# Методы понижения размерности

# Когда появляется много признаков?

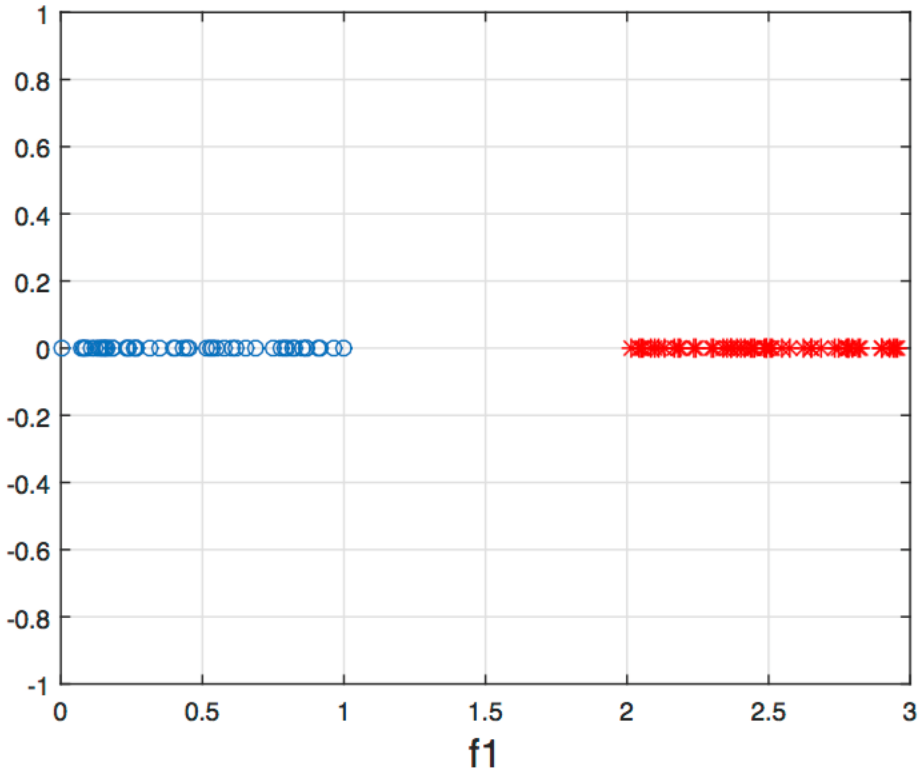
- Иногда сразу в исходных данных (ДНК)
- При обработке категориальных признаков
- При выделении признаков (например, при анализе текстов)

# Зачем понижать размерность?

- Чтобы уменьшить вычислительную сложность обучения и предсказания
- Чтобы избавиться от проклятия размерности (например, в kNN)
- Сделать модель интерпретируемой
- Улучшить качество модели

# Плохие признаки

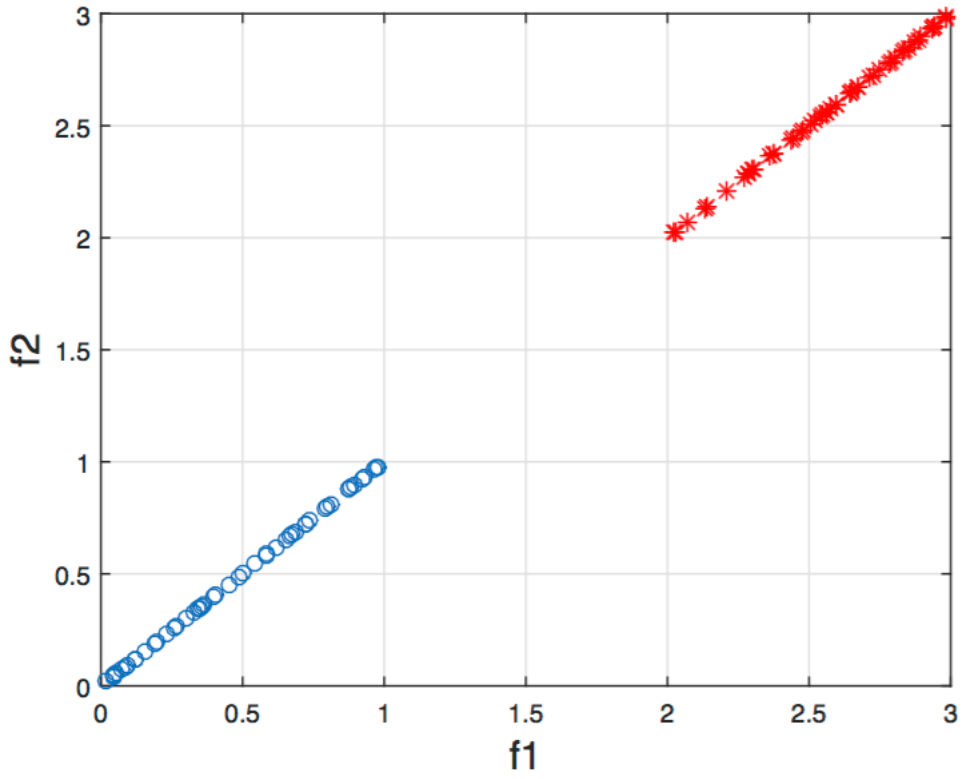
Информативный признак



# Плохие признаки

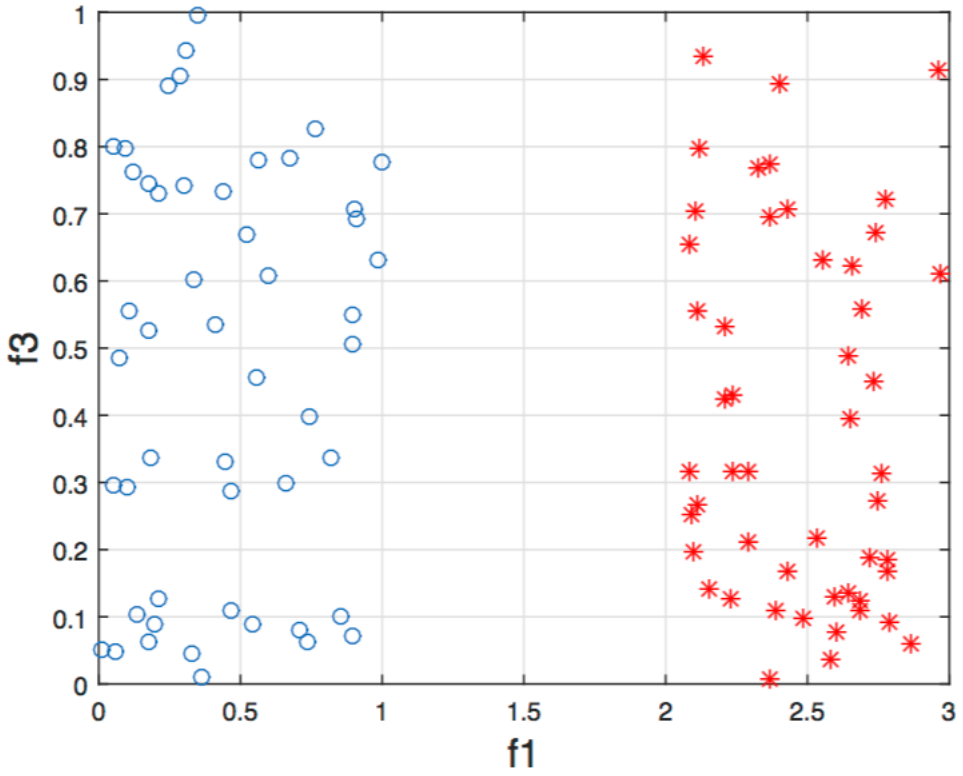
Коррелирующие признаки

f2 — избыточный признак



# Плохие признаки

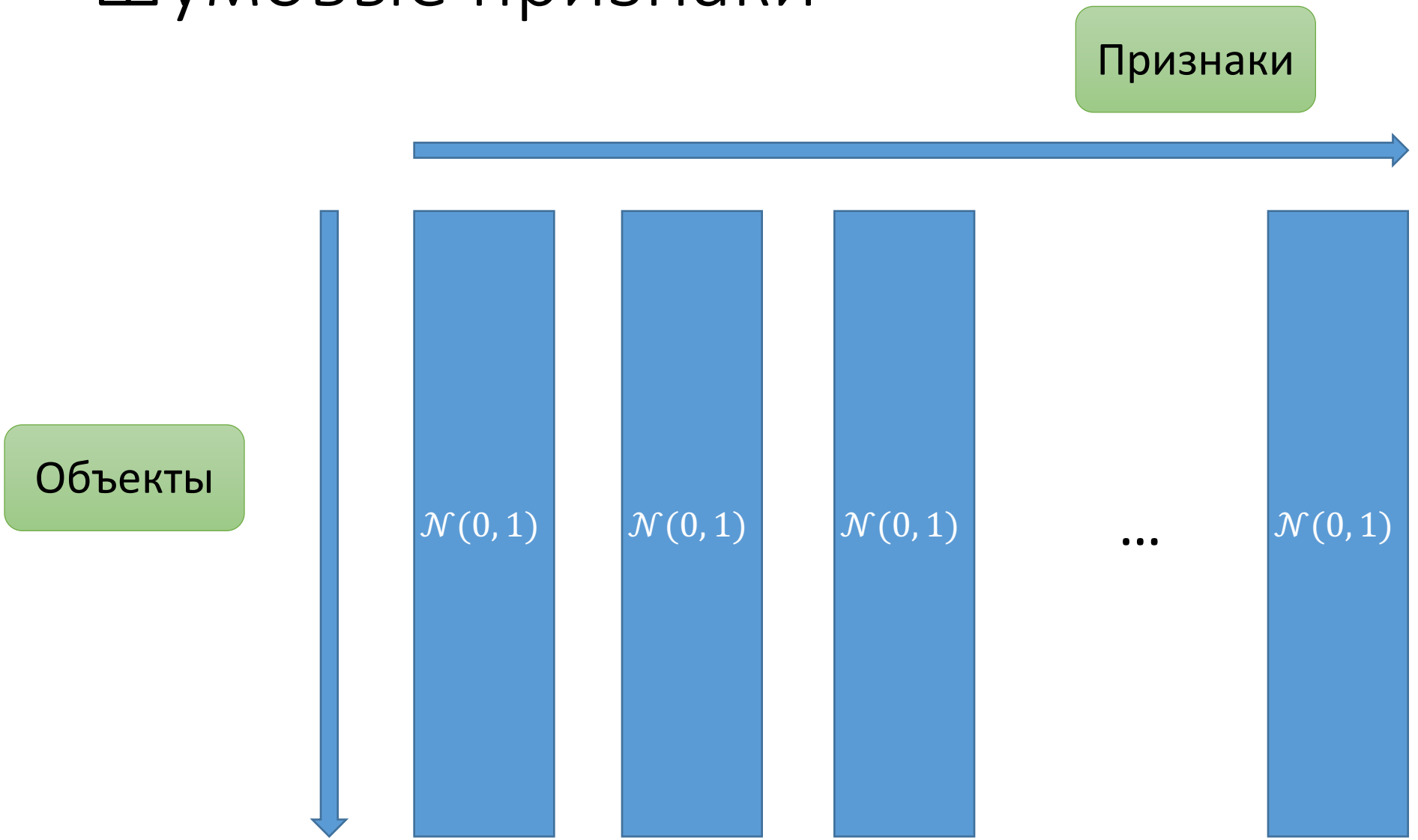
f3 — шумовой признак



# Шумовые признаки

- Признаки, которые никак не связаны с целевой переменной
- Но по обучающей выборке это не всегда можно понять

