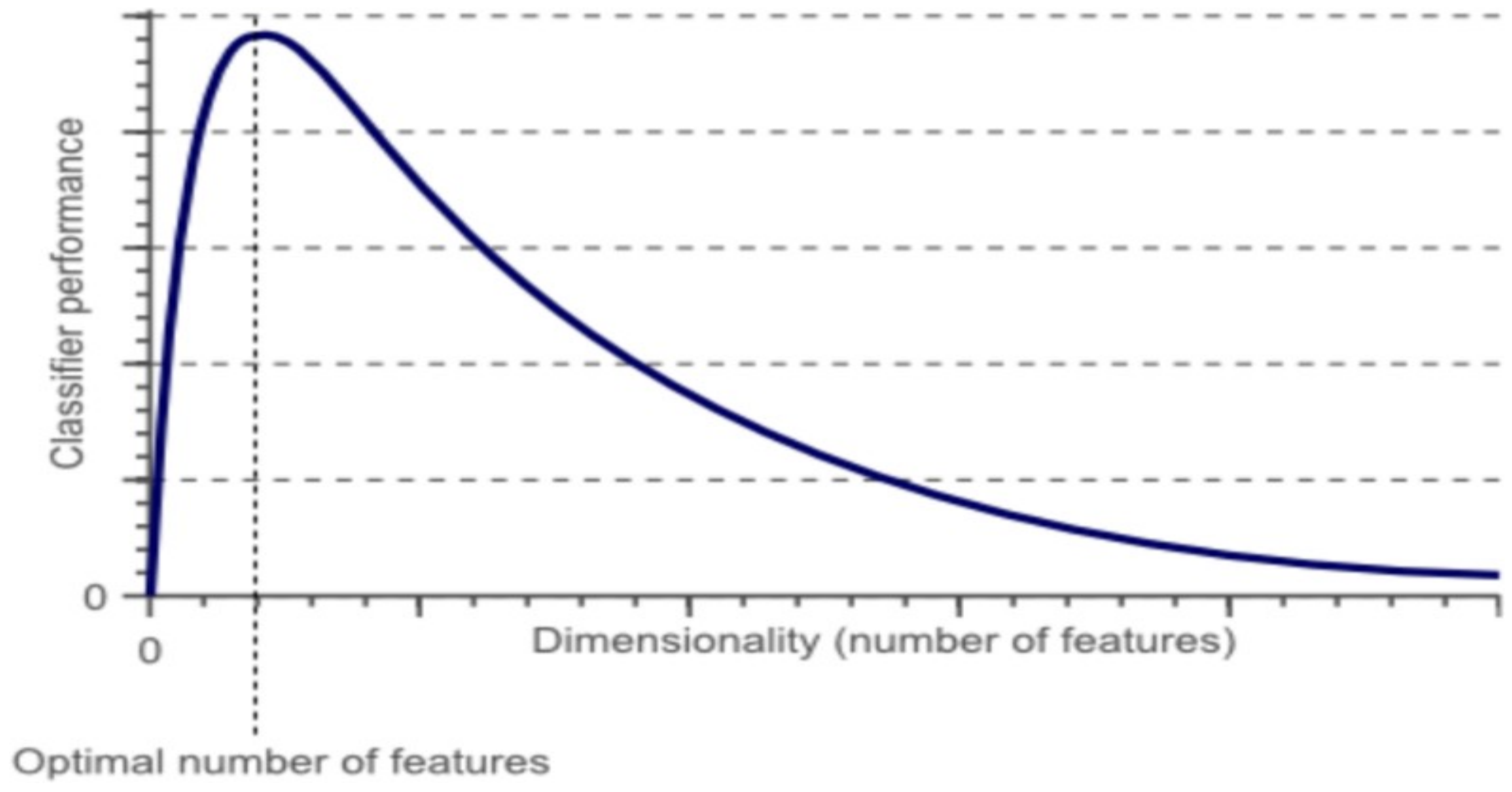
Занятие №11

Feature Selection

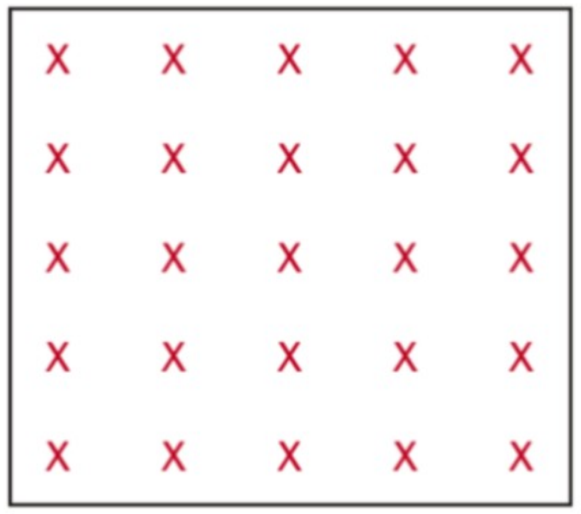
Содержание

Введение.
Зачем всё это?

