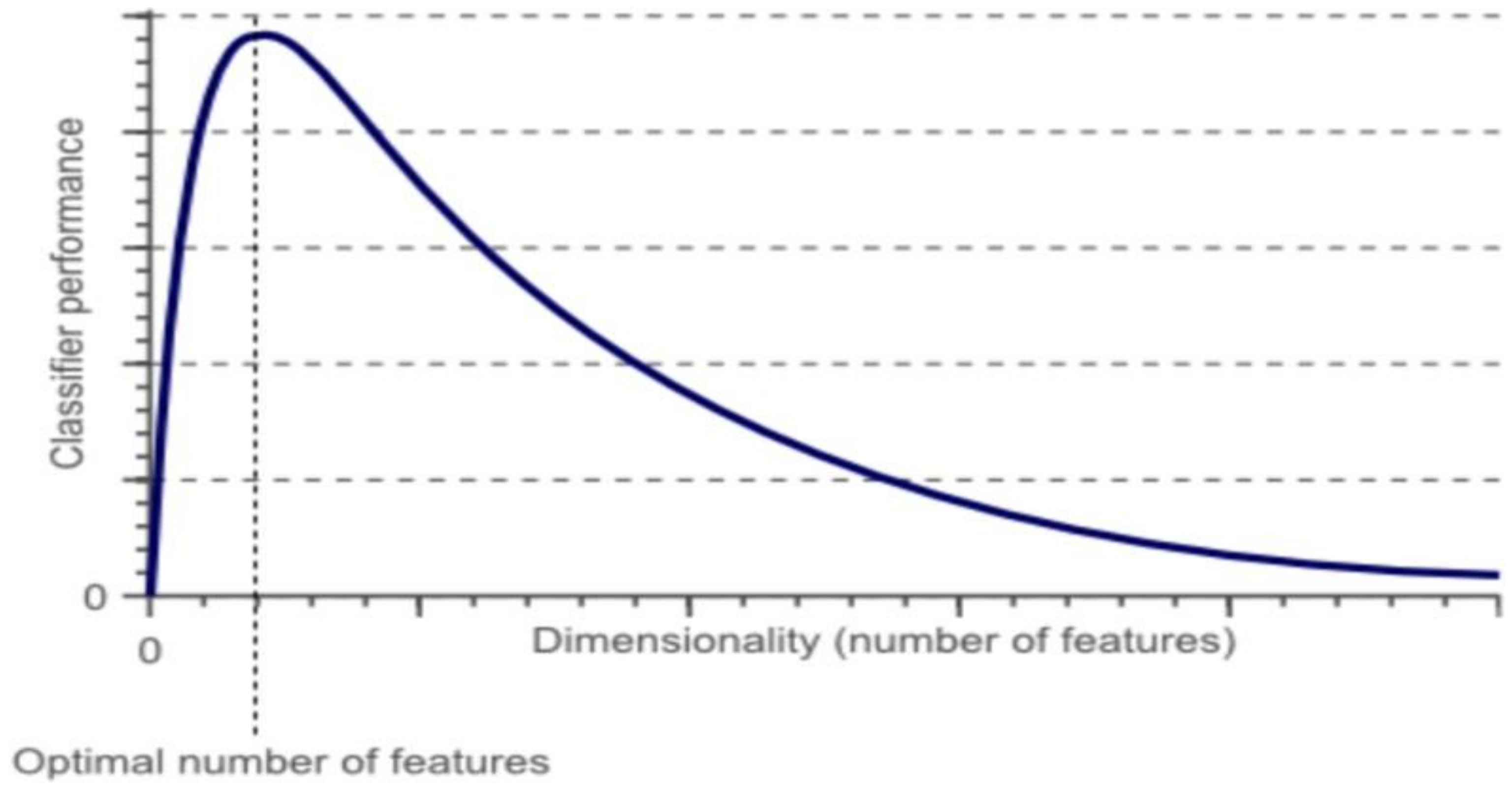
Занятие № 11

Feature Selection

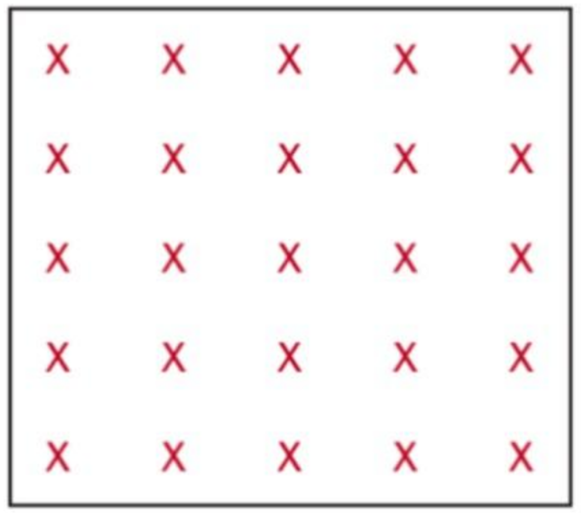
Содержание

Введение.
Зачем всё это?