# Шумовые признаки



# Шумовые признаки

- Генерируем случайные признаки
- Если их много, то некоторые будут хорошо коррелировать с ответами

| $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| -1 | 1.11 | **-0.5** | 0.42 | 0.33 |
| -1 | 1.22 | **-0.46** | -1.98 | -0.55 |
| 1 | -1.56 | **0.04** | 0.39 | -1.67 |
| 1 | -0.48 | **1.32** | 0.88 | -0.27 |

# Методы понижения размерности

- Отбор признаков (feature selection)
  - Выбрать $d$ самых важных признаков
- Извлечение признаков (feature extraction)
  - Найти $d$ новых признаков, выражающихся через исходные

# Методы понижения размерности

- Фильтрация (filter methods)
  - Понижение размерности без учёта модели
- Методы-обёртки (wrapper methods)
  - Выбор признаков, дающих лучшее качество для модели
- Понижение с помощью моделей (embedded methods)
  - Использование свойств моделей для оценивания важности признаков

# Одномерные методы

# Одномерные методы

- Оценивают важность каждого признака по отдельности
- Относятся к **методам фильтрации**
- Относятся к **методам отбора признаков**

# Дисперсия признаков

$$R_j = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(x_{ij} - \overline{x_j}\right)^2$$

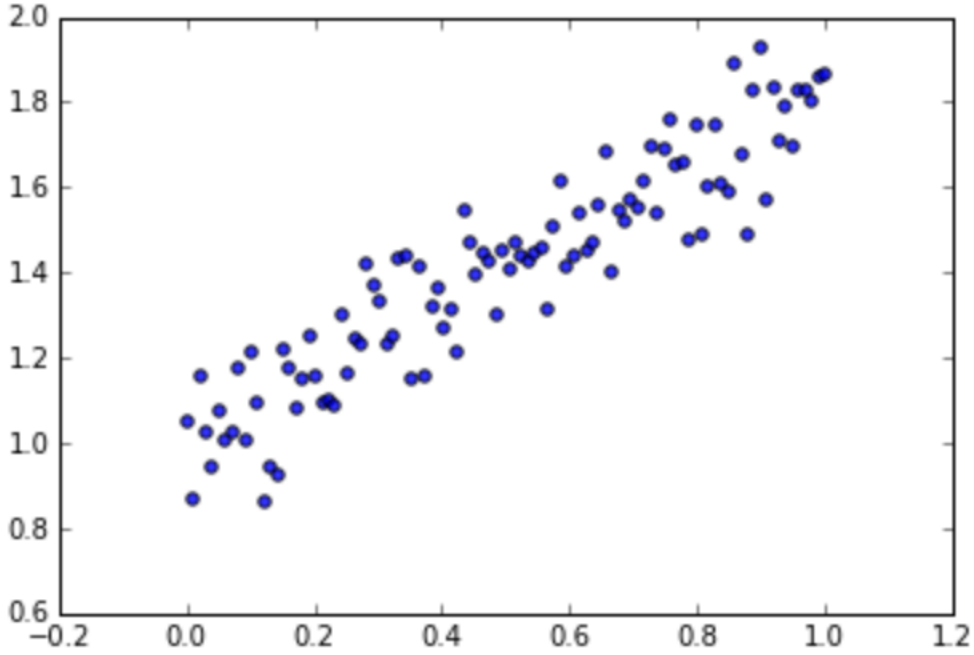
- Чем больше $R_j$, тем информативнее признак
- Никак не учитываются ответы
- Подходит для фильтрации константных и близких к ним признаков

# Корреляция

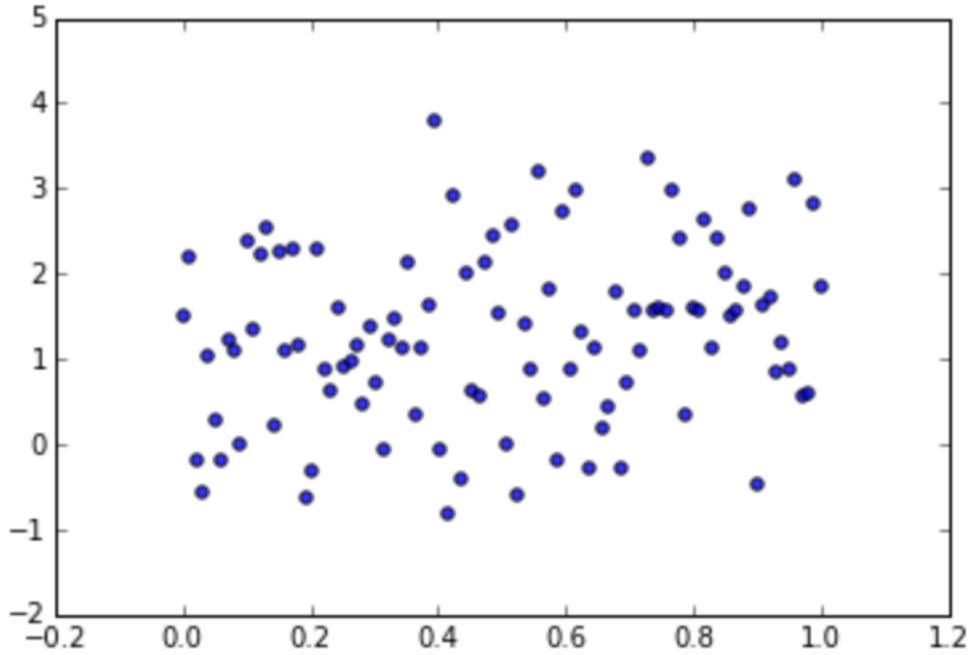
$$R_j = \frac{\sum_{i=1}^{\ell}\left(x_{ij} - \overline{x_j}\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell}\left(x_{ij} - \overline{x_j}\right)^2 \sum_{i=1}^{\ell}(y_i - \bar{y})^2}}$$

- Чем больше $\left|R_j\right|$, тем информативнее признак
- Учитывает только линейную связь