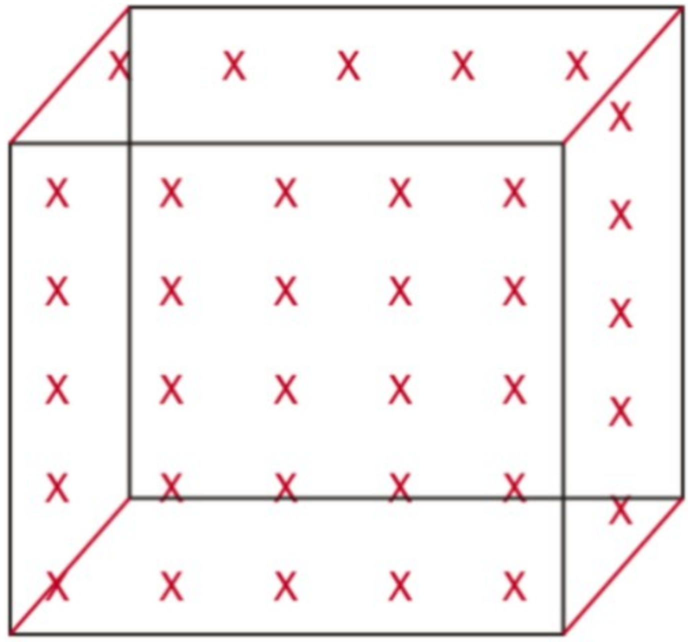
Проклятье размерности



Одно измерение - 5 точек



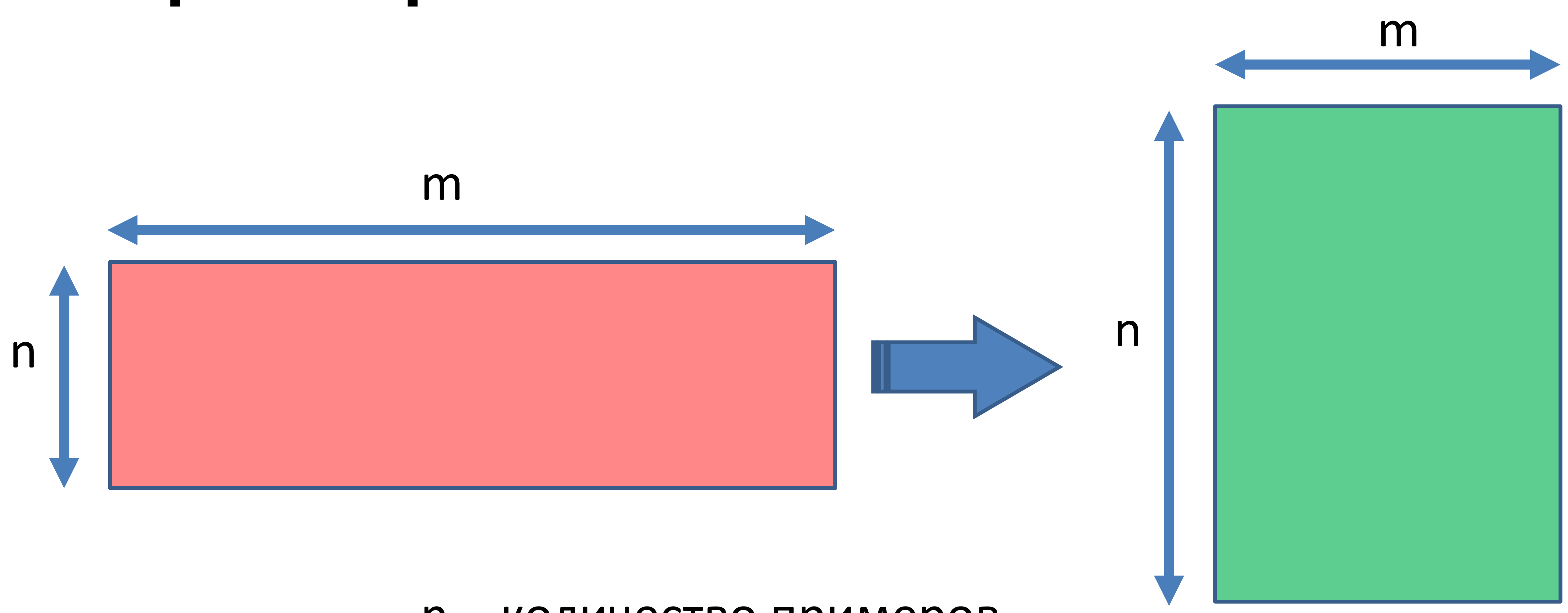
Два измерение - 25 точек



Три измерения - 125 точек

Методы отбора признаков

Позволит получить:
- упрощение моделей для того, чтобы сделать их проще для интерпретации исследователями или пользователями
- более короткое время тренировки
- уменьшения влияния проклятия размерности
- улучшение обобщения путём сокращения переобучения
- фильтрацию шумных признаков



n – количество примеров

m – количество признаков

Что можно сделать?

- Отобрать признаки
- Преобразовать признаки

Методы отбора признаков

Методы отбора

Задача – найти подмножество признаков на котором
выбранная модель покажет лучшее качество

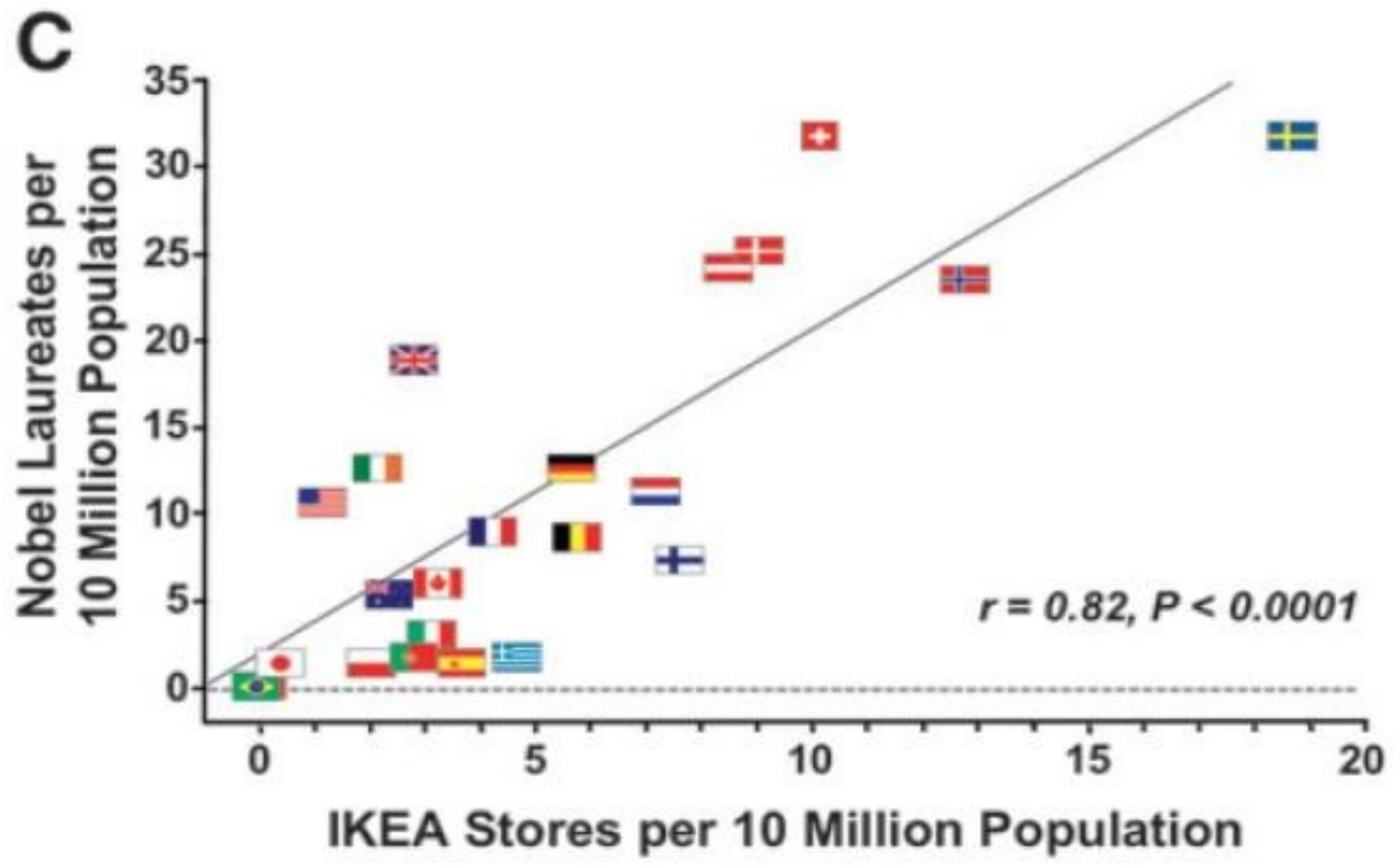
Фильтры
основаны на некоторых показателях, которые не зависит от метода классификации
(коэффициент корреляции, взаимная информация, WOE, IG)

Обертки
опираются на информацию о важности признаков полученную от других методов или моделей ML
(последовательный отбор и последовательное исключение признаков и др.)

