Понижение размерности

2021

Мврия Корлякова

ПЛАН

1. Селекция признаков
2. Редукция системы признаков

Постановка задачи

Задача обучения с учителем

$$X = (x^i, y_i)_{i=1}^l$$

$$a(x) = w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_d x^d$$

$$a(x) = sign(w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_d x^d)$$

$$a(x) :$$

$$Q(a, X) \to min$$

Этапы формирования модели

Анализ данных и их очистка

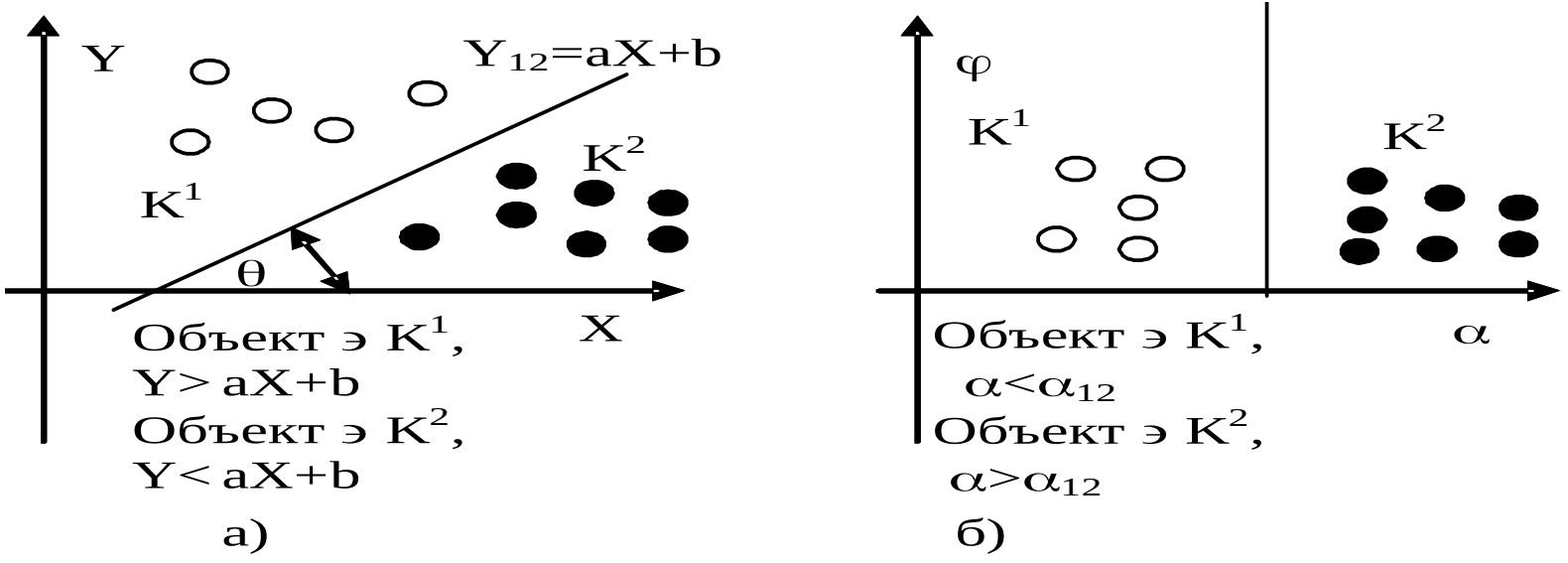
Генерация признаков – выявление признаков, которые наиболее полно описывают объект.

Селекция признаков – выявление признаков, которые имеют наилучшие классификационные свойства для конкретной задачи.

Построение модели.

