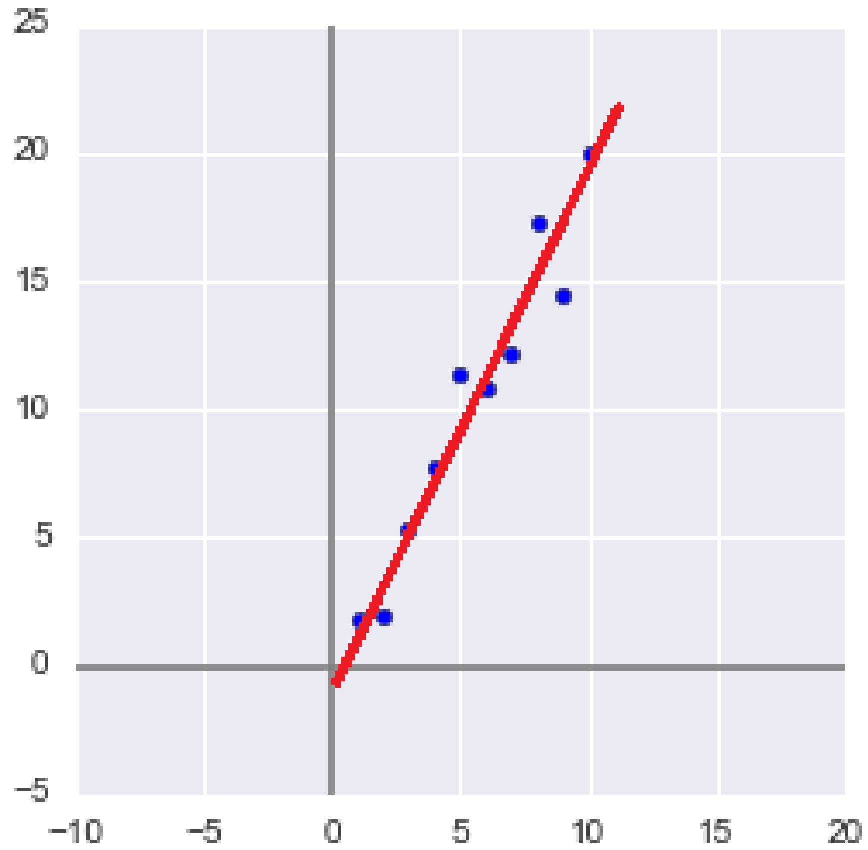
ПРИМЕР

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



ПРИМЕР

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



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. Базисные векторы этой гиперплоскости – это собственные векторы матрицы $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ (матрица ковариаций), соответствующие d её наибольшим собственным значениям.

Конструктивное построение базиса в PCA

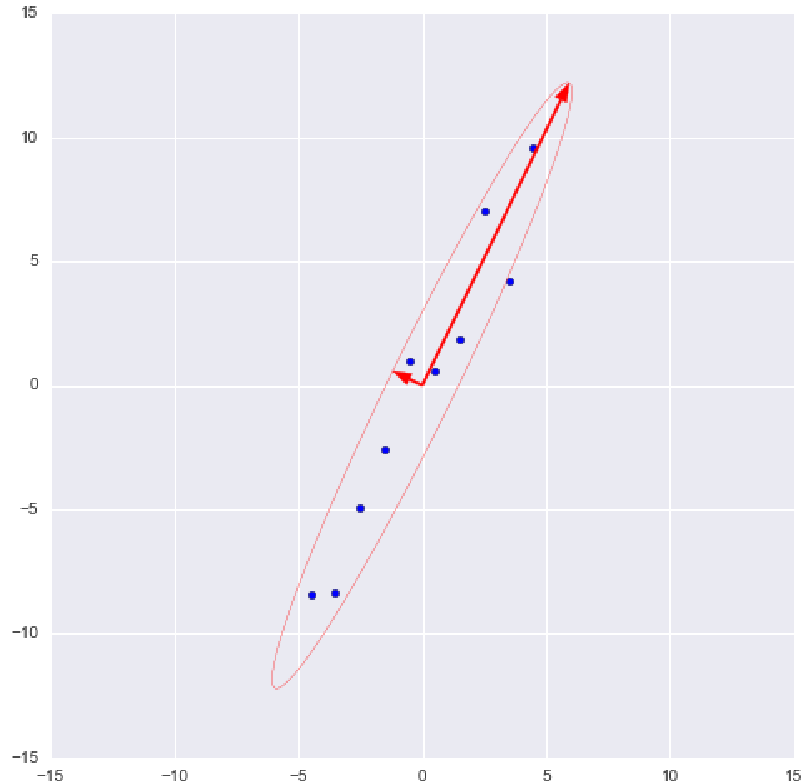
- Находим вектор $u_1 = argmax_u\big(D(Xu)\big)$ и нормируем его: $u_1 \rightarrow \frac{u_1}{||u_1||}$
- Находим вектор $u_2 = argmax_u\big(D(Xu)\big)$ такой, что $(u_1, u_2) = 0$ и нормируем его: $u_2 \rightarrow \frac{u_2}{||u_2||}$
- Находим вектор $u_3 = argmax_u\big(D(Xu)\big)$ такой, что $(u_1,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