Встроенные в алгоритмы
выполняют отбор признаков во время процедуры обучения классификатора,
и именно они явно оптимизируют набор используемых признаков для достижения
лучшей точности (регрессия с L1-регуляризация, Random Forest, SHAP)

Корреляция

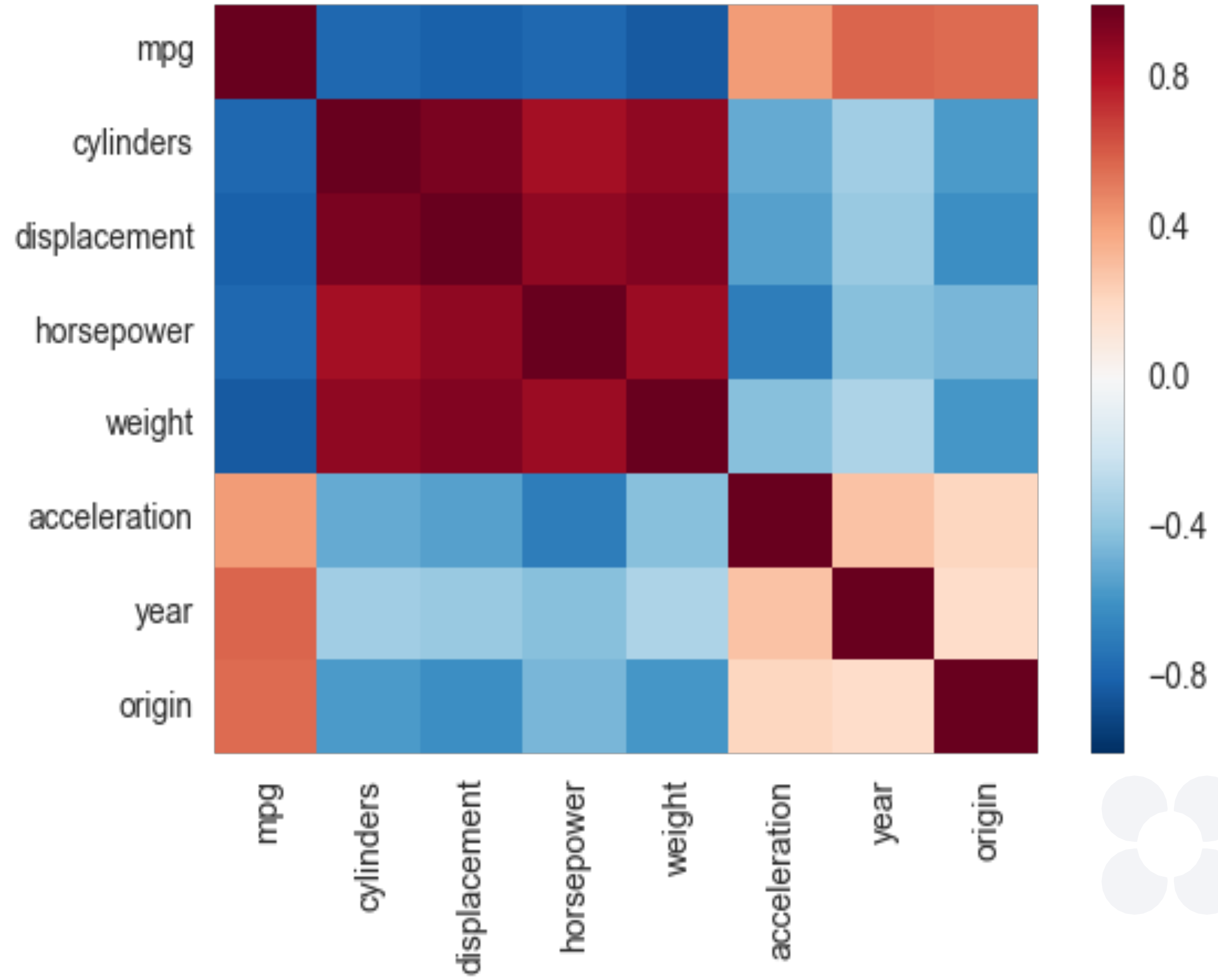
Корреля́ция — статистическая взаимосвязь двух или более случайных величин. При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.