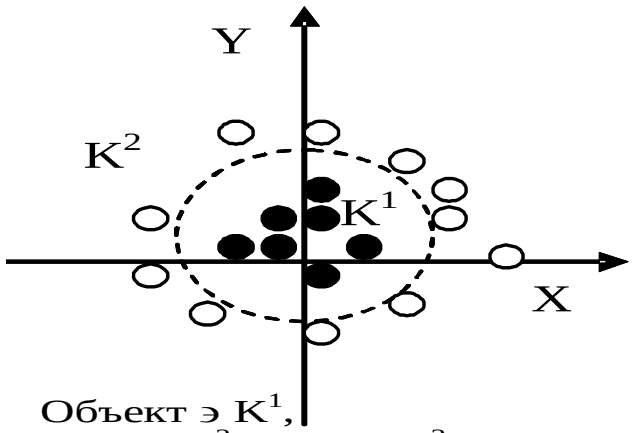
Оценка модели.

Изменение координат



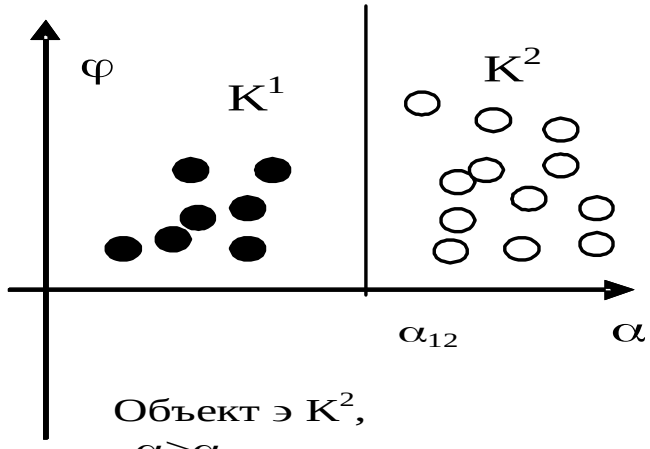
пространство объектов в исходной а) и развернутой на π/2-θ б) системе координат.

Изменение координат



Объект э K^1,
$b2(Y-b1)^2+a2(X-a1)^2<C$
Объект э K^2,
$b2(Y-b1)^2+a2(X-a1)^2>C$

а)



Объект э K^2,
$\alpha>\alpha_{12}$
Объект э K^1,
$\alpha<\alpha_{12}$

б)

пространство объектов в исходной а) и сферической б) системе координат.

ADD

Жадный алгоритм:

1) выбрать самый информативный признак :
текущая модель $\underline{X}$ ={x arg max I($\underline{X}$U$\underline{xi}$)}
2) пробуем добавить признак xi из множества оставшихся X* к $\underline{X}$ и оценим общую информативность I($\underline{X}$U$\underline{xi}$)
3) xj = arg max I($\underline{XUxi}$) , xi из X*
4) $\underline{X}$ =$\underline{X}$uxj, X* = X*\xj
5) |$\underline{X}$|<m, идем к 2)

Модели умеют считать важность признака

Линейная модель

$$a(x) = \sum_{i=1}^{n} w_i x^i$$

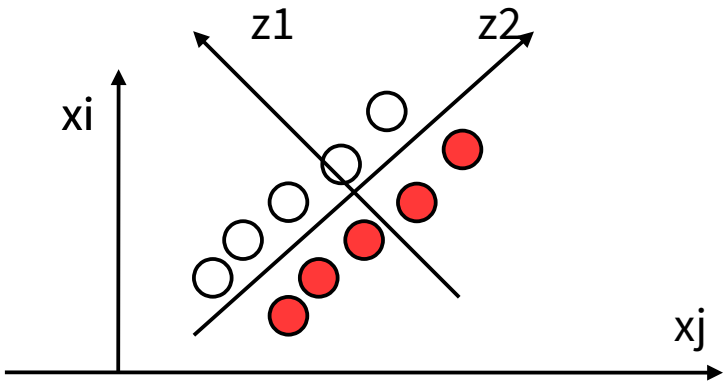
Деревья

$$Q(X_m, j, t) = H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

Понижение размерности: редукция

Новые координат

$$z_{ij} = \sum_{k=1}^{n} w_{jk} x_{ik}$$



Случайные проекции

$$d > \frac{8 \ln l}{\varepsilon^2},$$

l - количество объектов, ε - максимальное изменение расстояния между объектами

Метод главных компонент

По n – числу исходных признаков выделить k главных компонент, или обобщенных признаков.