И т.д.

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

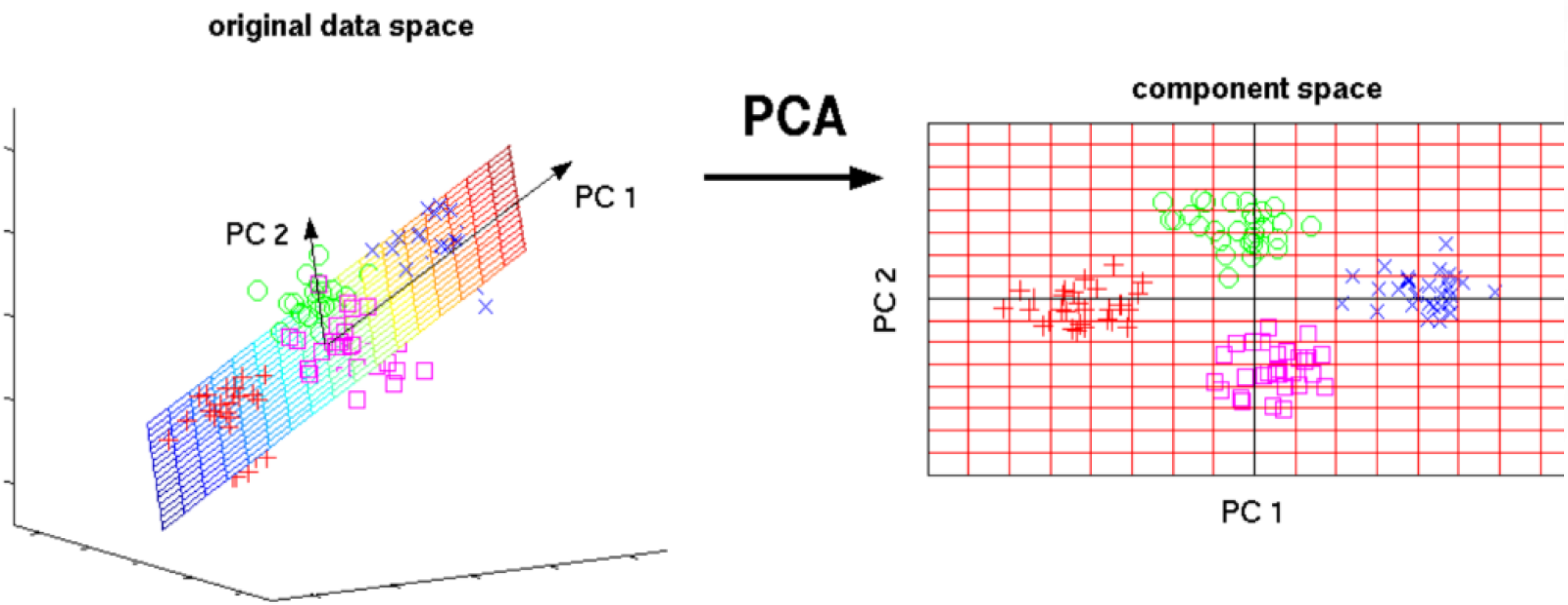
ГЕОМЕТРИЧЕСКИЙ СМЫСЛ PCA

- 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

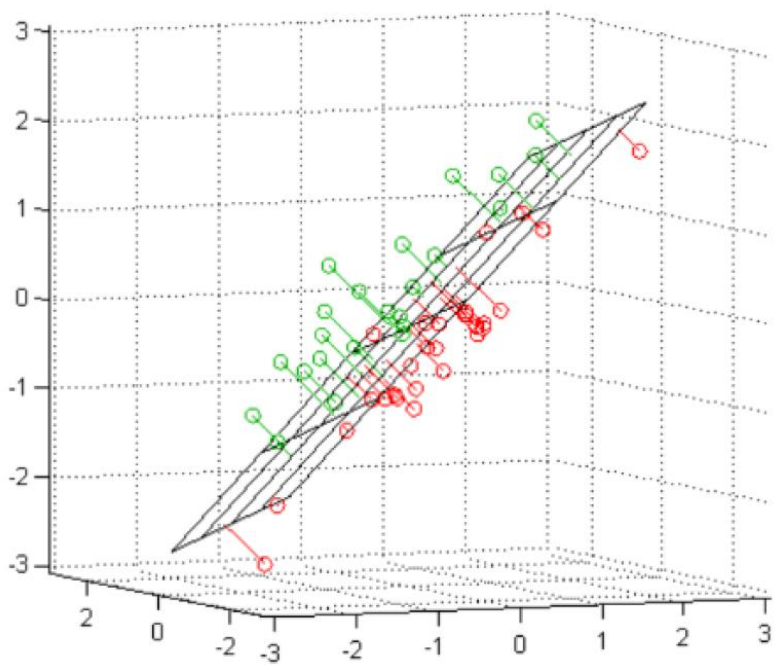


АЛЬТЕРНАТИВНАЯ ПОСТАНОВКА ЗАДАЧИ

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

$$\left\|X - ZU^T\right\|^2 \to \min_{Z,U}$$

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: $\lambda_1 \geq \lambda_2 \geq \cdots > \lambda_n \geq 0.$