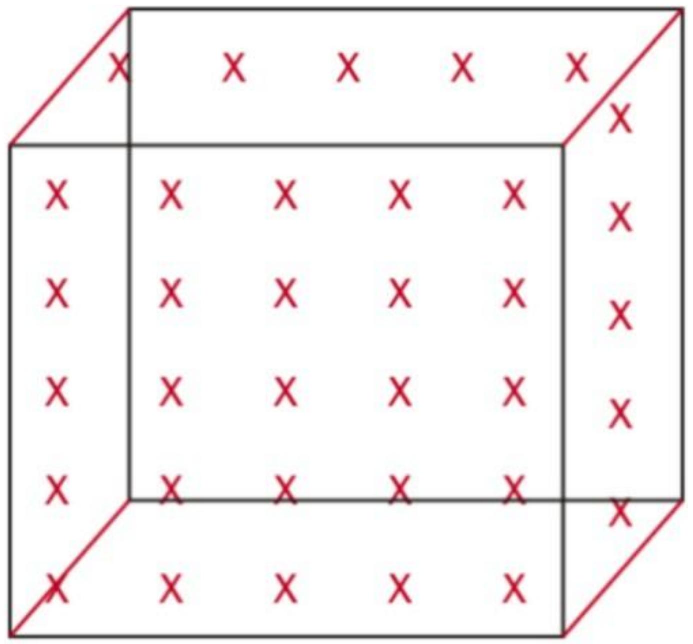
Проклятье размерности



Одно измерение - 5 точек



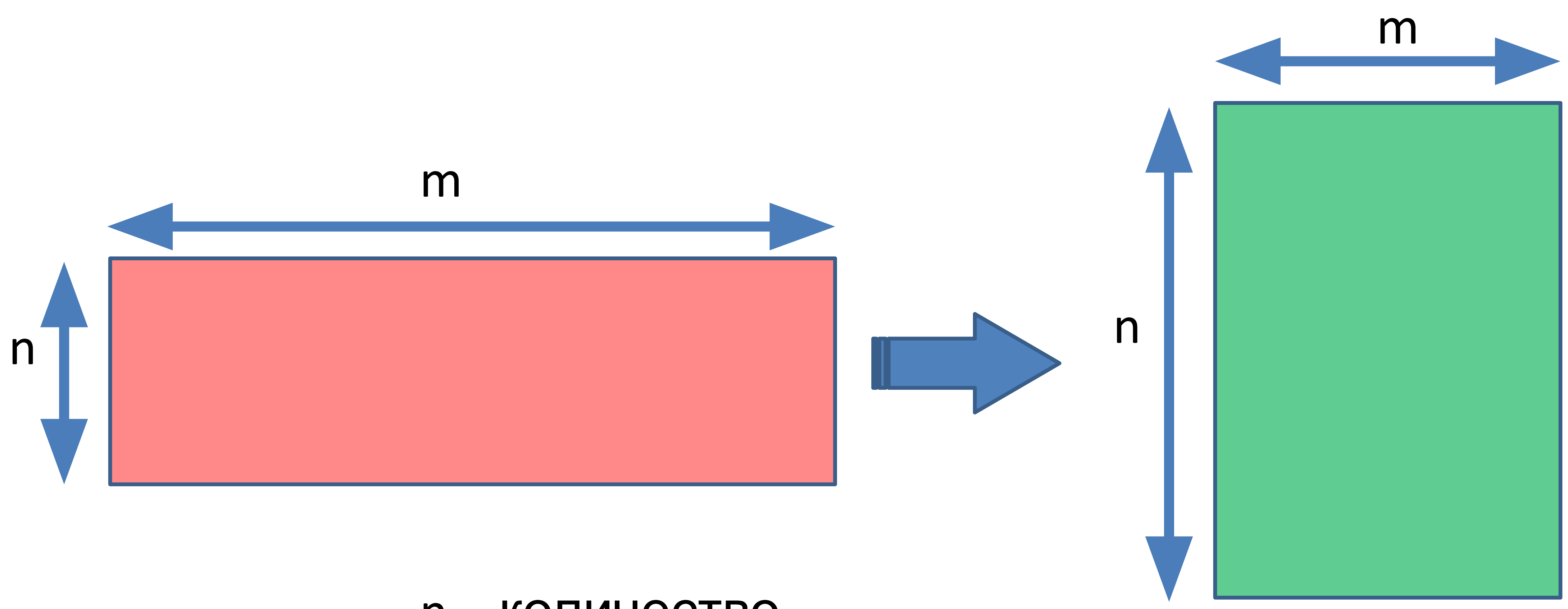
Два измерение - 25 точек



Три измерения - 125 точек

Методы отбора признаков

Позволит получить:
- упрощение моделей для повышения возможности 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