# Корреляция для регрессии

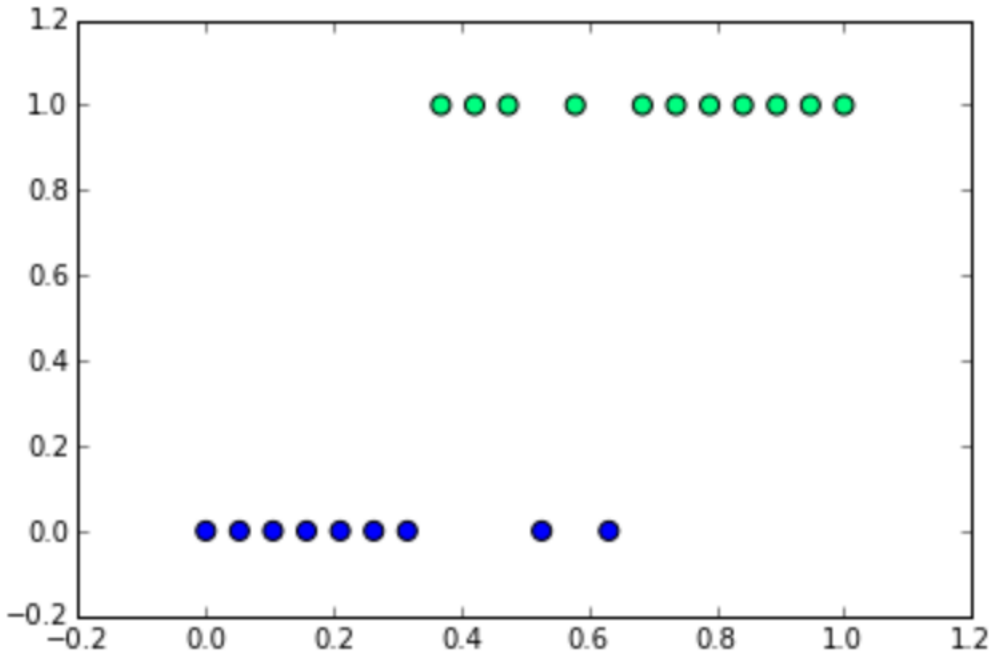



Корреляция
0.927

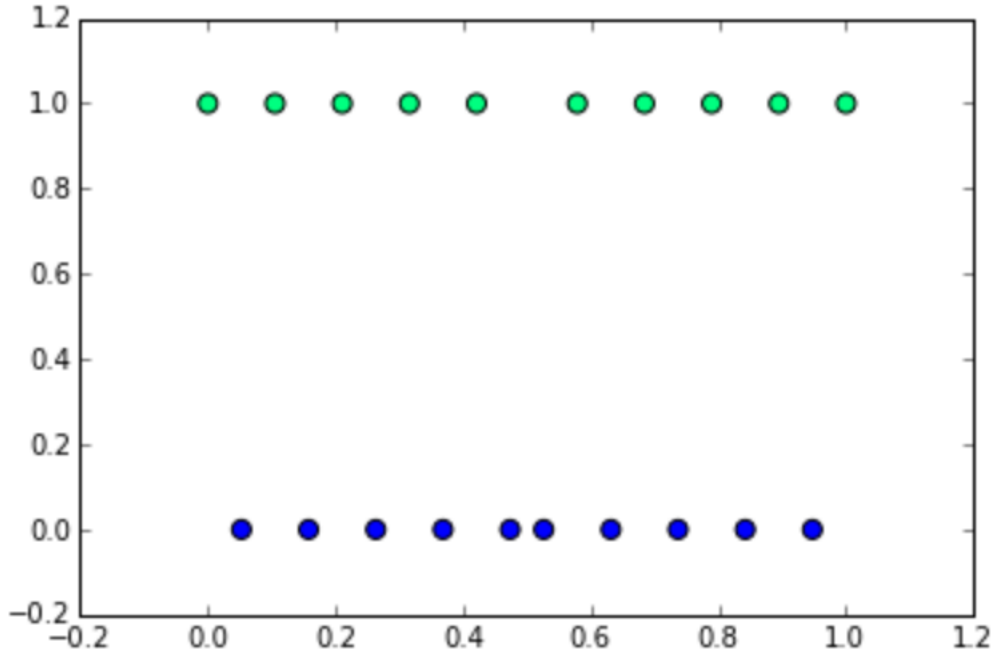



Корреляция
0.207

# Корреляция для классификации




Корреляция
0.741



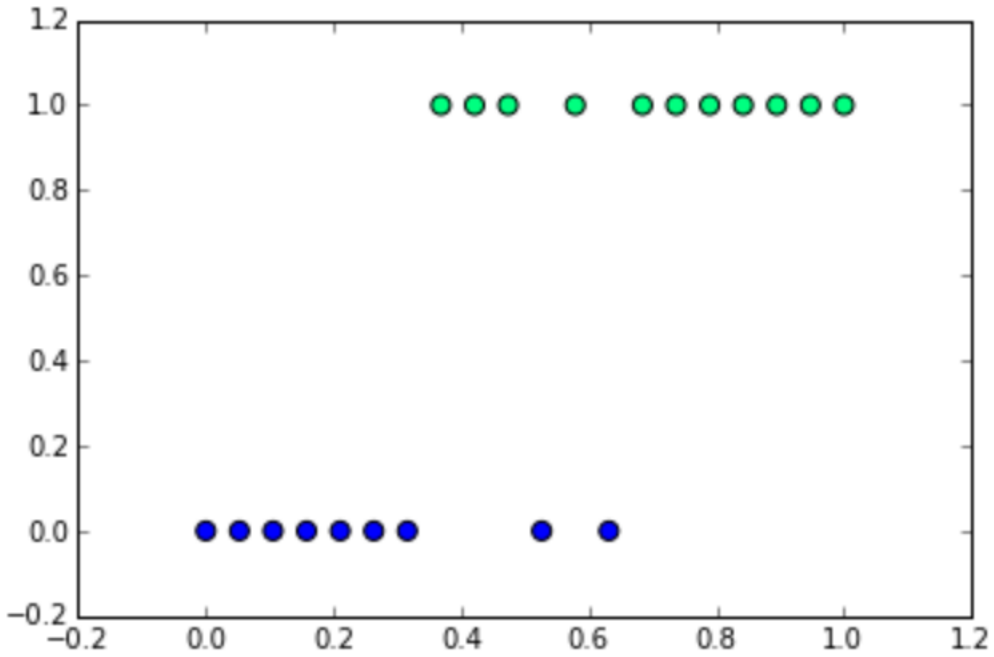

Корреляция
0.0

# T-score

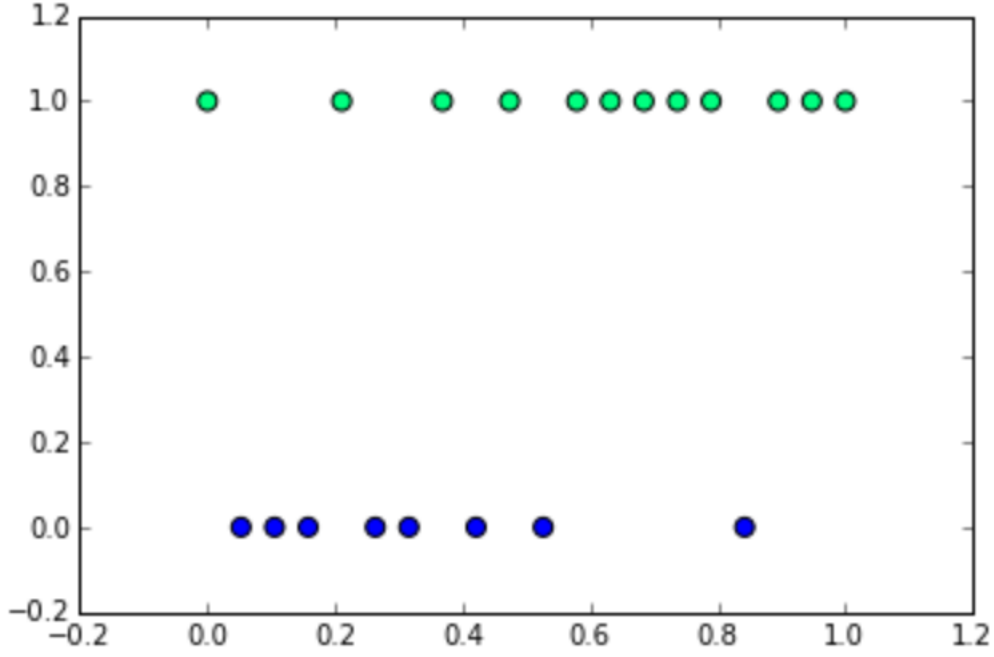
$$R_j = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Для задач бинарной классификации
- Чем больше $R_j$, тем информативнее признак
- $\mu_1, \mu_2$ — средние значения признаков в первом и втором классах
- $\sigma_1^2, \sigma_2^2$ — дисперсии
- $n_1, n_2$ — число объектов в первом и втором классах

# T-score




T-score
4.95




T-score
2.28

# F-score

$$R_j = \frac{\sum_{k=1}^{K} \frac{n_j}{K-1} \left(\mu_j - \mu\right)^2}{\frac{1}{\ell - K} \sum_{k=1}^{K} \left(n_j - 1\right) \sigma_j^2}$$

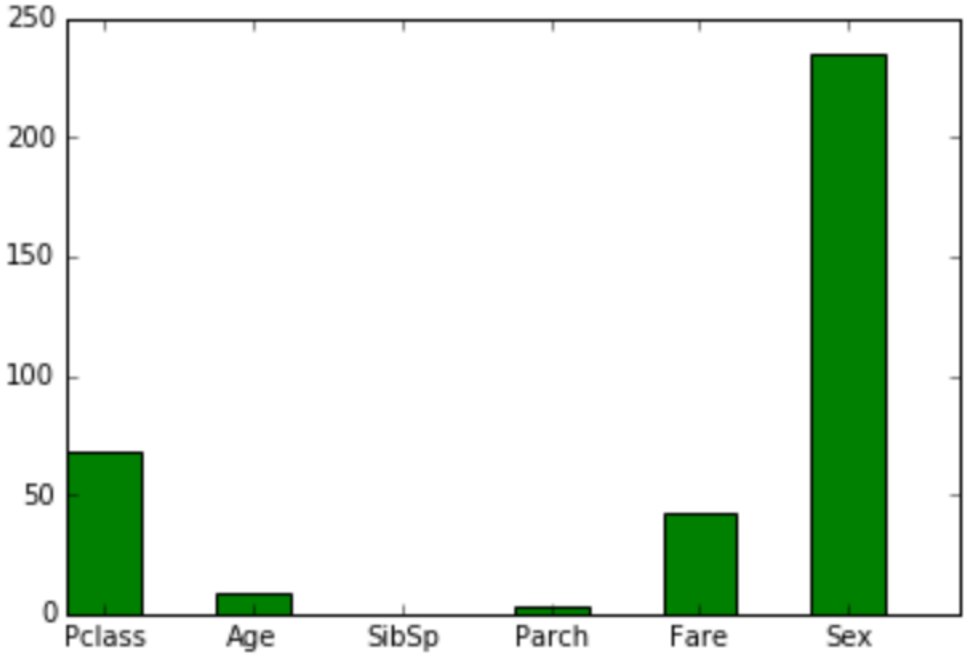
- Для задач многоклассовой классификации
- Чем больше $R_j$, тем информативнее признак
- $\mu_1, \ldots, \mu_K$ — средние значения признаков в классах
- $\mu$ — среднее значение признака по всей выборке
- $\sigma_1^2, \ldots, \sigma_K^2$ — дисперсии
- $n_1, \ldots, n_K$ — число объектов в первом и втором классах

# Пример: Titanic

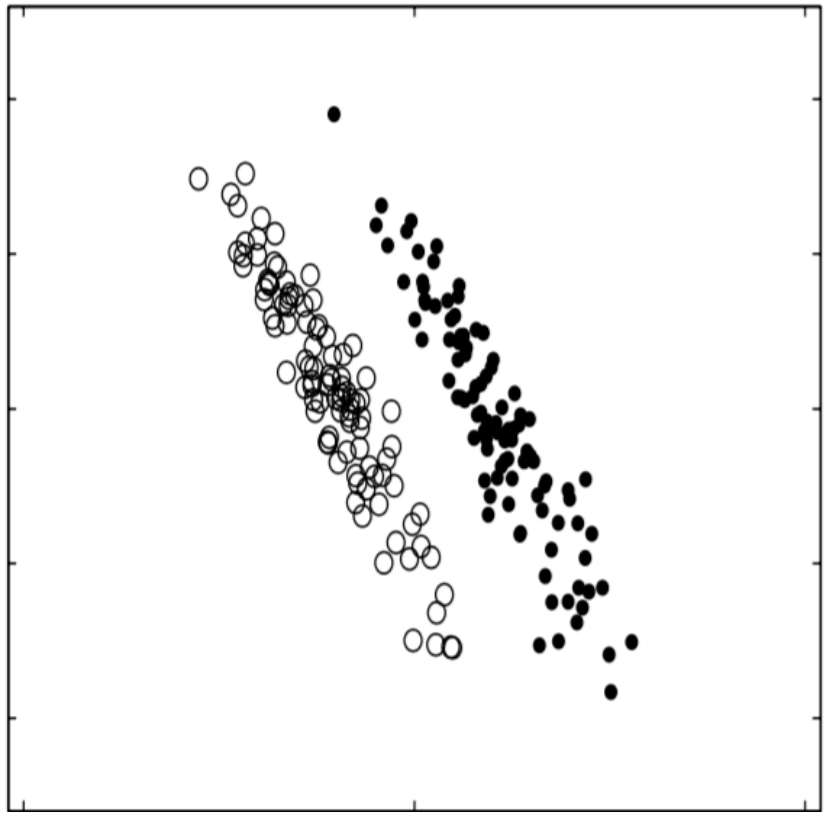
| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35 | 0 | 0 | 373450 | 8.0500 | NaN | S |