- Доля дисперсии, объяснённой j-й компонентой (explained variance ratio):

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

- Доля дисперсии, объясняемой первыми k компонентами:

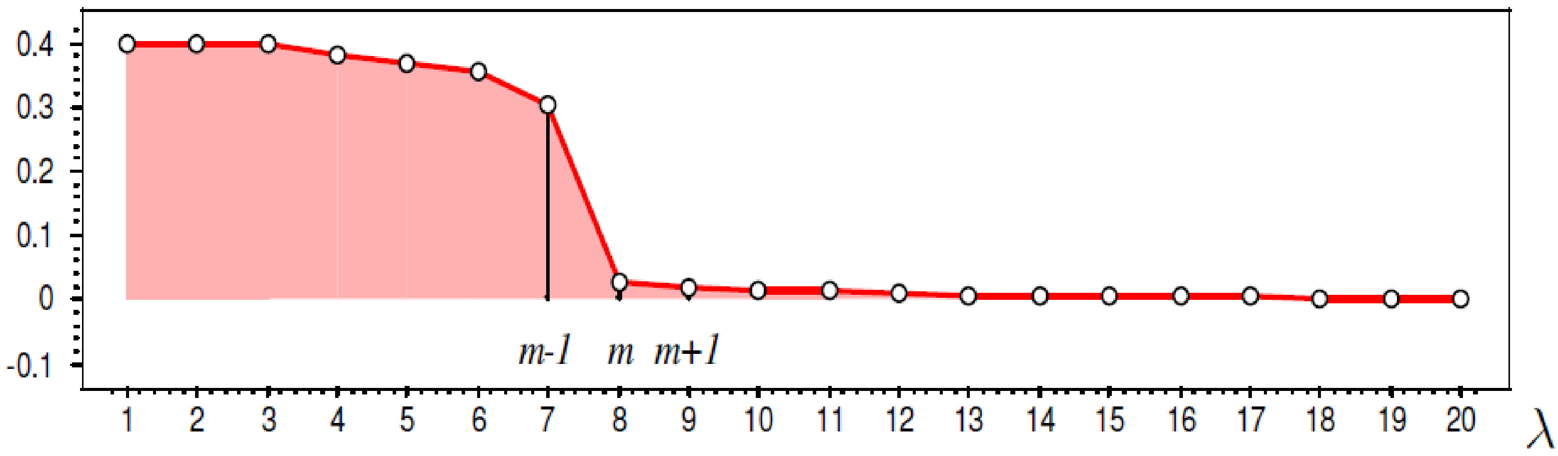
$$\boldsymbol{\delta = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_n}}$$

ВЫБОР ЧИСЛА ГЛАВНЫХ КОМПОНЕНТ

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

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

Критерий крутого склона:

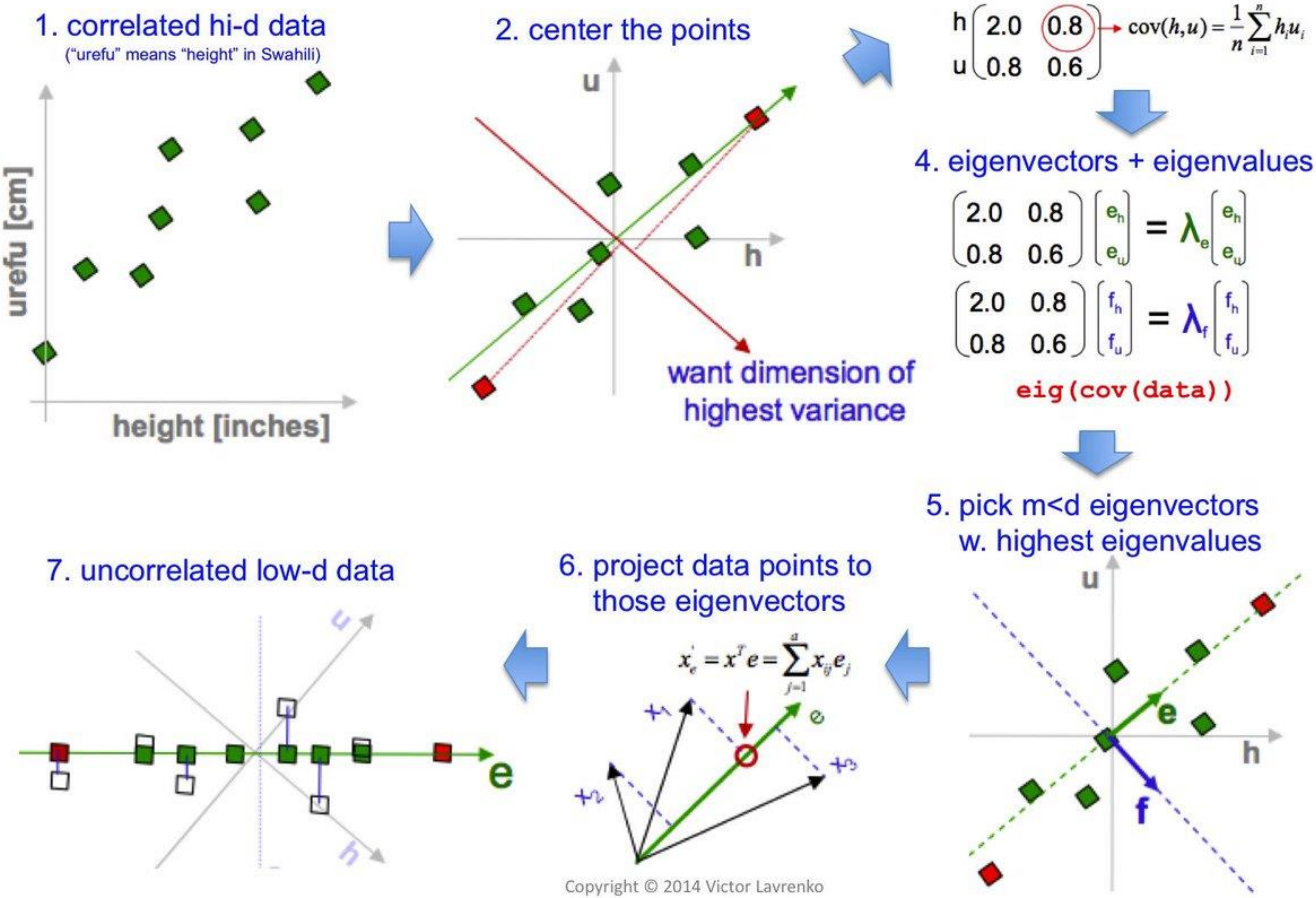


ПРИМЕР: FACES DATASET

FACES DATASET (ГЛАВНЫЕ КОМПОНЕНТЫ)

ВОССТАНОВЛЕННОЕ ИЗОБРАЖЕНИЕ

PCA in a nutshell



СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ (SINGULAR VALUE DECOMPOSITION , SVD)