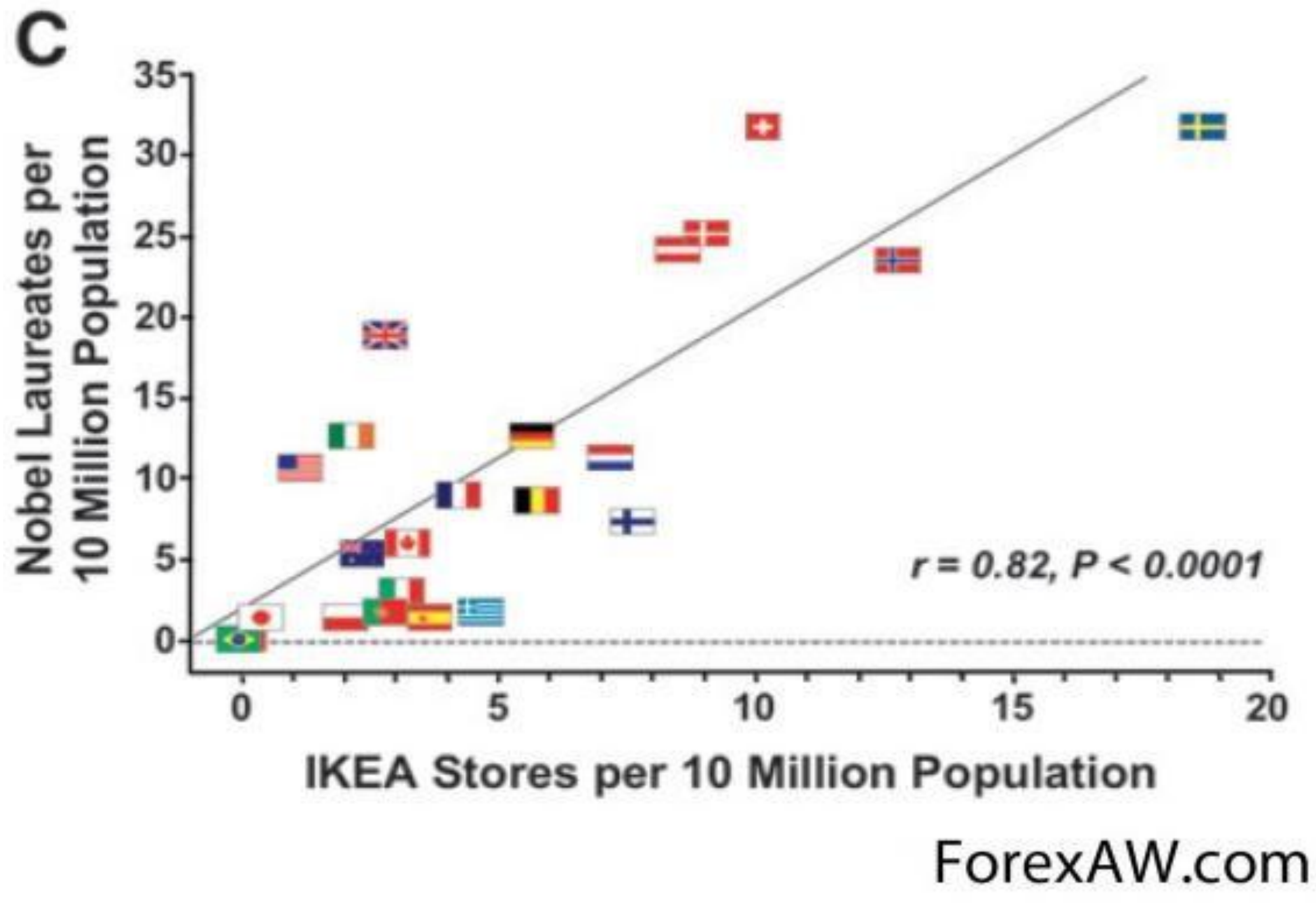
Фильтры
основаны на некоторых показателях, которые не зависит от метода классификации
(коэффициент корреляции, взаимная информация, WOE, IG)

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)

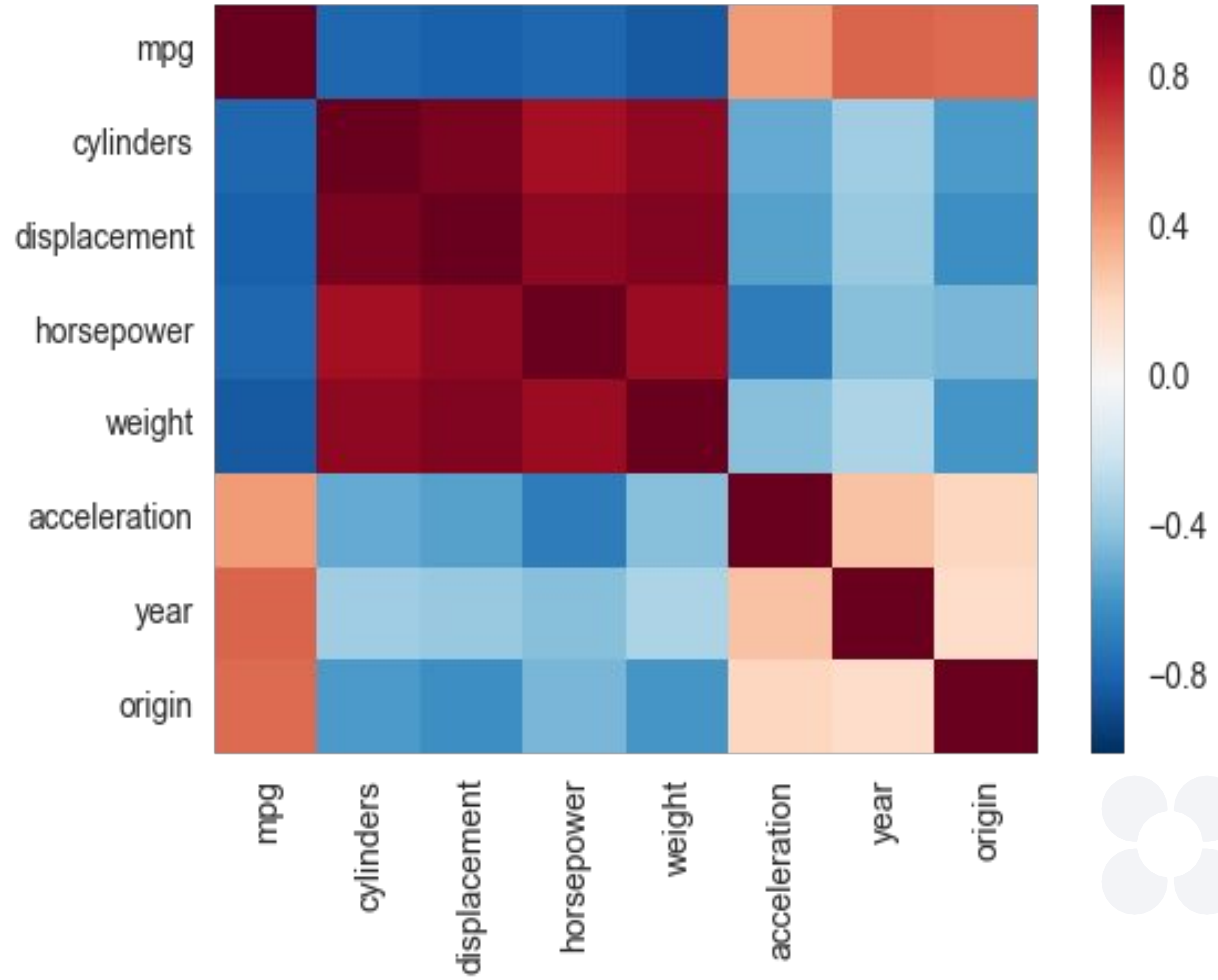
Корреляция

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.