# Пример: Titanic

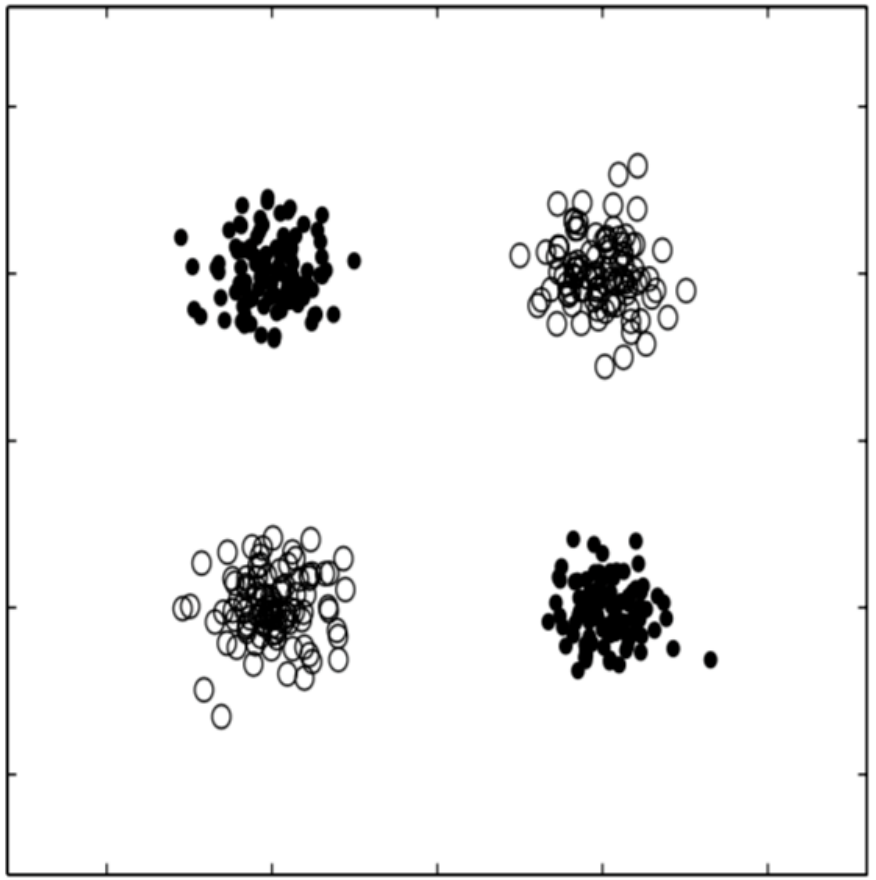
- Вычислим T-score для всех признаков
- Действительно, пол сильнее всего коррелирует с выживаемостью пассажиров



# Проблемы

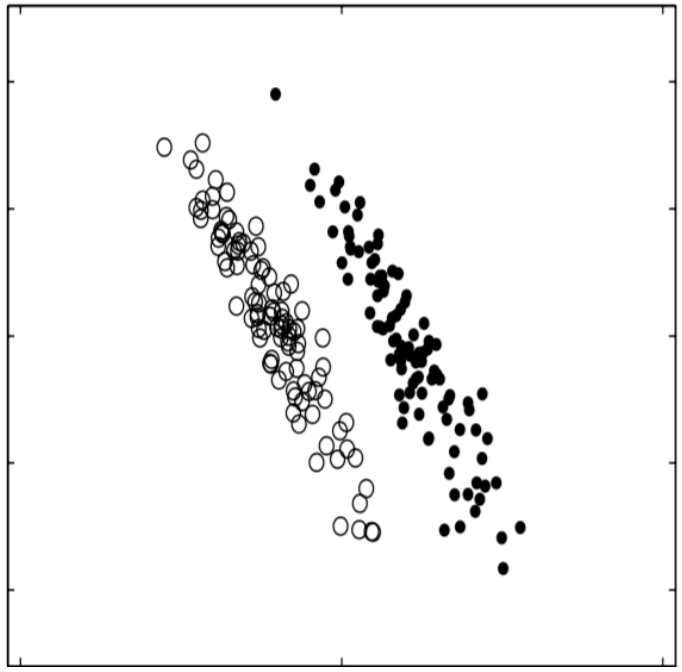


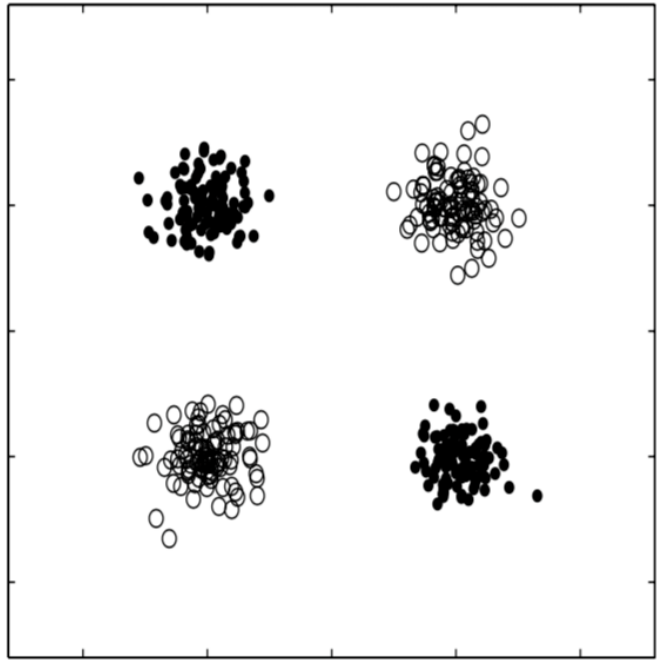
# Проблемы



# Проблемы

- Одномерные критерии не работают, если целевая переменная зависит от совокупностей признаков





# Отбор с помощью моделей

# Отбор с помощью моделей

- Оценивают важность признаков, используя модели машинного обучения
- Относятся к **методам отбора признаков**

# Линейные модели

$$a(x) = \sum_{j=1}^{d} w_j x^j$$

- Если признаки масштабированы, то веса можно использовать как показатели информативности
- Для повышения числа нулевых весов — $L_1$-регуляризация

# $L_1$-регуляризация

$$Q(a, X) + \lambda \sum_{j=1}^{d} |w_j| \to \min_w$$

- Чем выше $\lambda$, тем больше весов зануляется
- Позволяет построить модель, использующую только самые важные признаки

# Решающие деревья

- Поиск лучшего разбиения:

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

- $H(X)$ — критерий информативности (MSE, энтропийный)

# Решающие деревья

- Чем сильнее уменьшили $H(X)$, тем лучше признак
- Уменьшение критерия:

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

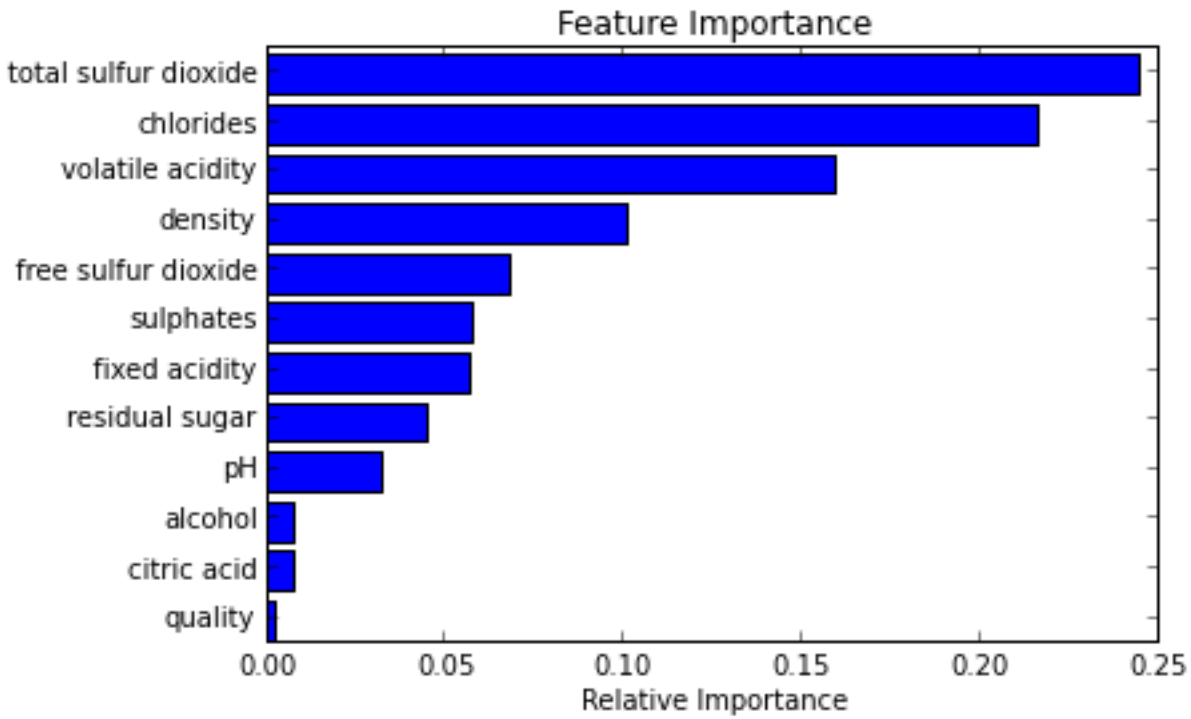
- Важность признака $R_j$: просуммируем уменьшения по всем вершинам, где разбиение делалось по признаку $j$

# Случайный лес

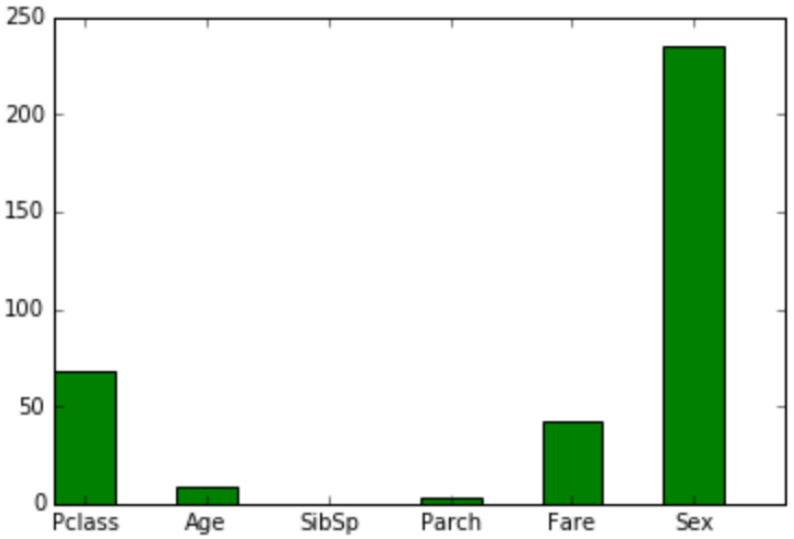
- Сумма важностей $R_j$ по всем деревьям
- Чем больше, тем важнее признак
- Учитывается важность признаков в совокупности