ForexAW.com

Коэффициент корреляции Пирсона

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

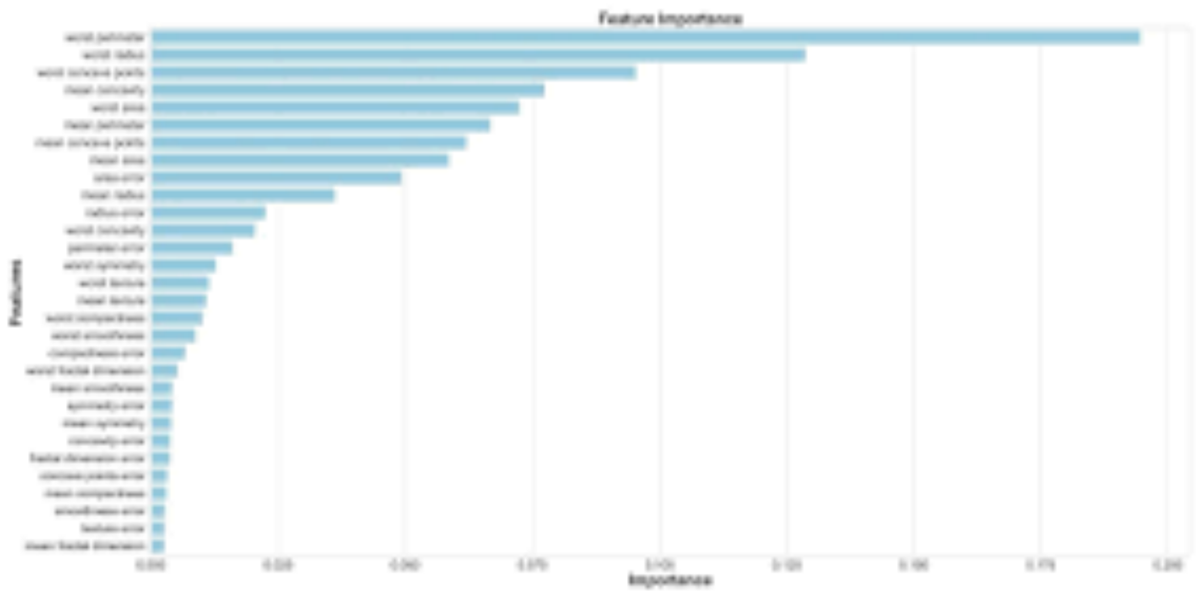


Ковариация

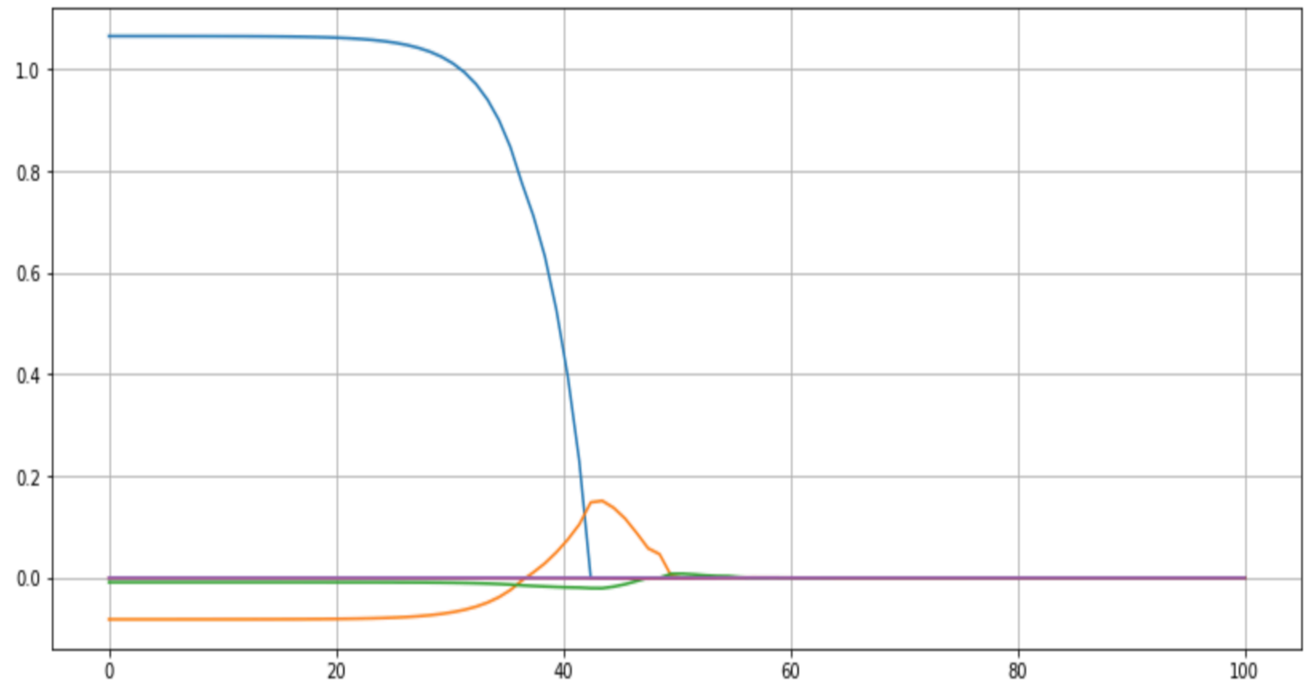
$$\mathrm{cov}_{XY} = \mathbf{M}\left[(X - \mathbf{M}(X))(Y - \mathbf{M}(Y))\right] = \mathbf{M}(XY) - \mathbf{M}(X)\mathbf{M}(Y)$$

Википедия http://ru.wikipedia.org/wiki/Корреляция

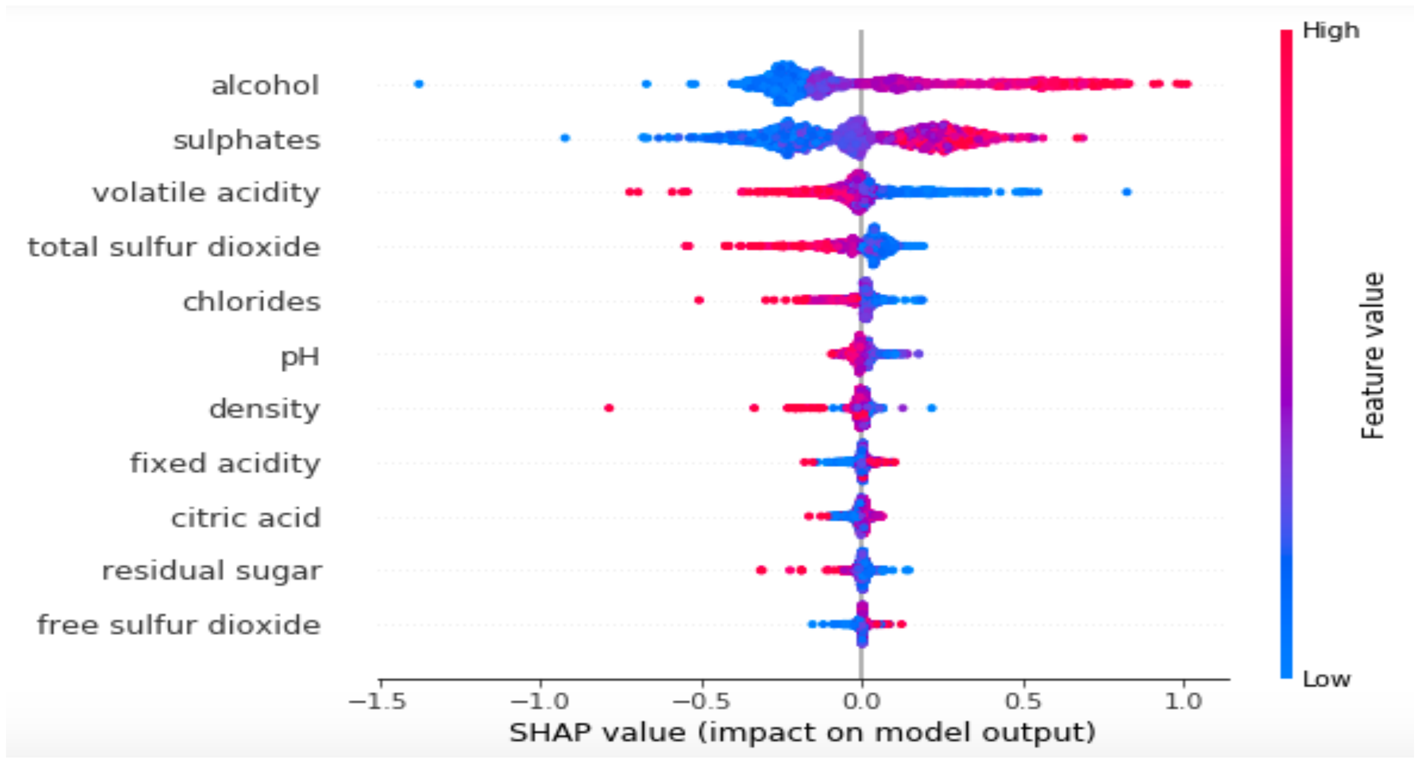
Random Forest



L1 - регуляризация



SHAP (SHapley Additive exPlanations)