Коэффициент корреляции Пирсона

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

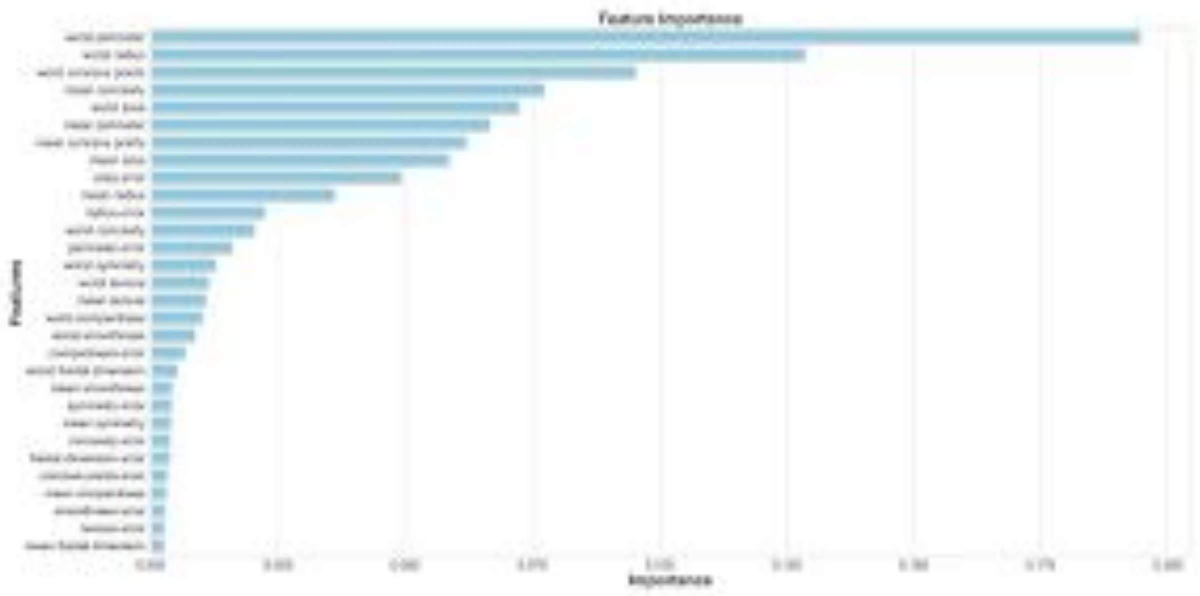


Ковариация

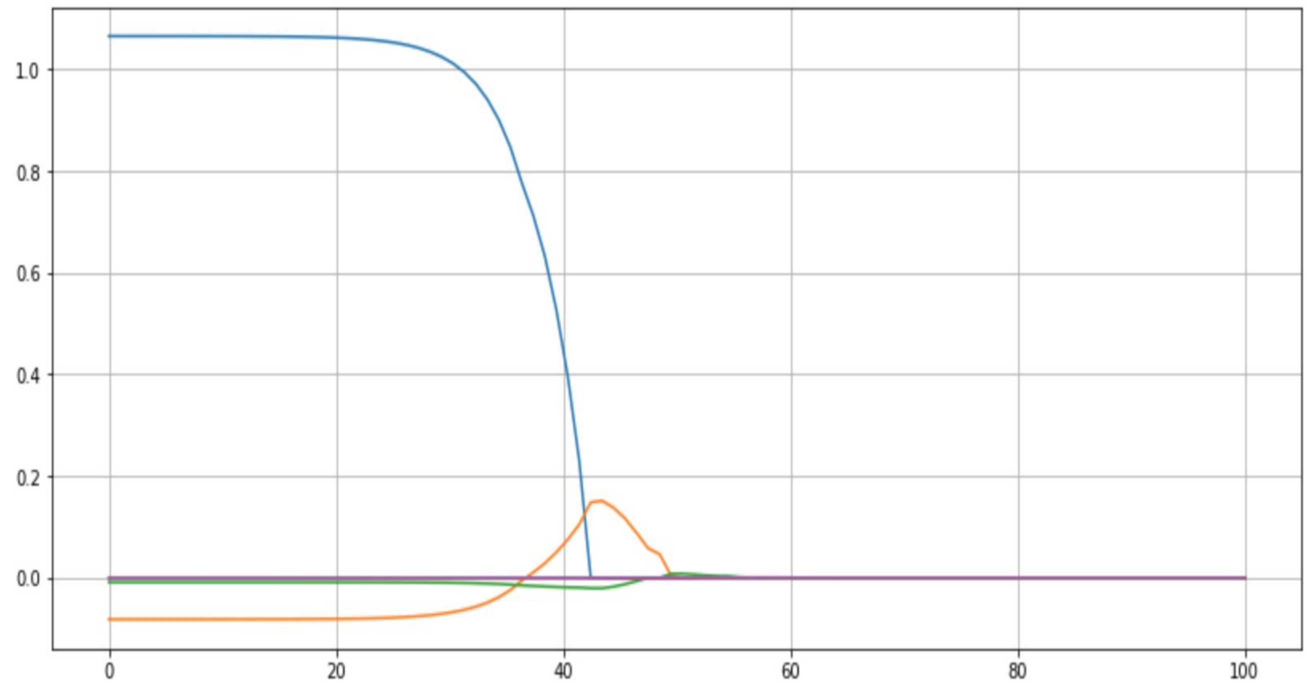
$$\mathrm{cov}_{XY} = \mathbf{M}\left[(X-\mathbf{M}(X))(Y-\mathbf{M}(Y))\right] = \mathbf{M}(XY) - \mathbf{M}(X)\mathbf{M}(Y)$$