# Случайные леса и отбор признаков

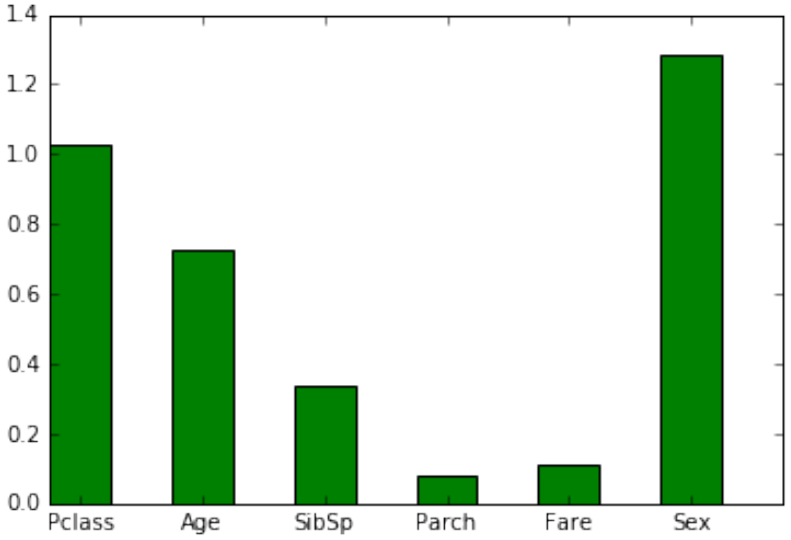
- Классификация вин на белые и красные



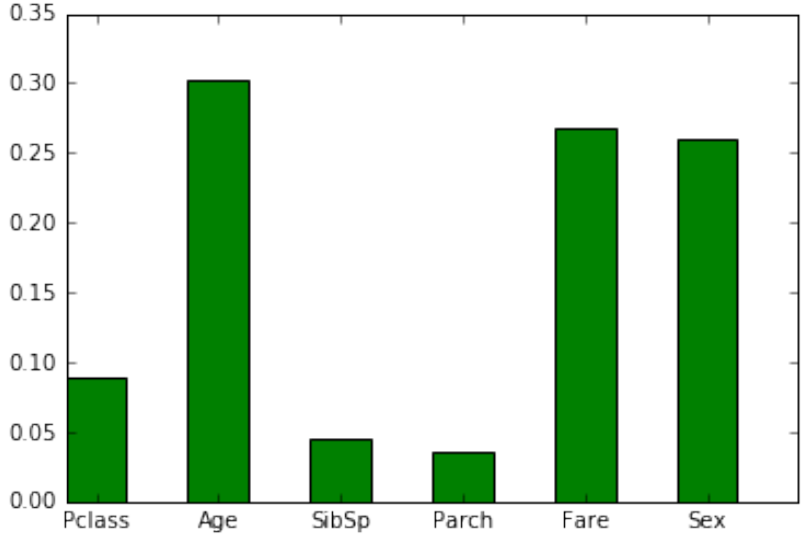
# Пример: Titanic




Одномерный отбор



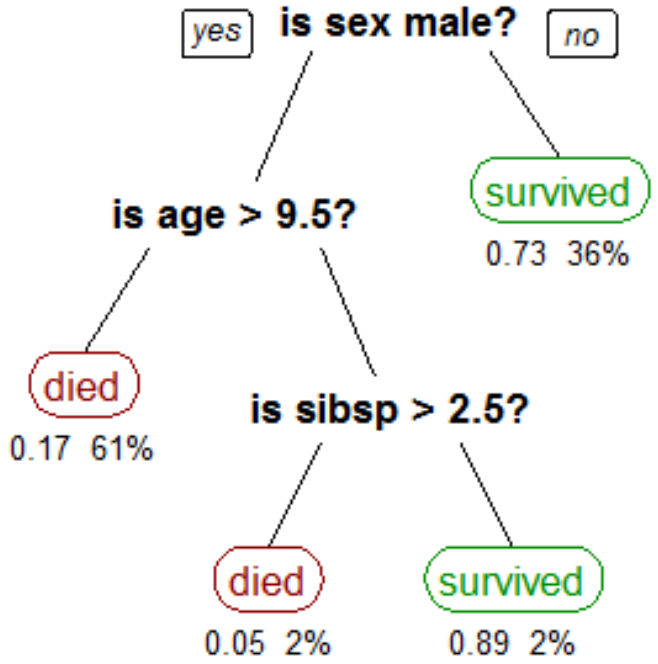

Логистическая регрессия




Случайный лес

# Пример: Titanic

- Модели выделяют признак Age как важный
- Ответы зависят от возраста только в совокупности с полом

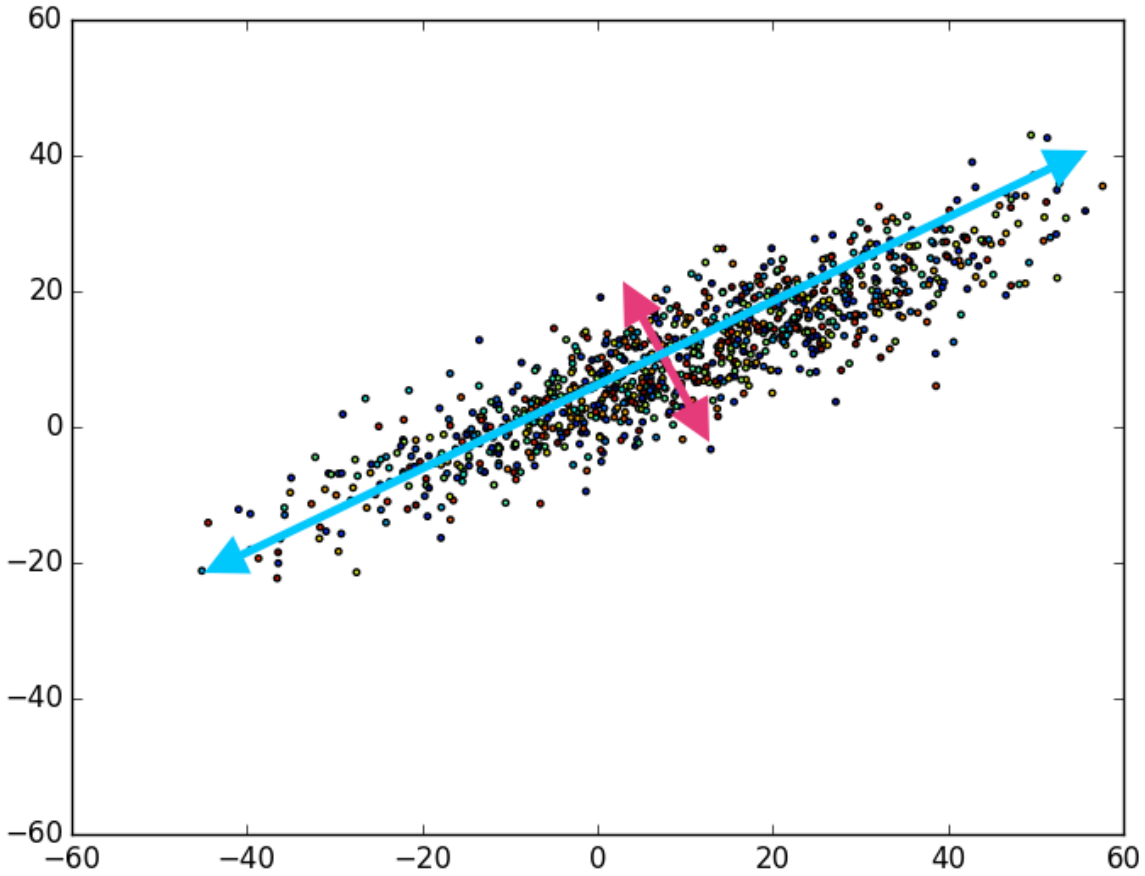


# Метод главных компонент

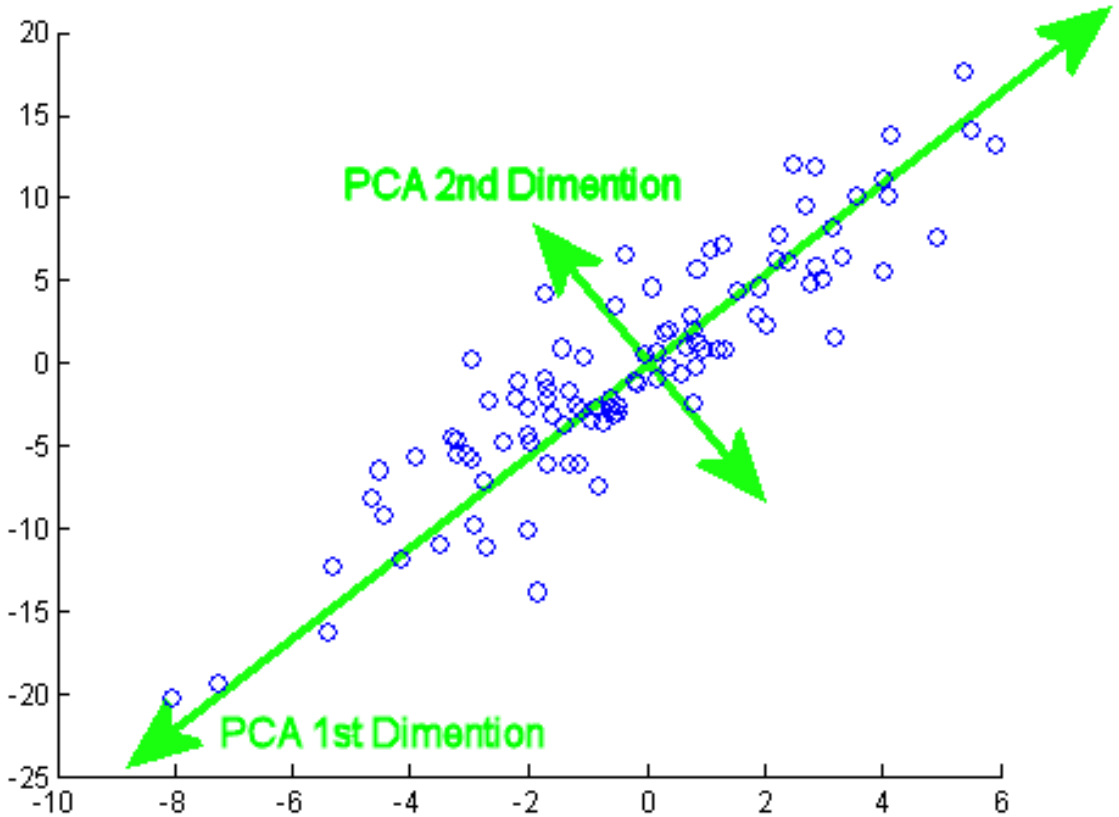
# Метод главных компонент

- Principal component analysis (PCA)
- Проецирует данные в пространство меньшей размерности
- Относится к **методам фильтрации**
- Относится к **методам извлечения признаков**

# Извлечение признаков



# Извлечение признаков



# Извлечение признаков

- Порождение новых признаков
- Их должно быть меньше
- Они должны содержать как можно больше информации из исходных признаков