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

$$\boldsymbol{A = U\Sigma V^T},$$

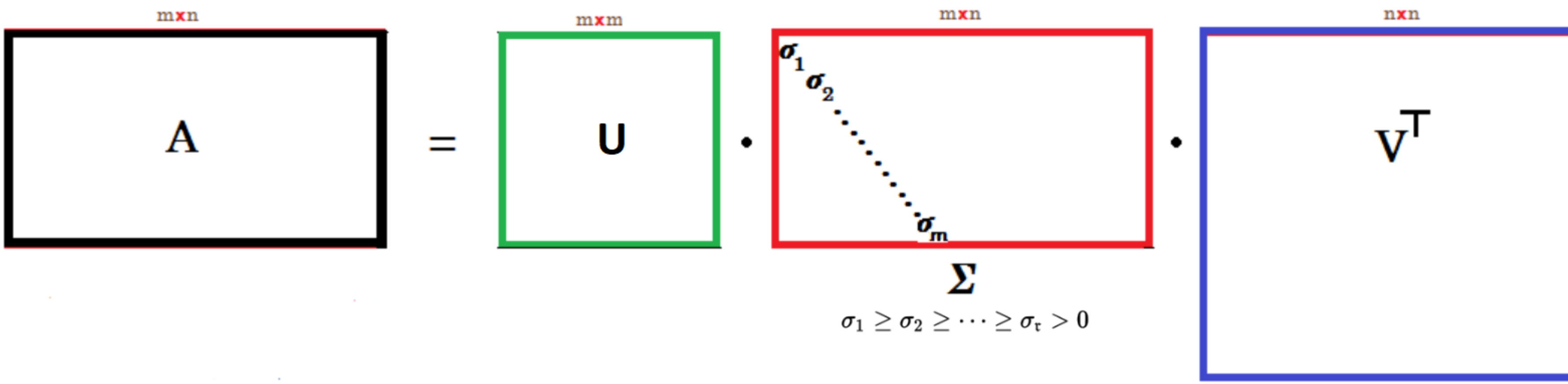
- где $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ - ортогональные матрицы,
- $\Sigma \in \mathbb{R}^{m \times n}$ - диагональная матрица с ненулевыми элементами $\sigma_i = \sqrt{\lambda_i}$, где λ_i - собственные значения матрицы $A^T A$.

При этом

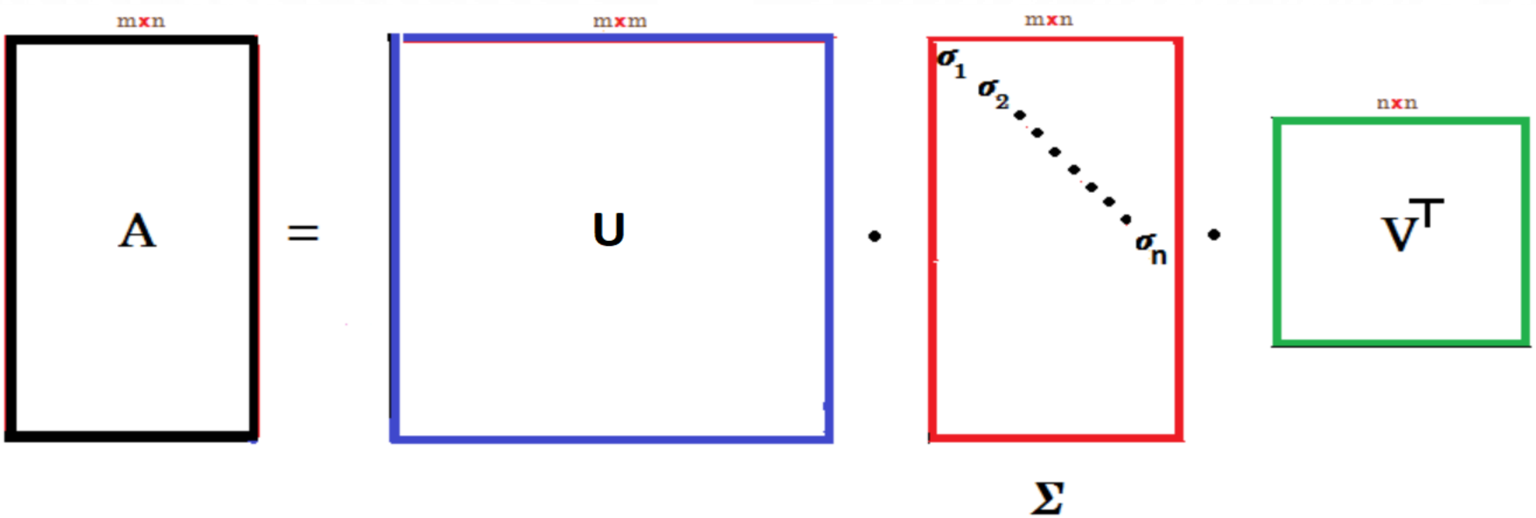
- Столбцы матрицы U являются собственными векторами матрицы AA^T
- Столбцы матрицы V являются собственными векторами матрицы $A^T A$.