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

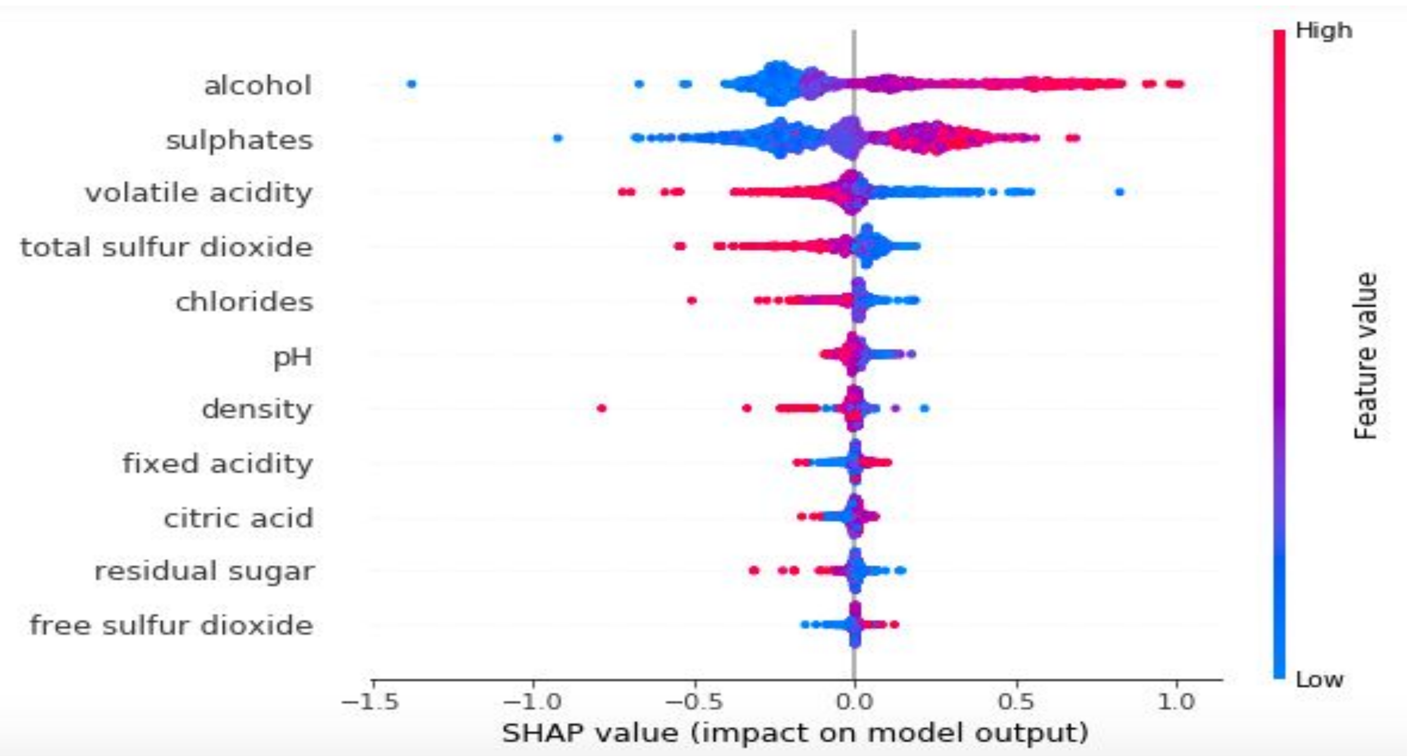
Random Forest



L1 - регуляризация



SHAP (SHapley Additive exPlanations)