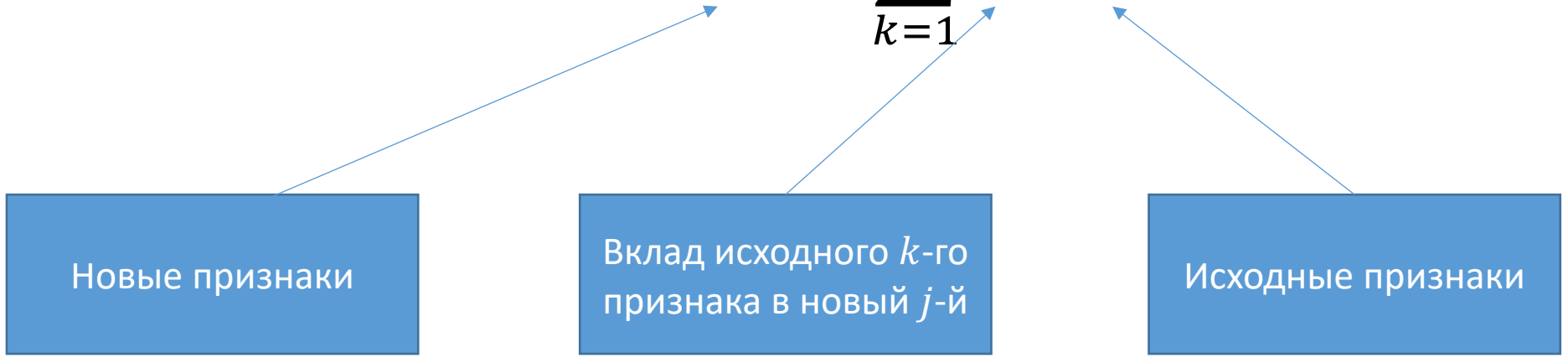
# Извлечение признаков

- Линейные методы
- Каждый новый признак — линейная комбинация исходных

# Извлечение признаков

- Исходные признаки: $x_{ik}$, $D$ штук
- Новые признаки: $z_{ij}$, $d$ штук
- Линейный подход:

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



# Метод главных компонент

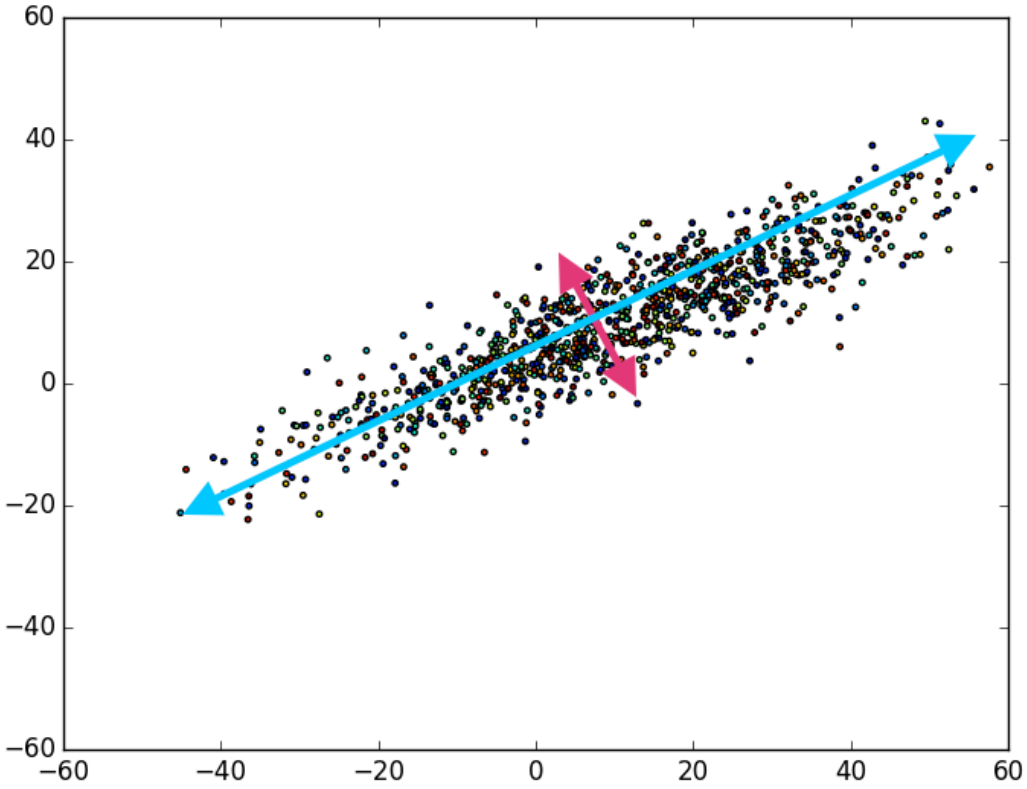
- Матричная запись:

$$Z = XW^T$$

- $j$-й столбец $W$ — коэффициенты при исходных признаках для вычисления нового $j$-го признака

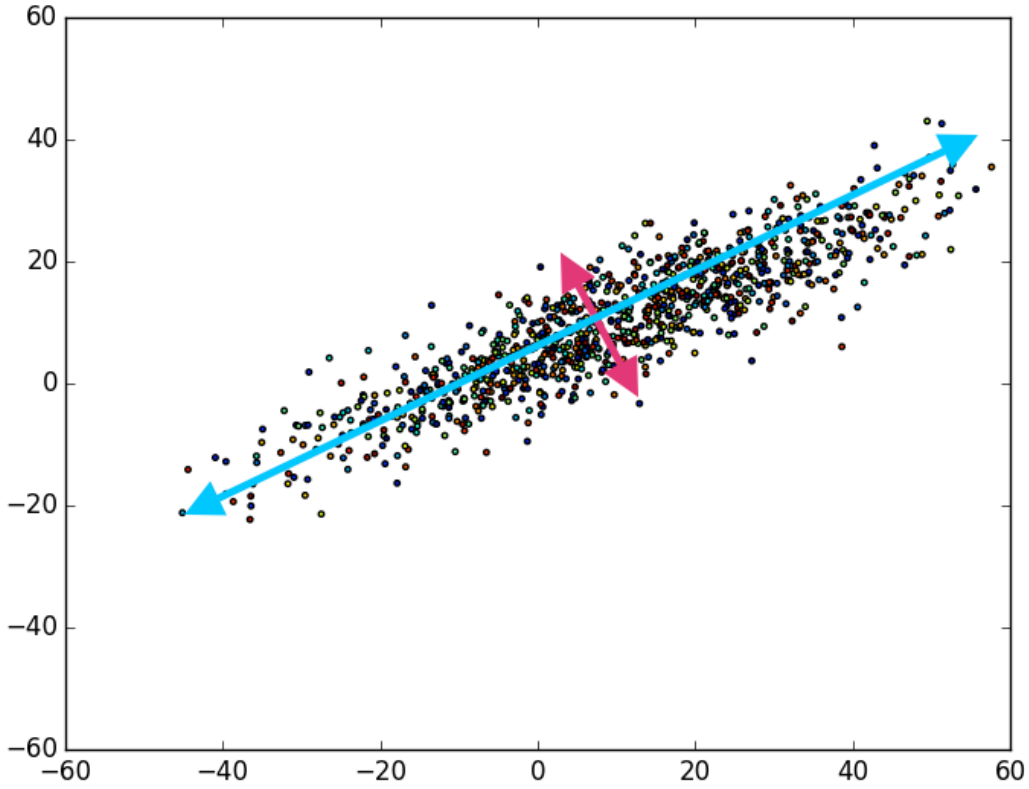
# Метод главных компонент

- Геометрический смысл — поиск гиперплоскости для проецирования выборки
- Как выбирать гиперплоскость?

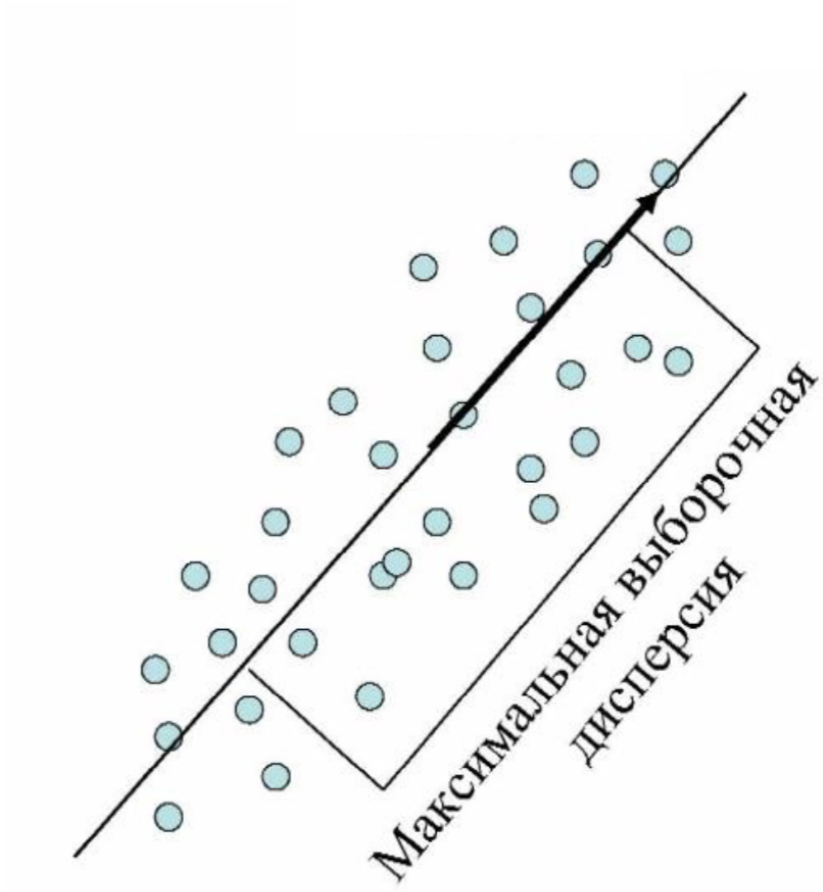


# Метод главных компонент

- Чем выше дисперсия выборки после проецирования, тем лучше
- Дисперсия — мера количества информации



# Метод главных компонент



# Максимизация дисперсии

$$\begin{cases} \sum_{j=1}^{d} w_j^T X^T X w_j \to \max\limits_{W} \\ W^T W = I \end{cases}$$

# Максимизация дисперсии

$$\begin{cases} \sum_{j=1}^{d} w_j^T X^T X w_j \to \max_W \\ W^T W = I \end{cases}$$