SINGULAR VALUE DECOMPOSITION

- При $m \leq n$:



- При $m > n$:



СВЯЗЬ SVD И PCA

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

Тогда:

- Столбцы матрицы $\boldsymbol{V}$ – это собственные векторы матрицы $X^T X$, т.е. векторы $v_1, \dots, v_n$ − главные компоненты.
- Столбцы матрицы $\boldsymbol{U\Sigma}$ – это новые признаки, то есть, проекции исходных признаков на главные компоненты $Z = Xv$

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

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

СВЯЗЬ SVD И PCA

- Столбцы матрицы V – это собственные векторы матрицы $X^T X$, т.е. векторы $v_1, \ldots, v_n$ $-$ главные компоненты.
- Столбцы матрицы $U\Sigma$ – это новые признаки $z = Xv$

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

- Сингулярные числа матрицы Σ – 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 Σ, 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 Σ – это корни из собственных чисел матрицы $X^T X$

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

- 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, …, т.е.

$$\sigma_i u_i = Xv_i \Rightarrow u_i = \frac{1}{\sigma_i} Xv_i$$

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

ПОИСК АНОМАЛИЙ

ПОИСК АНОМАЛИЙ С ПОМОЩЬЮ МОДЕЛЕЙ ML

Идея: *можно настроить модель машинного обучения так, чтобы на нормальных объектах она принимала значения, близкие к нулю (или, например, положительные значения). Тогда если прогноз на объекте сильно отличается от прогноза на обучающей выборке, то такой объект можно считать аномальным.*

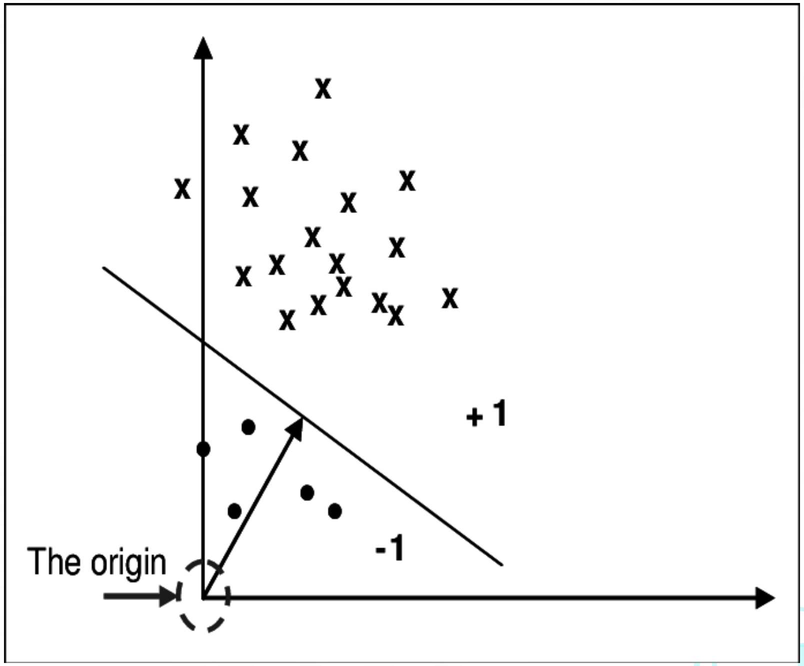
ONE-CLASS SVM

Метод строит линейную функцию $a(x) = sign(w, x)$ так, чтобы она отделяла выборку от начала координат с максимальным отступом, а именно:

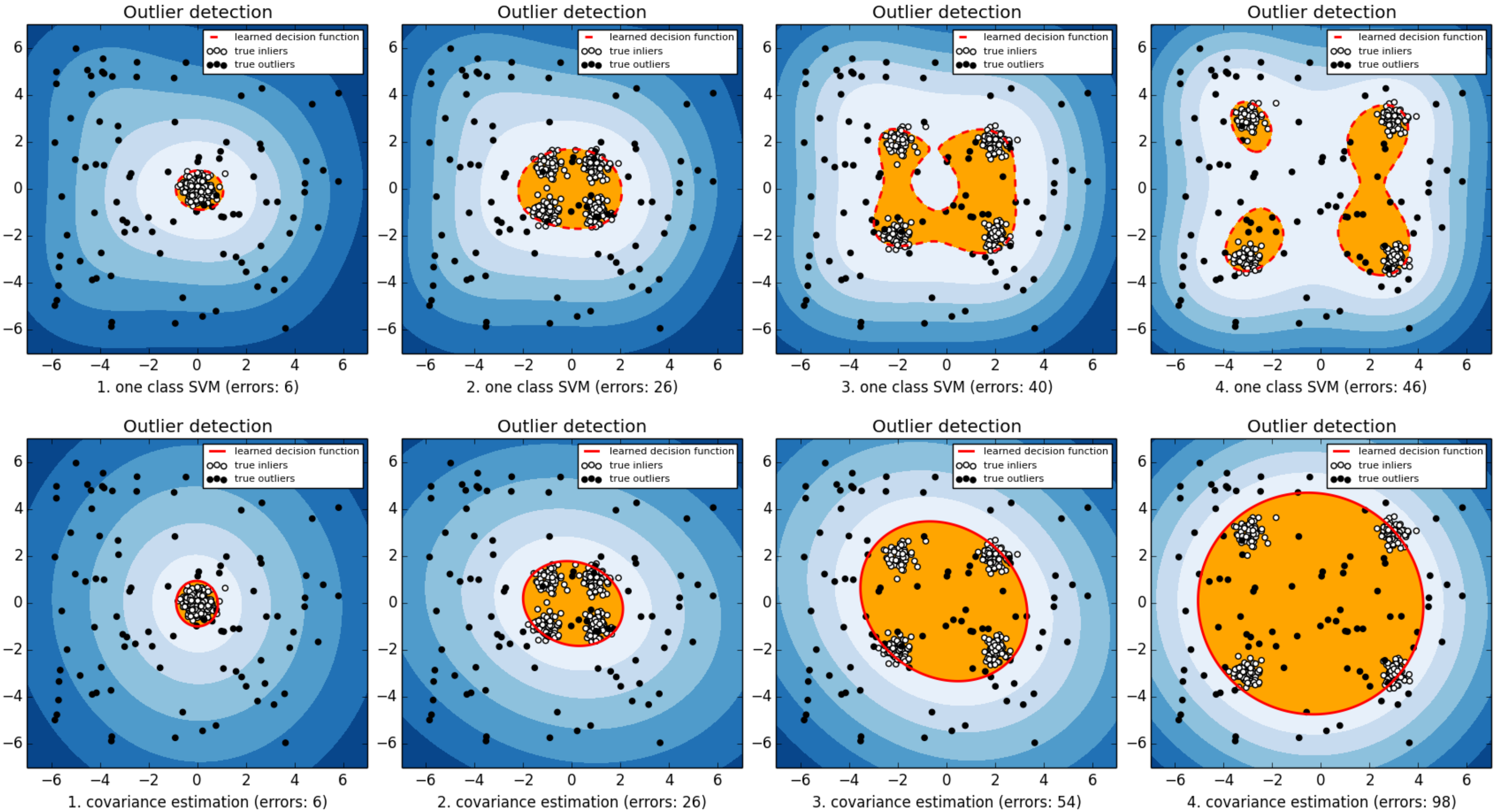
- $a(x)$ отделяет как можно больше объектов выборки от нуля: $a(x) = +1$ на области как можно меньшего объема, содержащей как можно больше объектов выборки
- имеет большой отступ от 0.

Тогда объекты с $a(x) = -1$

– это аномалии.