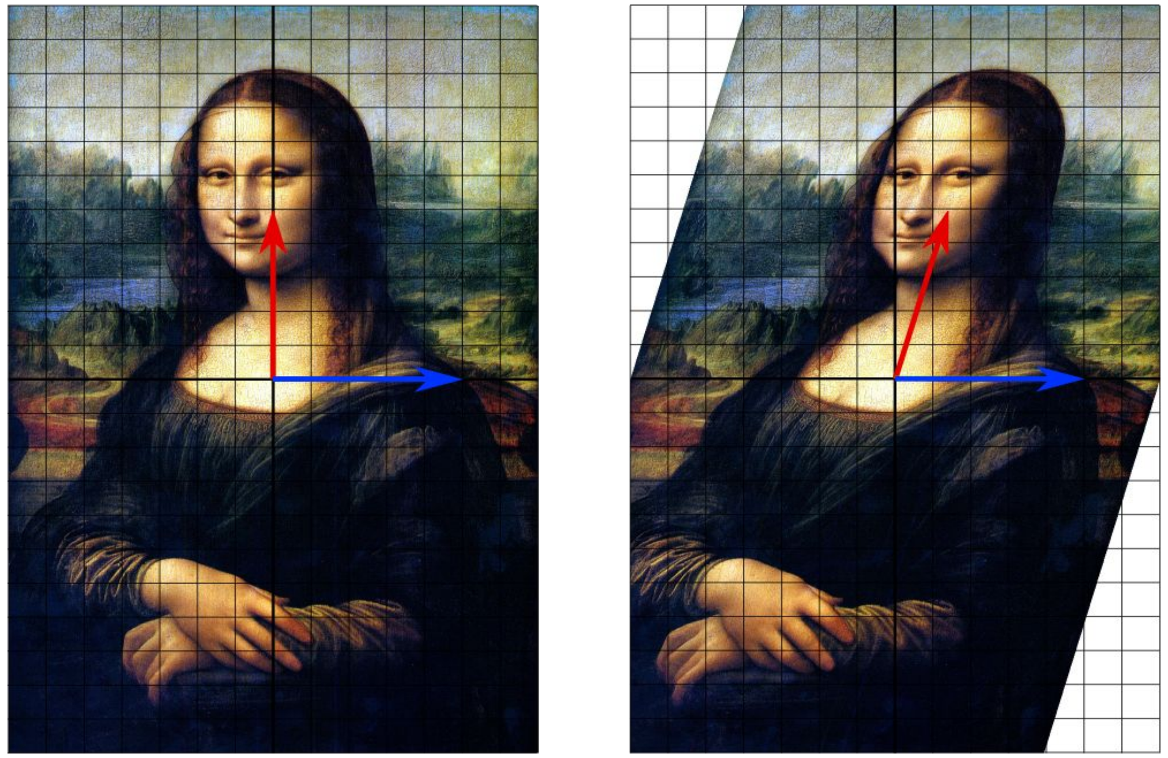


Преобразование признаков

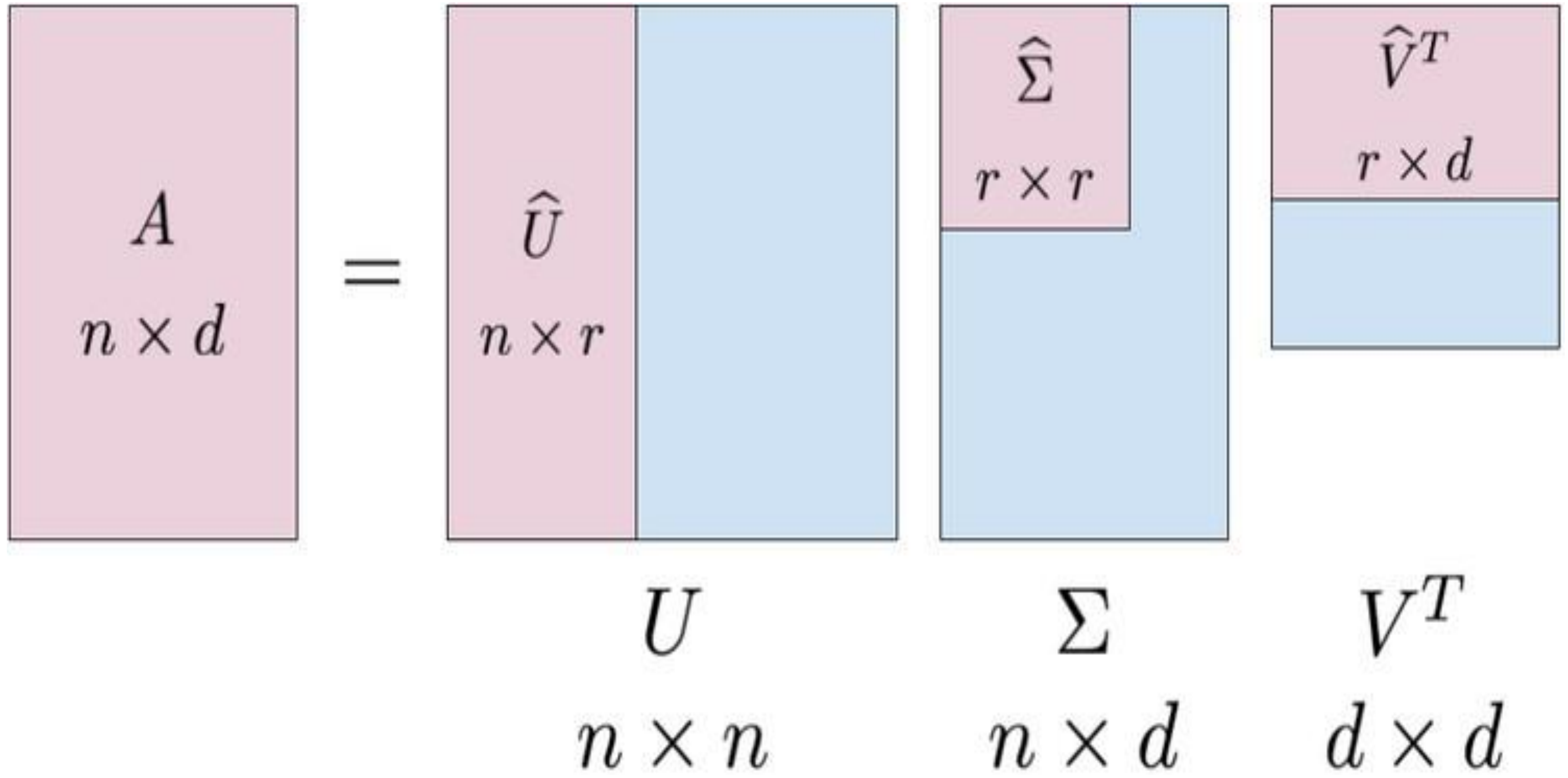
Метод главных компонент (principal component analysis, PCA): позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры

Собственный вектор

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

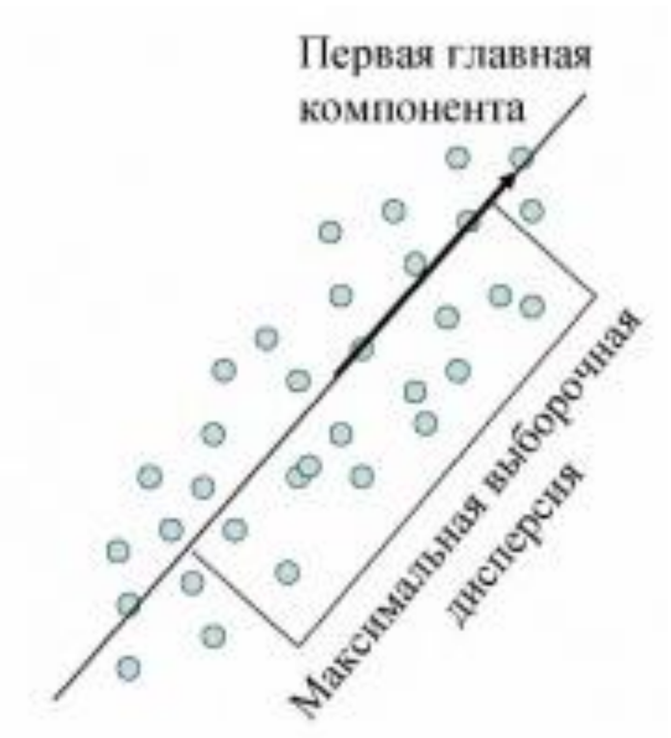


Сингулярное разложение (SVD)

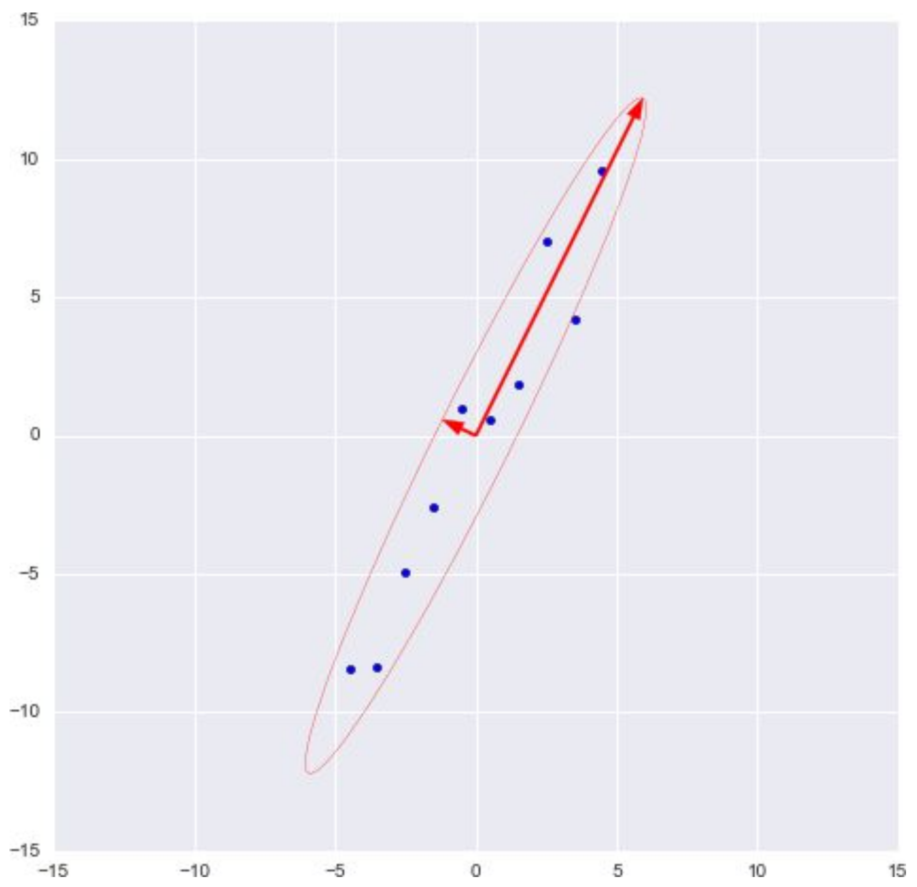


PCA

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.



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

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

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

https://cmi.to/методы/метод-главных-компонент/

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.

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

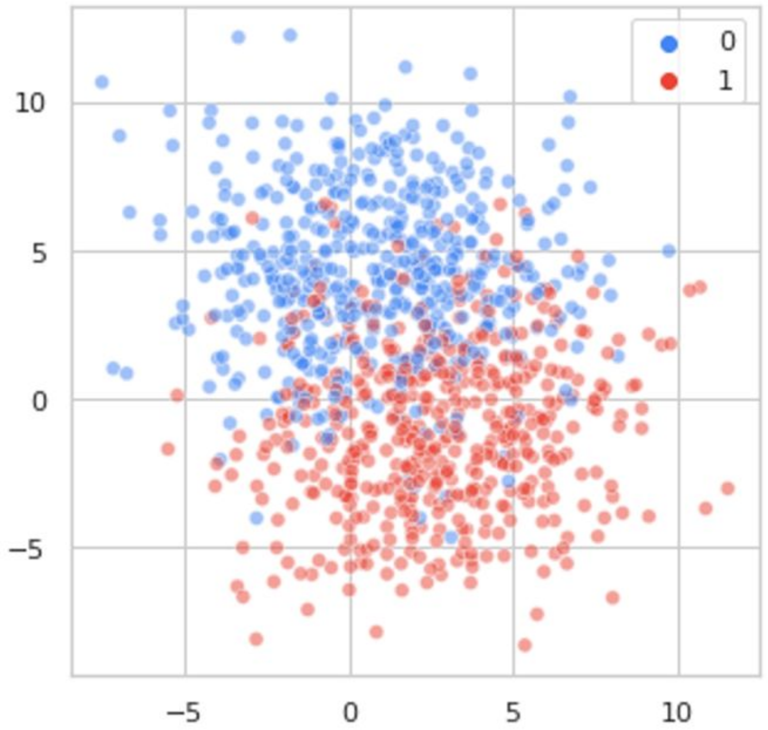
Третий шаг - построить пространство более низкой размерности, которое максимизирует дисперсию между классами и минимизирует дисперсию внутри класса.
P - проекция пространства нижней размерности