Пространство главных компонент ортогонально.

Модель метода главных компонент основана на допущении, что значения множества взаимосвязанных признаков порождают некоторый общий результат

Метод главных компонент

principal component analysis, PCA $Z = XW^T$

X **- матрица "объекты-признаки", где по строкам отложены объекты, а по столбцам - значения признаков,**

Z **- матрица новых признаков,**

W^T **- транспонированная матрица весов (W — ортогональна)**

$$\|ZW - X\|^2 \to \min_{Z,W}$$

$$rank(X) \geq d$$

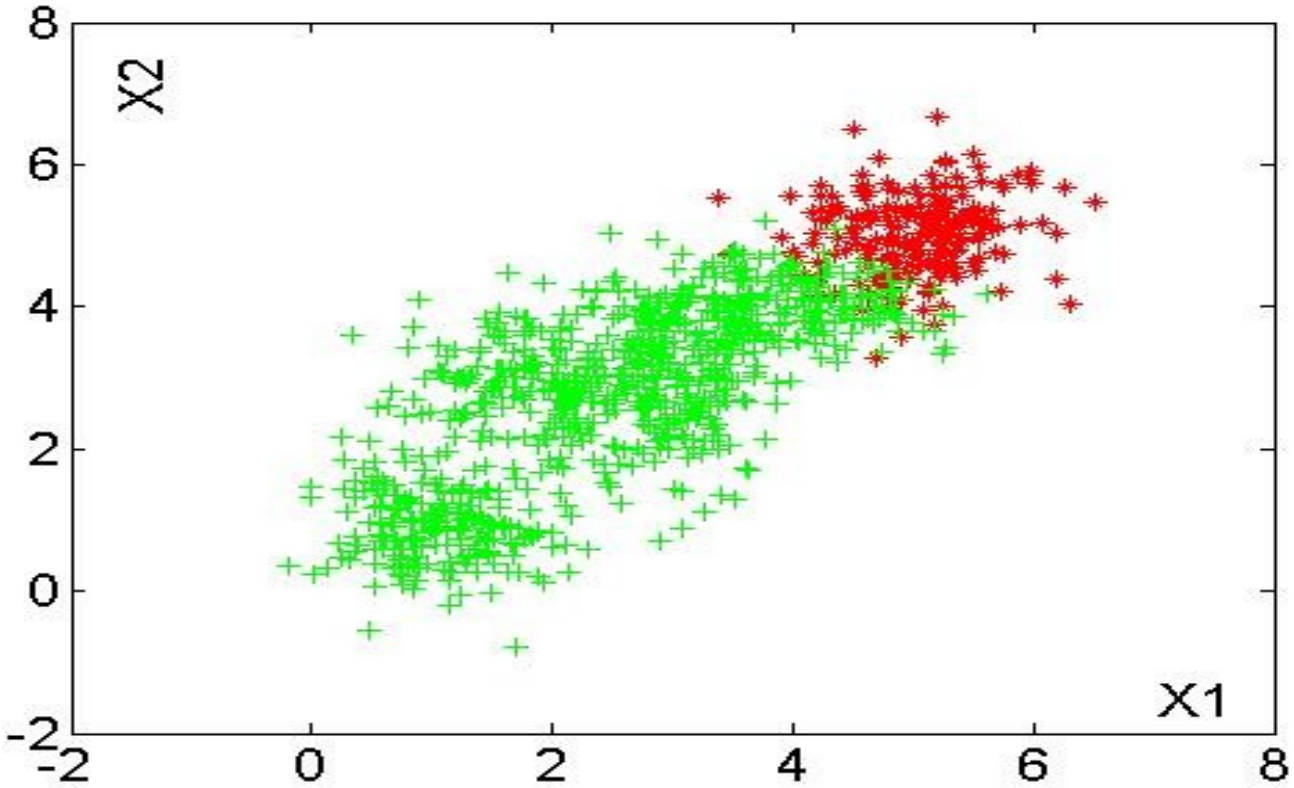
d - число новых признаков,

Метод главных компонент

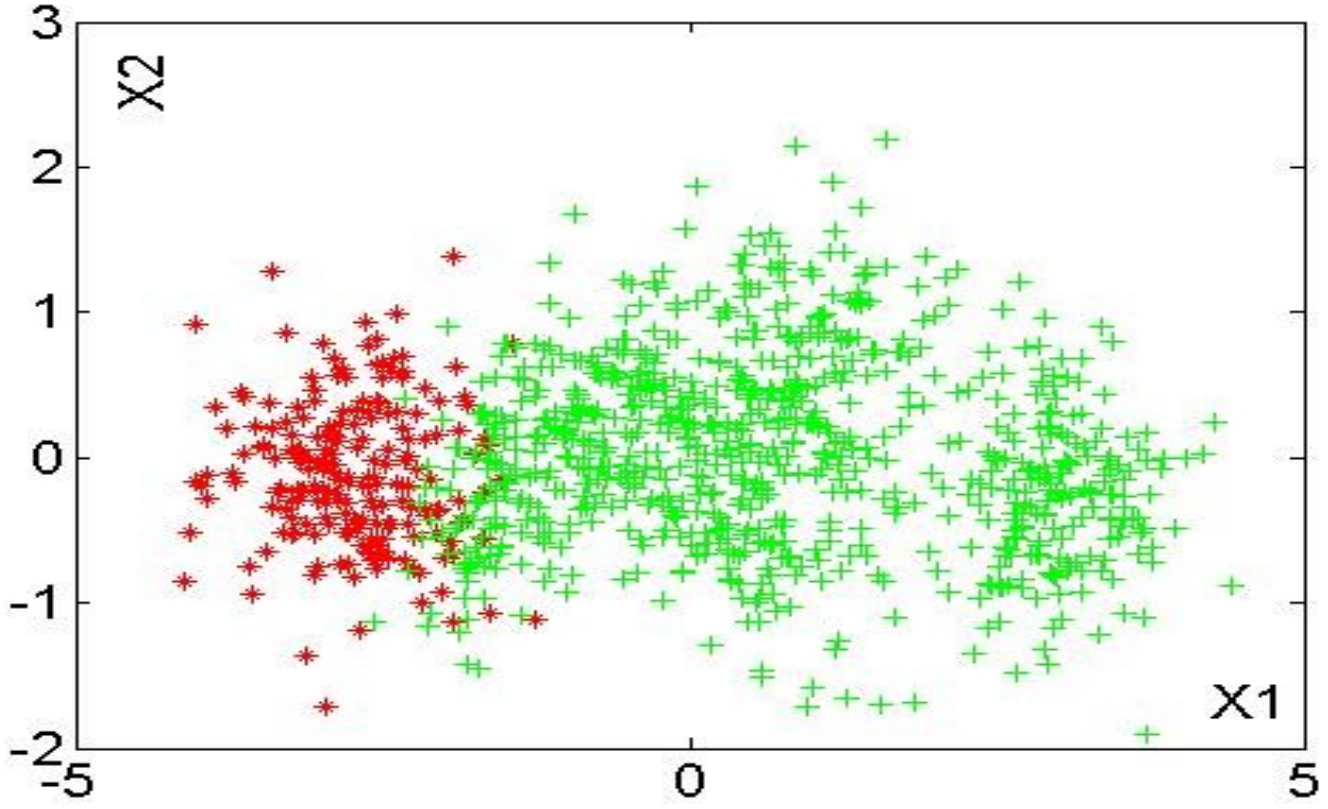
principal component analysis, PCA

- найти собственные значения матрицы $X^T X$;
- отобрать d максимальных;
- составить матрицу W^T,
- получить $Z = XW$.