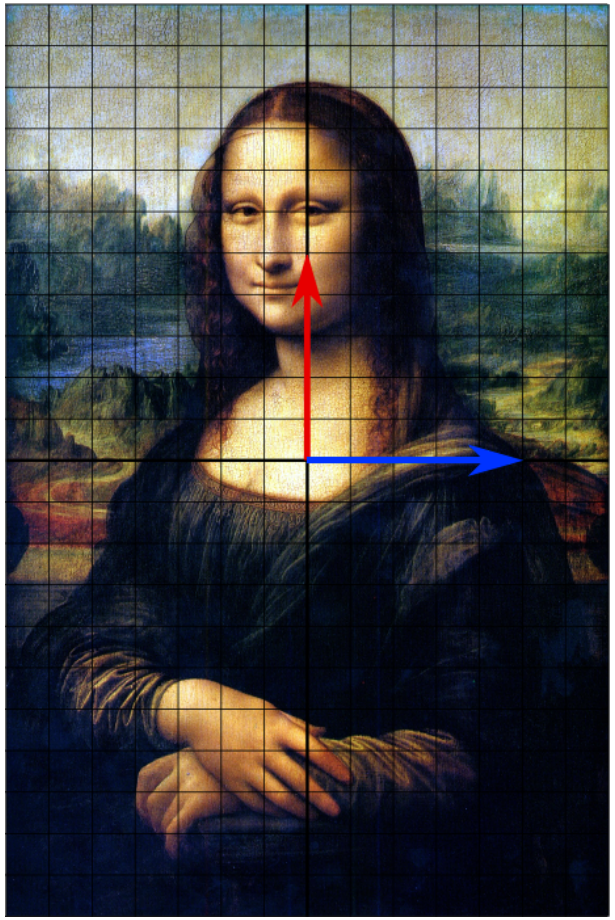
Дисперсия выборки

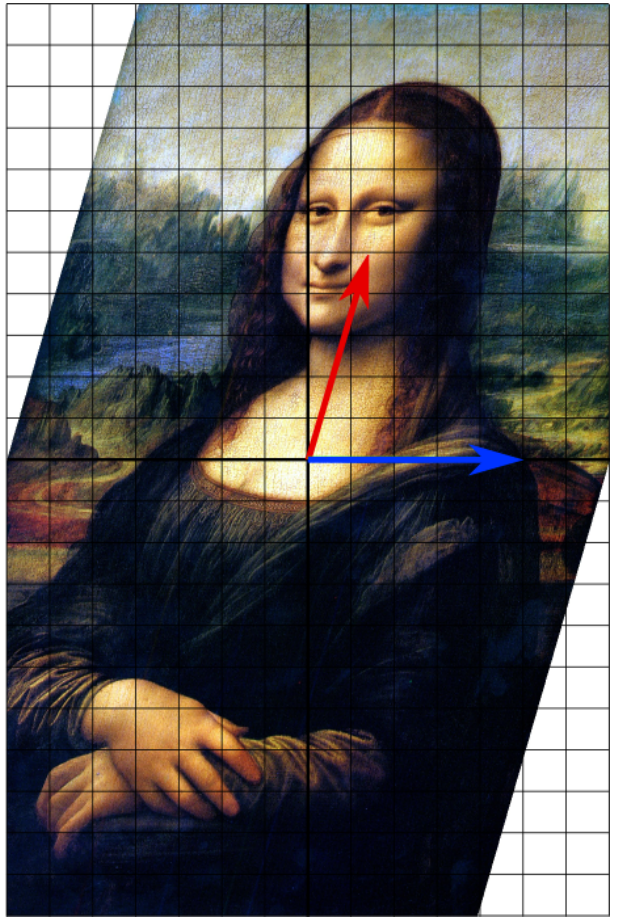
Векторы в W образуют ортонормированный базис

# Собственные векторы

- $A$ — матрица размера $n \times n$
- Пусть $Ax = \lambda x$
- Тогда $x$ — собственный вектор, $\lambda$ — собственное значение
- $x$ — вектор, который не меняет направление под воздействием матрицы

# Собственные векторы

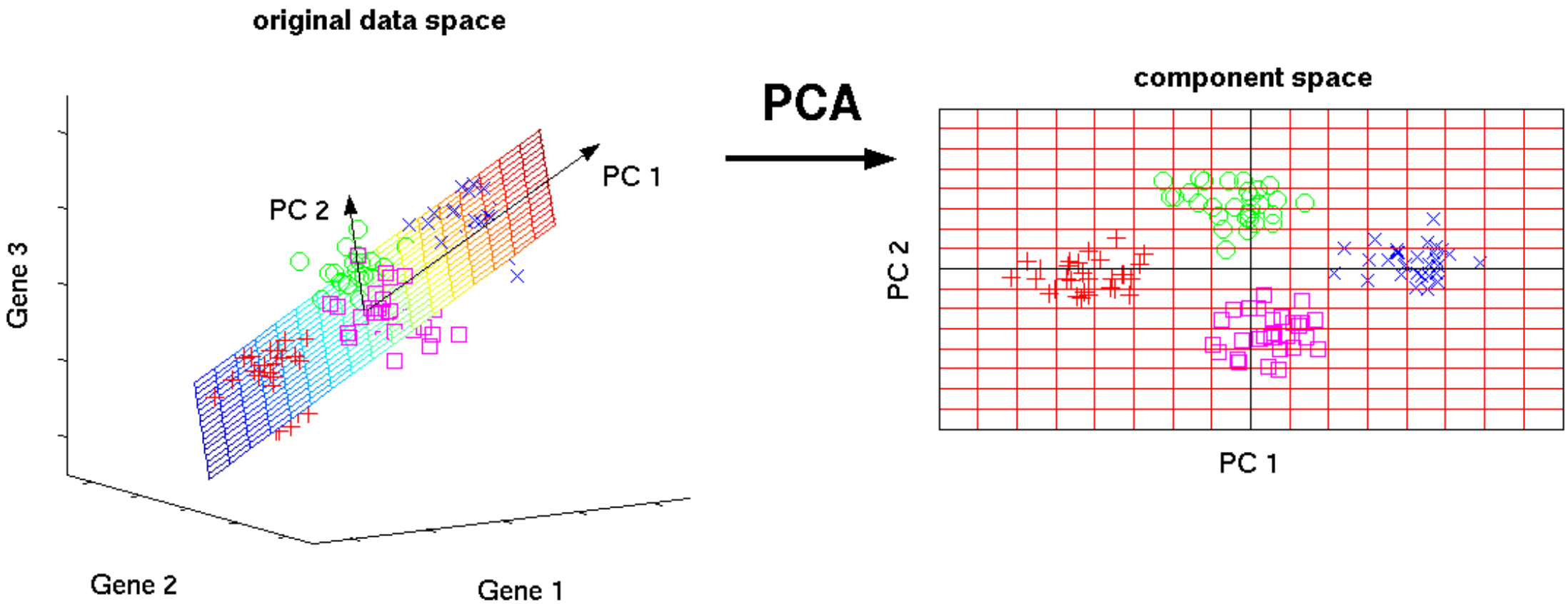




# Решение

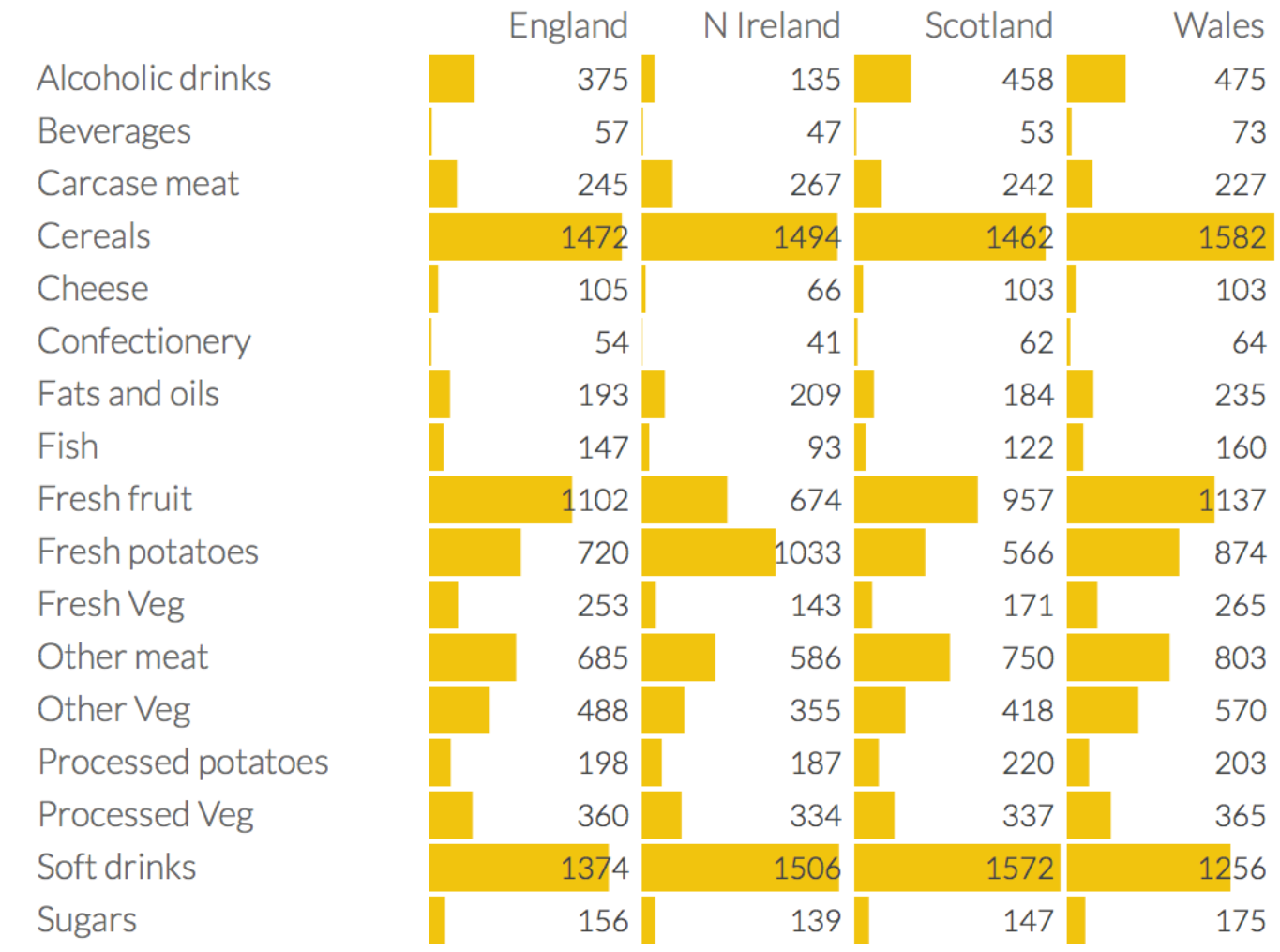
- Столбцы $W$ — собственные векторы матрицы $X^T X$, соответствующие наибольшим собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_d$
- $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$ — доля дисперсии, сохранённой при понижении размерности