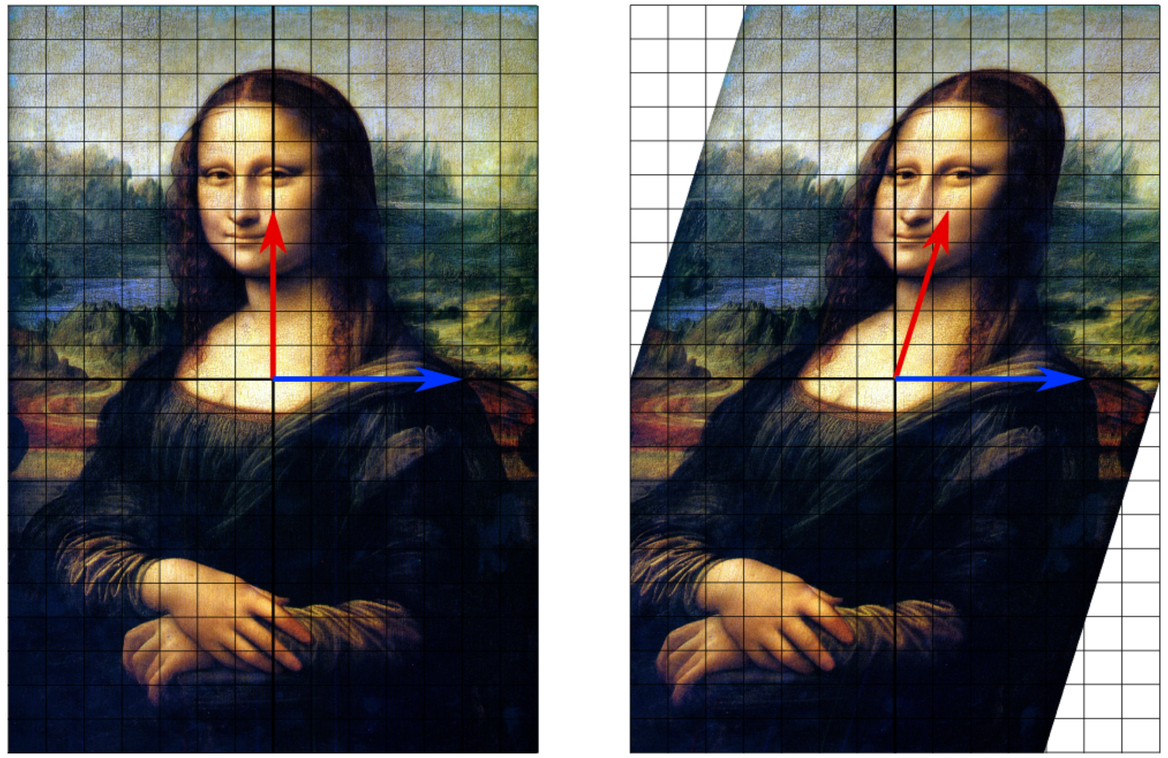


Преобразование признаков

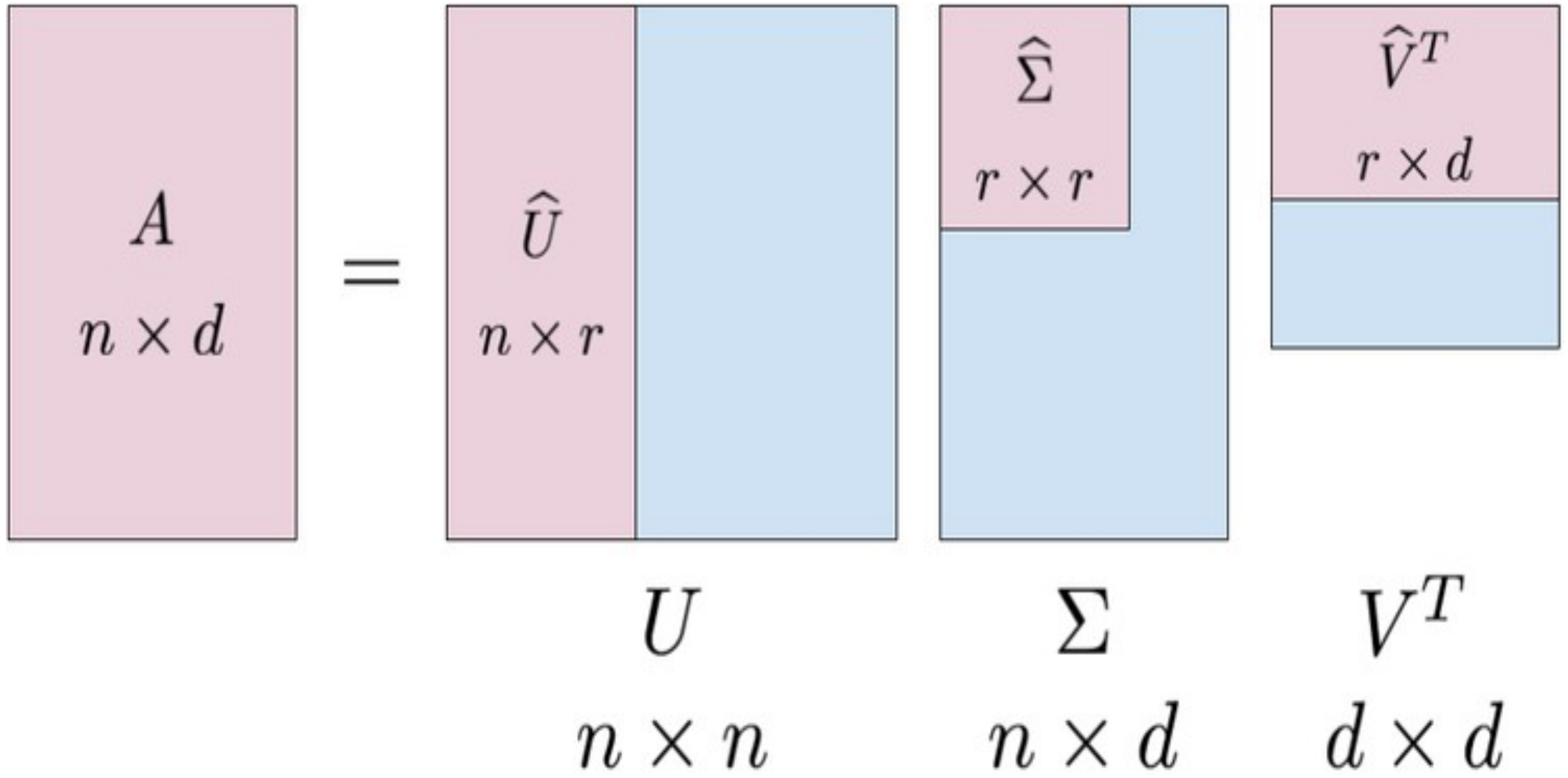
Метод главных компонент (principal component analysis, PCA):
позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры

Собственный вектор

$$M\vec{x} = \lambda\vec{x}$$

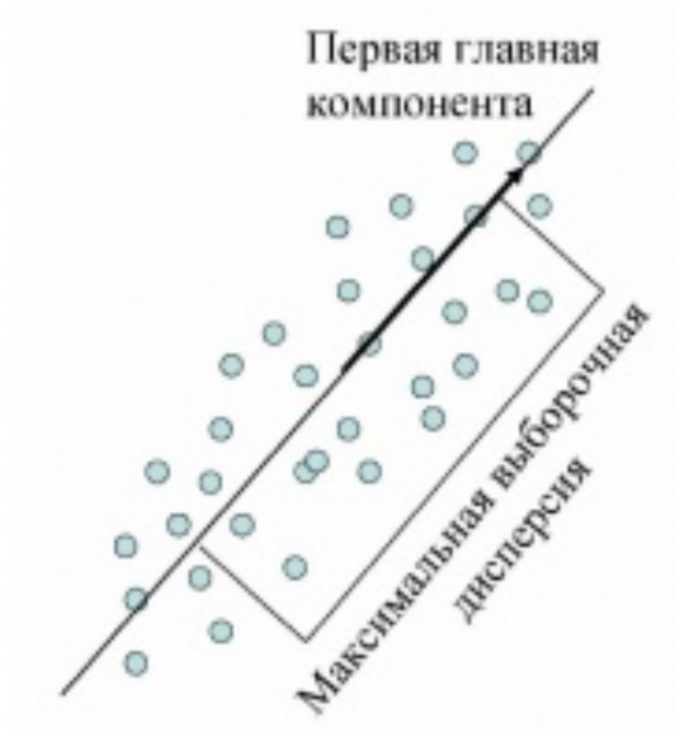


Сингулярное разложение (SVD)

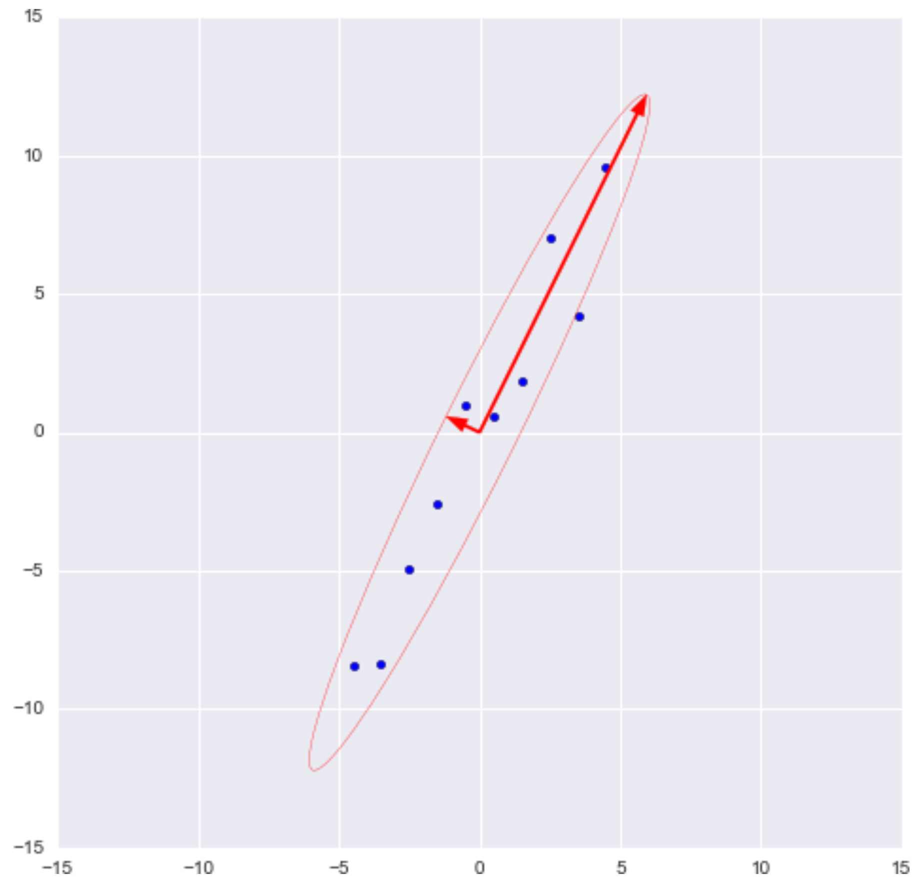


PCA

Зачем он нужен? Он уменьшает размерность с минимумом потери информации



- перевести данные в пространство меньшей размерности
- найти такое преобразование при котором разброс данных и дисперсия в ортогональной проекциях максимален
- корреляция между отдельными координатами обратятся в ноль.



$$Cov(X_i, X_j) = E\Big[\big(X_i - E(X_i)\big)\cdot\big(X_j - E(X_j)\big)\Big] = E(X_iX_j) - E(X_i)\cdot E(X_j)$$

$$Var(X^*) = \Sigma^* = E(X^*\cdot X^{*T}) = E\Big((\vec{v}^T X)\cdot(\vec{v}^T X)^T\Big) =$$

$$= E(\vec{v}^T X\cdot X^T\vec{v}) = \vec{v}^T E(X\cdot X^T)\vec{v} = \vec{v}^T\Sigma\vec{v}$$

LDA

Линейный дискриминантный анализ

Метод уменьшения размерности, используемый в качестве этапа предварительной обработки в приложениях машинного обучения и классификации.

Первый шаг - вычислить разделимость между разными классами (то есть расстояние между средними значениями разных классов), также называемое межклассовой дисперсией.

$$S_b = \sum_{i=1}^{g} N_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T$$

Второй шаг - вычислить расстояние между средним значением и выборкой каждого класса, которое называется внутриклассовой дисперсией.