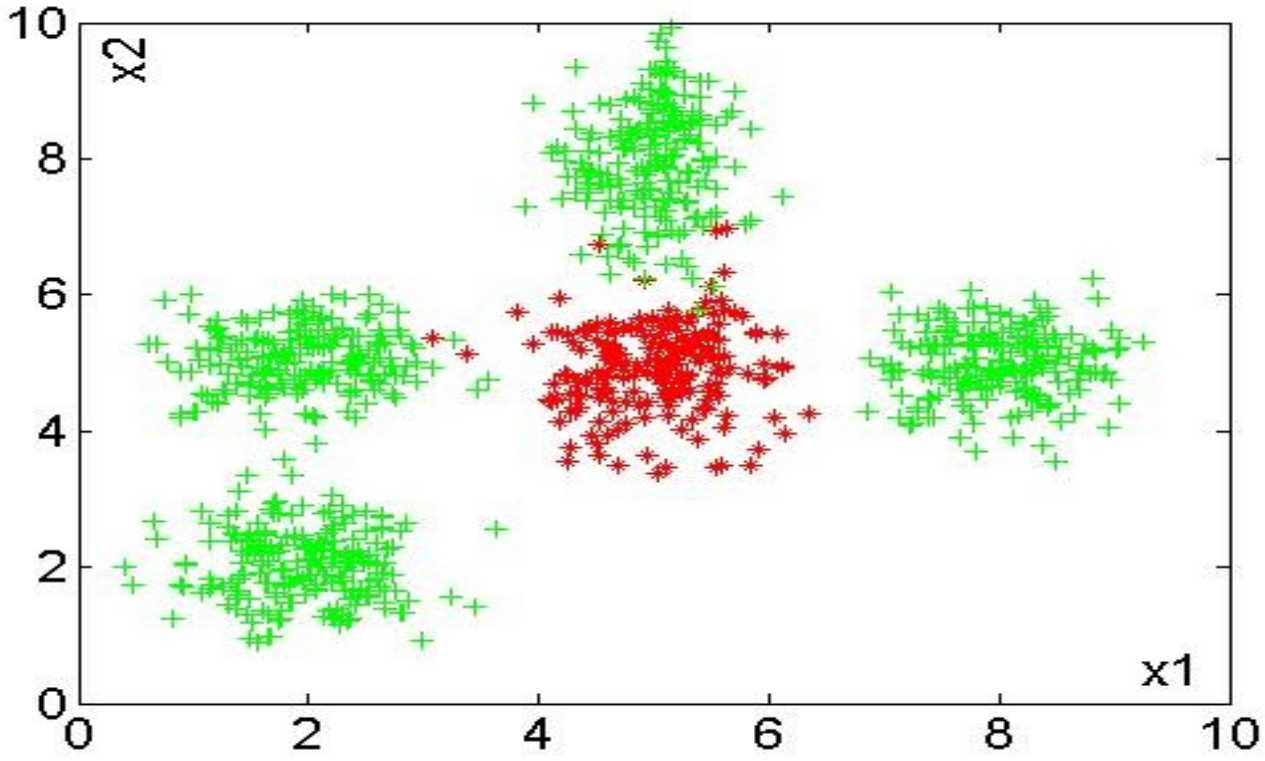
До преобразования



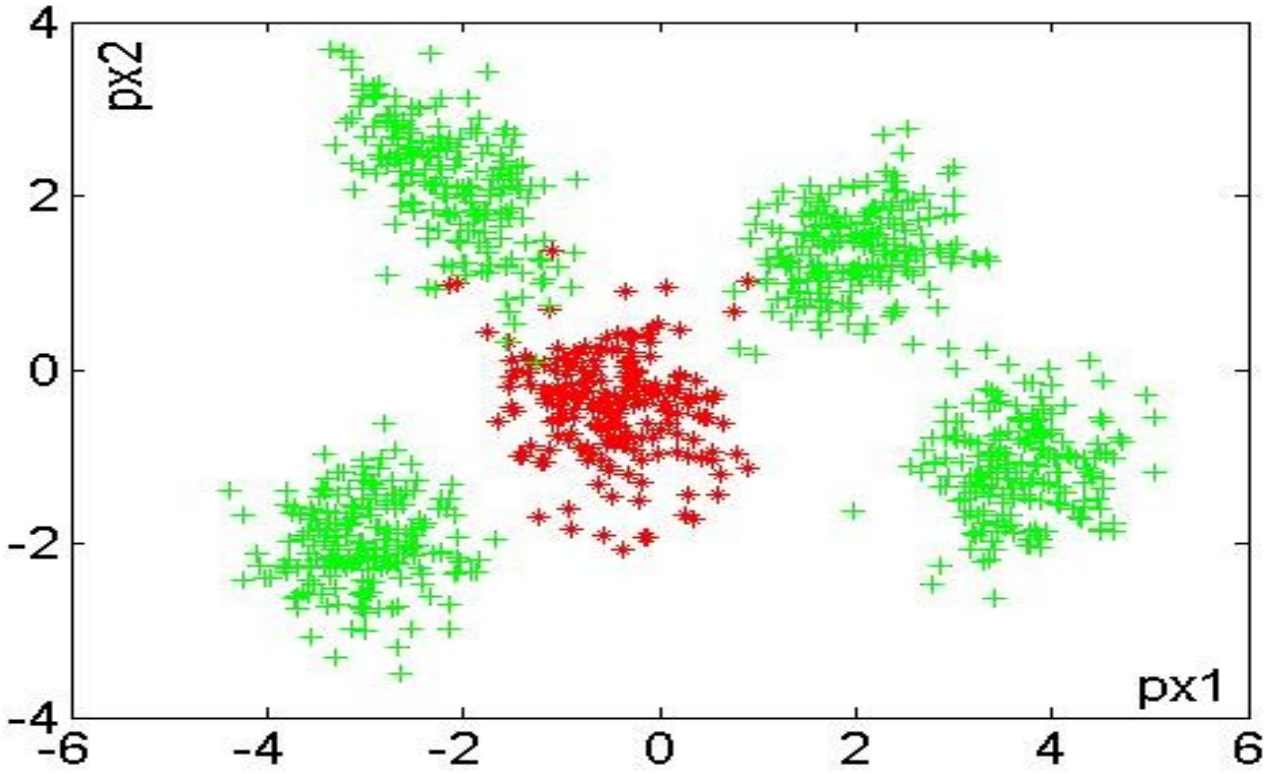
После преобразования



До преобразования



После преобразования



SVD

singular value decomposition, SVD

$$X = UDV^T.$$

U **- это собственные векторы матрицы** XX^T

столбцы ортогональной матрицы V **- собственные векторы матрицы** X^TX**,**

на главной диагонали диагональной D **собственные значения матриц** XX^T **и** X^TX

$$X = ZW = UDV^T$$

При $d = n$

SVD

singular value decomposition, SVD

$$X = U D V^T.$$

U **- это собственные векторы матрицы** XX^T

столбцы ортогональной матрицы V **- собственные векторы матрицы** $X^T X$**,**

на главной диагонали диагональной D **собственные значения матриц** XX^T **и** $X^T X$

$$X = ZW = UDV^T \qquad W = V^T \qquad Z = UD.$$

При $d = n$

SVD

- найти сингулярное разложение вектора X.
- сформировать из столбцов матрицы V, матрицу весов W;
- получить $Z = XW$.