# Метод главных компонент



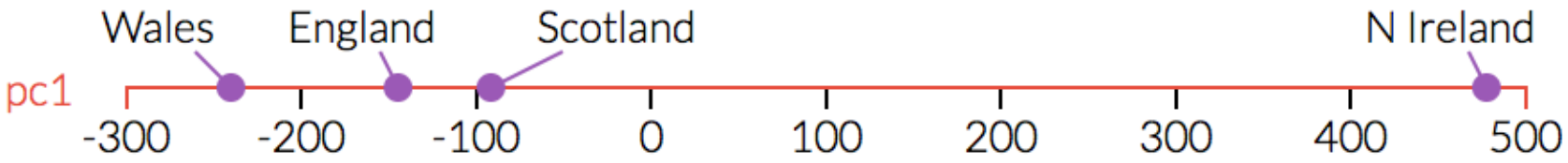
# Рацион в Великобритании

- Данные — среднее потребление продуктов в неделю в каждой провинции
- Не очень удобно смотреть на них

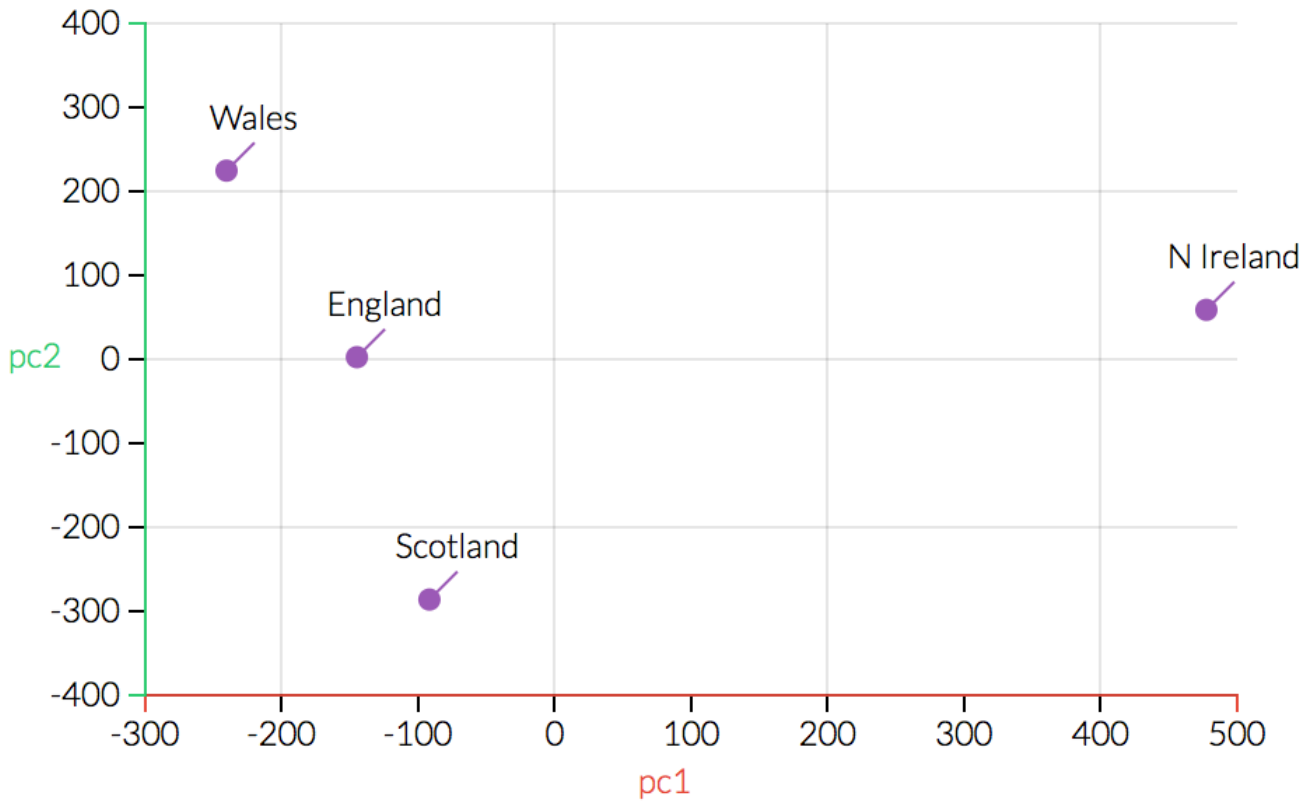


| | England | N Ireland | Scotland | Wales |
|---|---|---|---|---|
| Alcoholic drinks | 375 | 135 | 458 | 475 |
| Beverages | 57 | 47 | 53 | 73 |
| Carcase meat | 245 | 267 | 242 | 227 |
| Cereals | 1472 | 1494 | 1462 | 1582 |
| Cheese | 105 | 66 | 103 | 103 |
| Confectionery | 54 | 41 | 62 | 64 |
| Fats and oils | 193 | 209 | 184 | 235 |
| Fish | 147 | 93 | 122 | 160 |
| Fresh fruit | 1102 | 674 | 957 | 1137 |
| Fresh potatoes | 720 | 1033 | 566 | 874 |
| Fresh Veg | 253 | 143 | 171 | 265 |
| Other meat | 685 | 586 | 750 | 803 |
| Other Veg | 488 | 355 | 418 | 570 |
| Processed potatoes | 198 | 187 | 220 | 203 |
| Processed Veg | 360 | 334 | 337 | 365 |
| Soft drinks | 1374 | 1506 | 1572 | 1256 |
| Sugars | 156 | 139 | 147 | 175 |

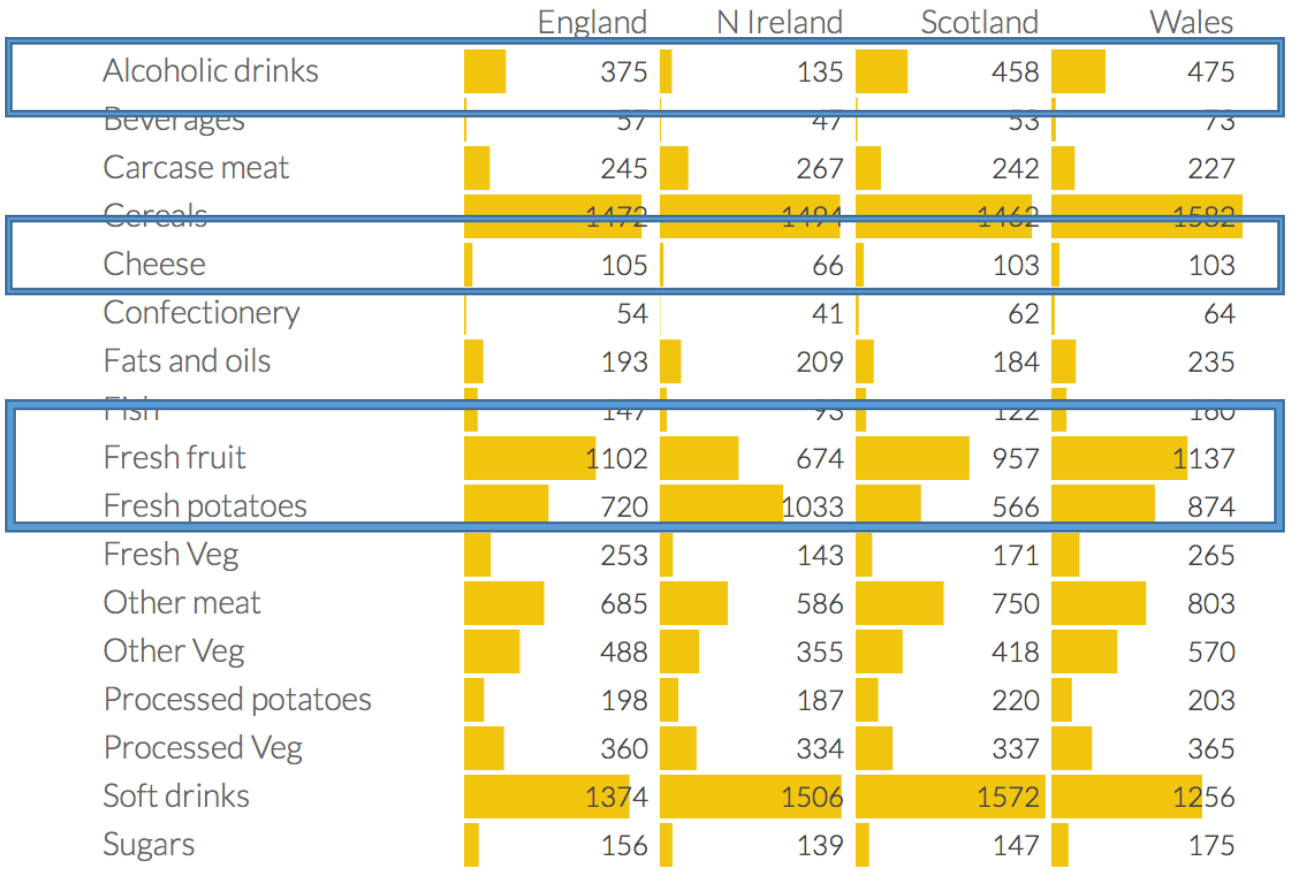
# Рацион в Великобритании



# Рацион в Великобритании



# Рацион в Великобритании



| | England | N Ireland | Scotland | Wales |
|---|---|---|---|---|
| Alcoholic drinks | 375 | 135 | 458 | 475 |
| Beverages | 57 | 47 | 53 | 73 |
| Carcase meat | 245 | 267 | 242 | 227 |
| Cereals | 1472 | 1494 | 1462 | 1582 |
| Cheese | 105 | 66 | 103 | 103 |
| Confectionery | 54 | 41 | 62 | 64 |
| Fats and oils | 193 | 209 | 184 | 235 |
| Fish | 147 | 93 | 122 | 160 |
| Fresh fruit | 1102 | 674 | 957 | 1137 |
| Fresh potatoes | 720 | 1033 | 566 | 874 |
| Fresh Veg | 253 | 143 | 171 | 265 |
| Other meat | 685 | 586 | 750 | 803 |
| Other Veg | 488 | 355 | 418 | 570 |
| Processed potatoes | 198 | 187 | 220 | 203 |
| Processed Veg | 360 | 334 | 337 | 365 |
| Soft drinks | 1374 | 1506 | 1572 | 1256 |
| Sugars | 156 | 139 | 147 | 175 |

# Ограничения

- Иногда выборка может лучше проецироваться не на прямую, а на некоторую кривую
- Существуют и другие методы уменьшения размерности

# Подробнее об PCA

- https://habr.com/ru/post/304214/
- http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D1%8B%D1%85_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%82