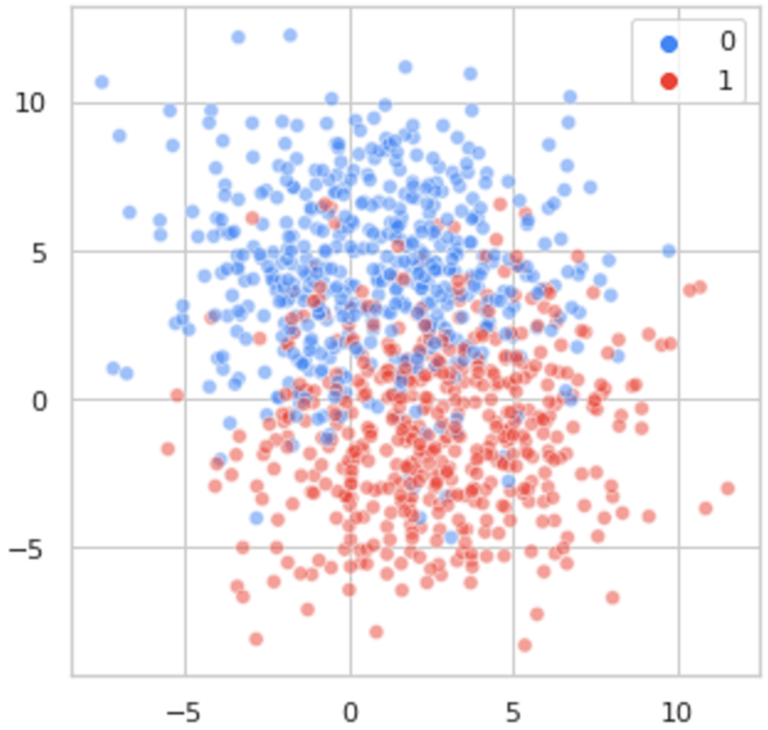
$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x}_i)(x_{i,j} - \overline{x}_i)^T$$

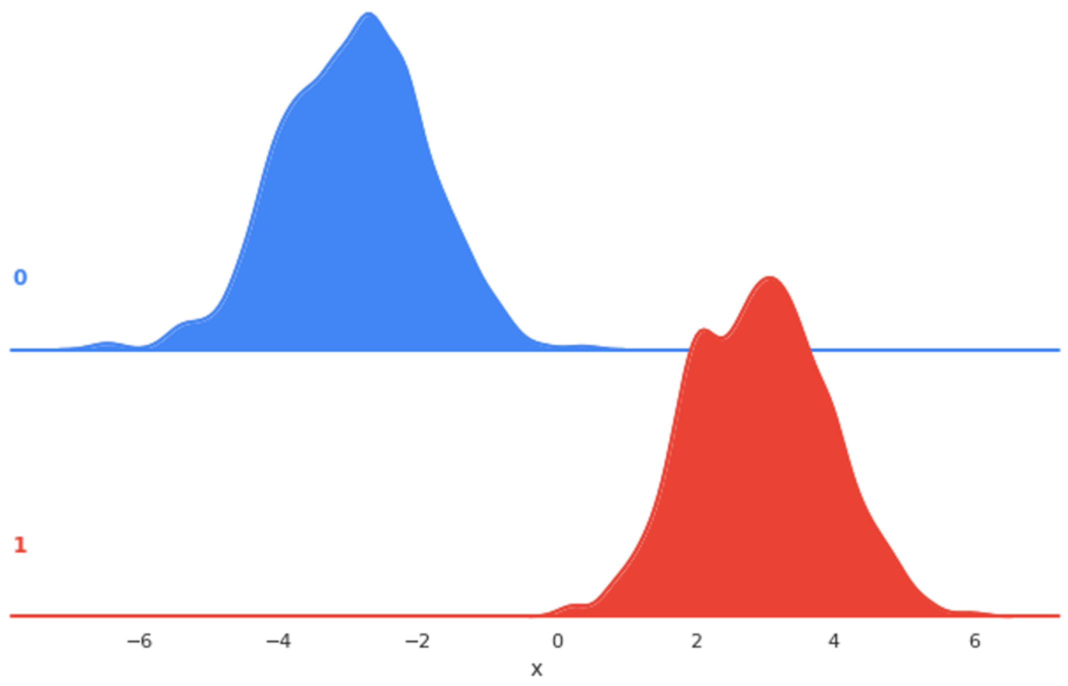
Третий шаг - построить пространство более низкой размерности, которое максимизирует дисперсию между классами и минимизирует дисперсию внутри класса.
P - проекция пространства нижней размерности, которая называется критерием Фишера