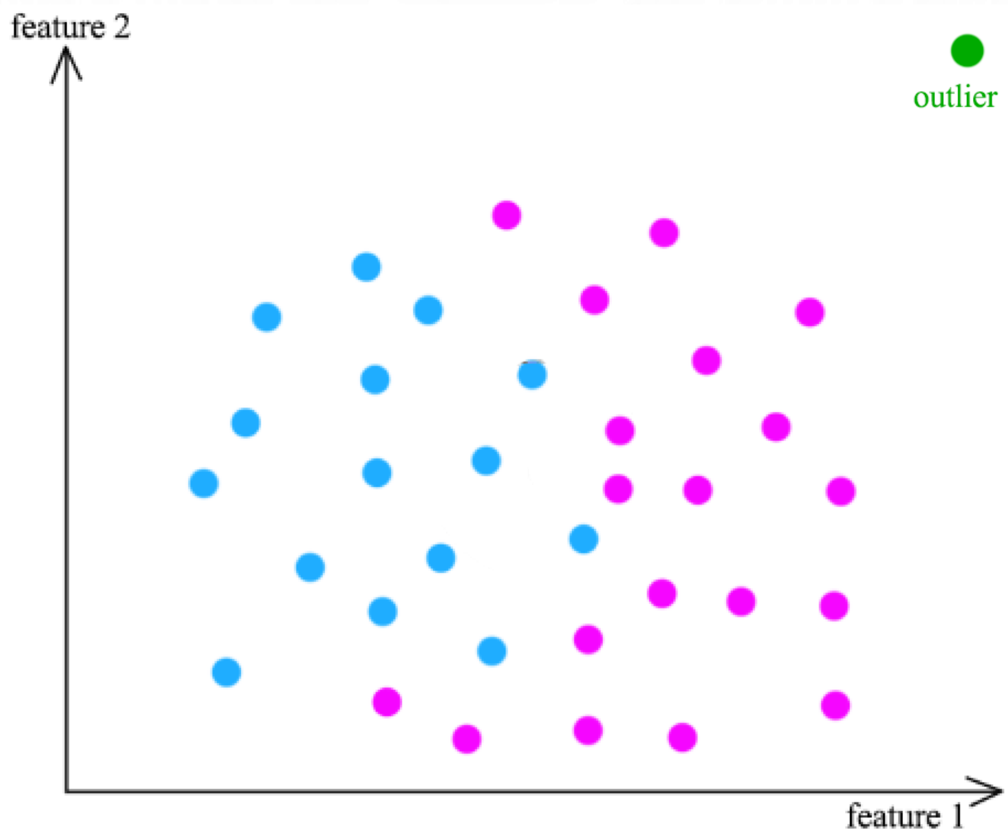


ONE-CLASS SVM С RBF-ЯДРОМ



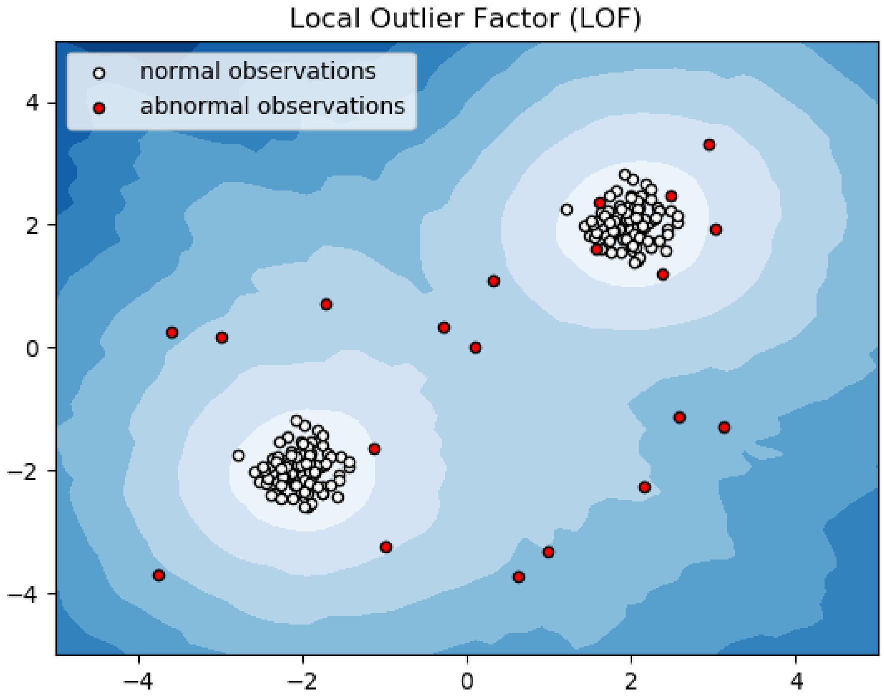
ПОИСК ВЫБРОСОВ С ПОМОЩЬЮ KNN

- Вычисляем среднее расстояние от каждой точки до её ближайших k соседей
- Точки с наибольшим средним расстоянием – выбросы



LOCAL OUTLIER FACTOR

- Задаем плотность распределения в точке, используя k ближайших соседей
- Точки, плотность распределения в которых значительно меньше, чем у соседей – выбросы.



ССЫЛКИ

- https://dyakonov.org/2017/04/19/поиск-аномалий-anomaly-detection/
- https://scikit-learn.org/stable/modules/outlier_detection.html
- https://github.com/yzhao062/pyod