$$P_{lda} = \arg\max_{P} \frac{\left|P^T S_b P\right|}{\left|P^T S_w P\right|}$$

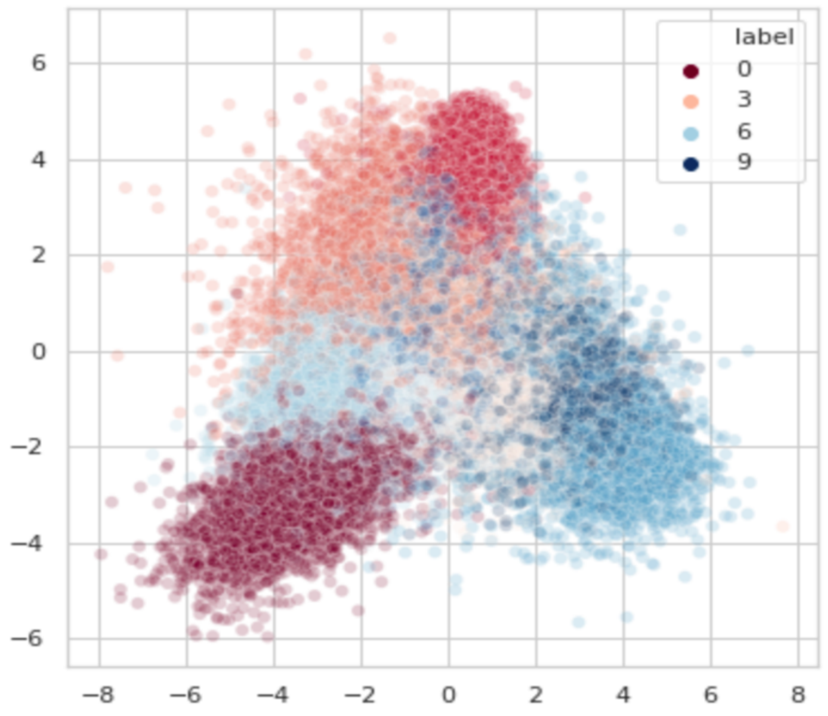
LDA

Отображение распределение
в 1- мерное пространство

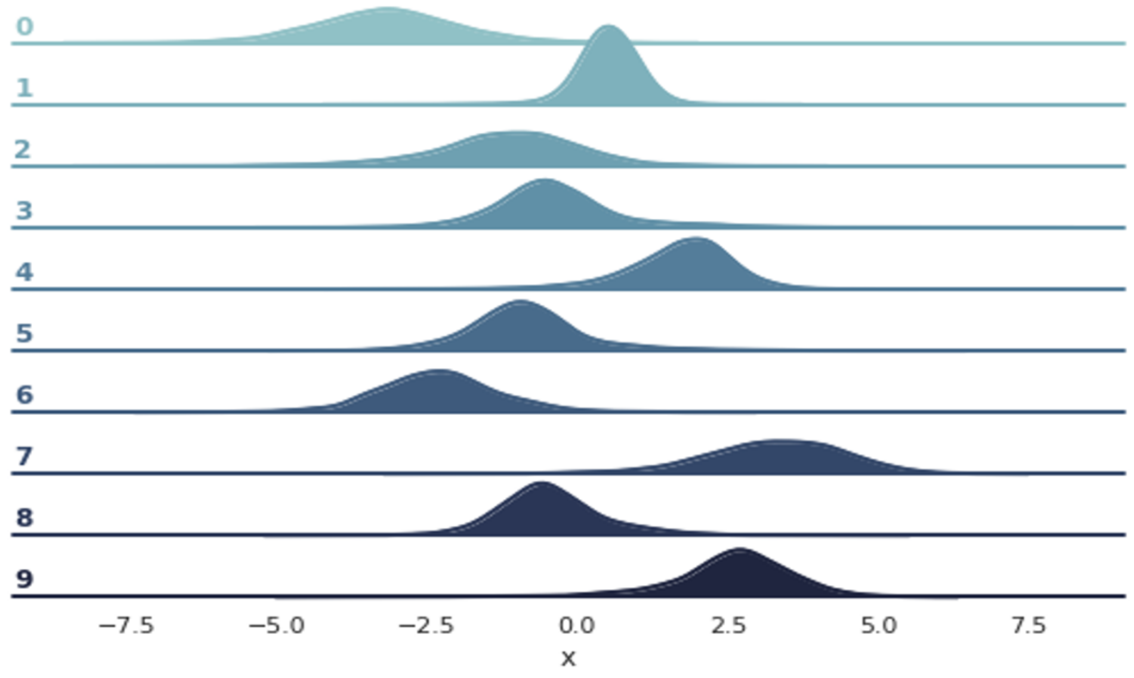




Two-Dimensional Representation

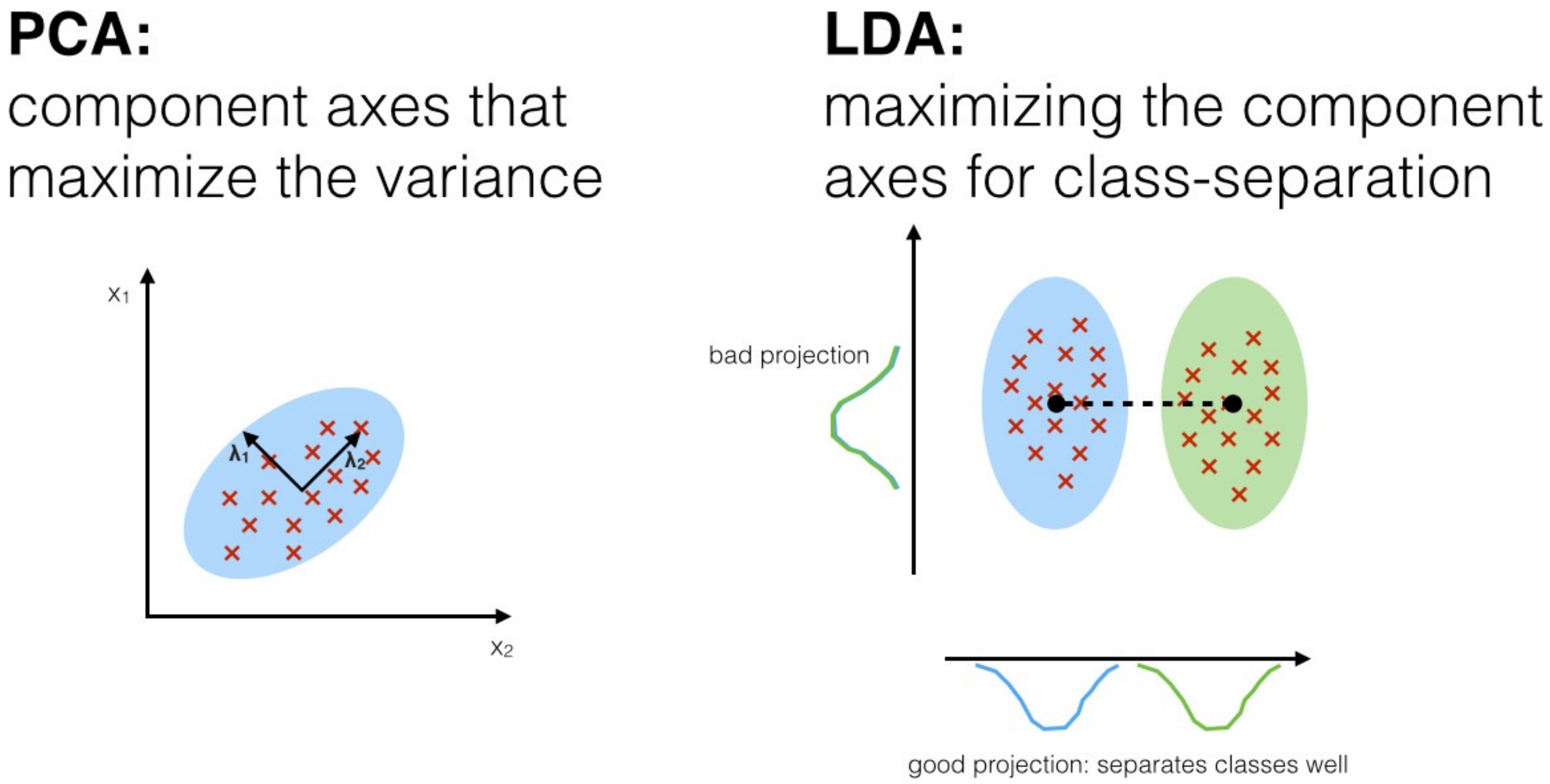


One-Dimensional Representation



Отображение картинок MNIST
в 2- и 1- мерное пространство

Сравнение LDA и PCA



ПРАКТИКА

Спасибо
за
внимание!