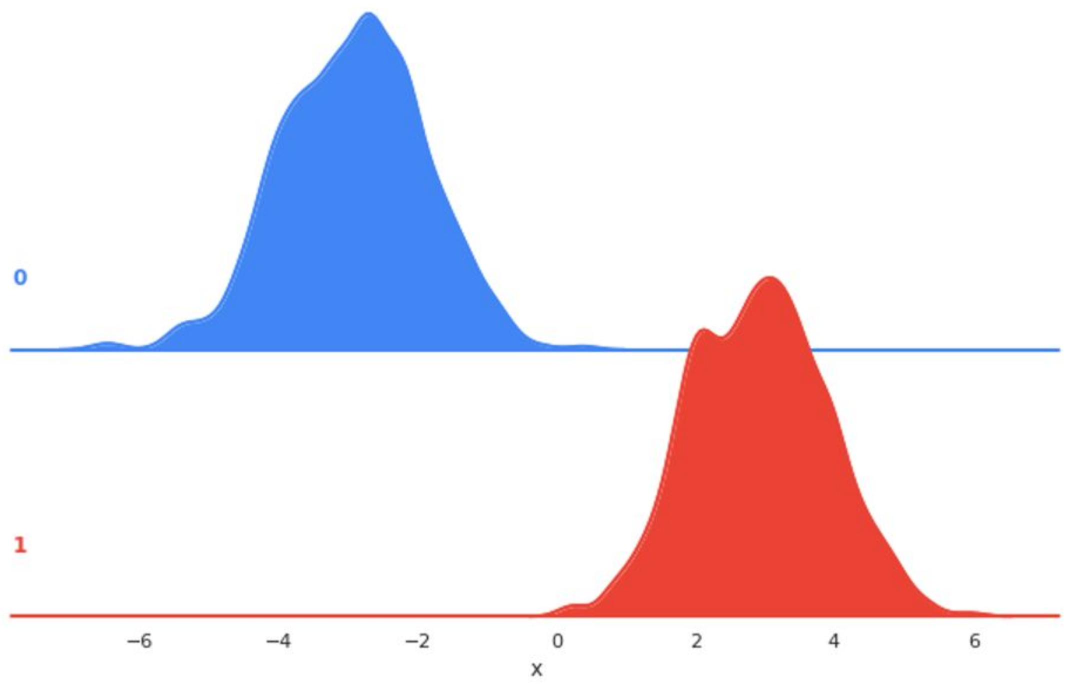
$$P_{lda} = \arg\max_{P} \frac{\left|P^T S_b P\right|}{\left|P^T S_w P\right|}$$

https://cmi.to/методы/линейный-дискриминантный-анализ/

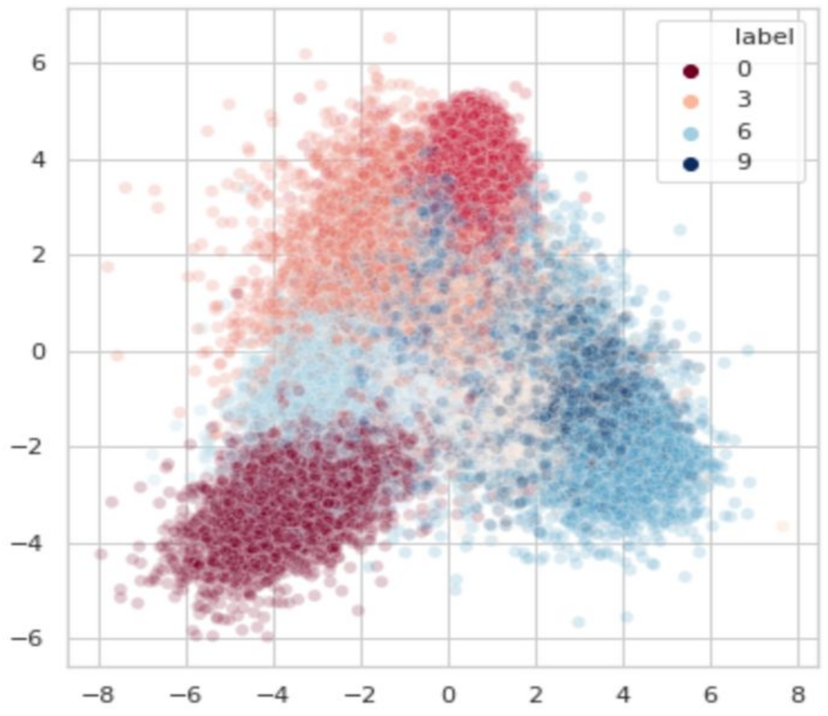
LDA

Отображение распределение
в 1- мерное пространство

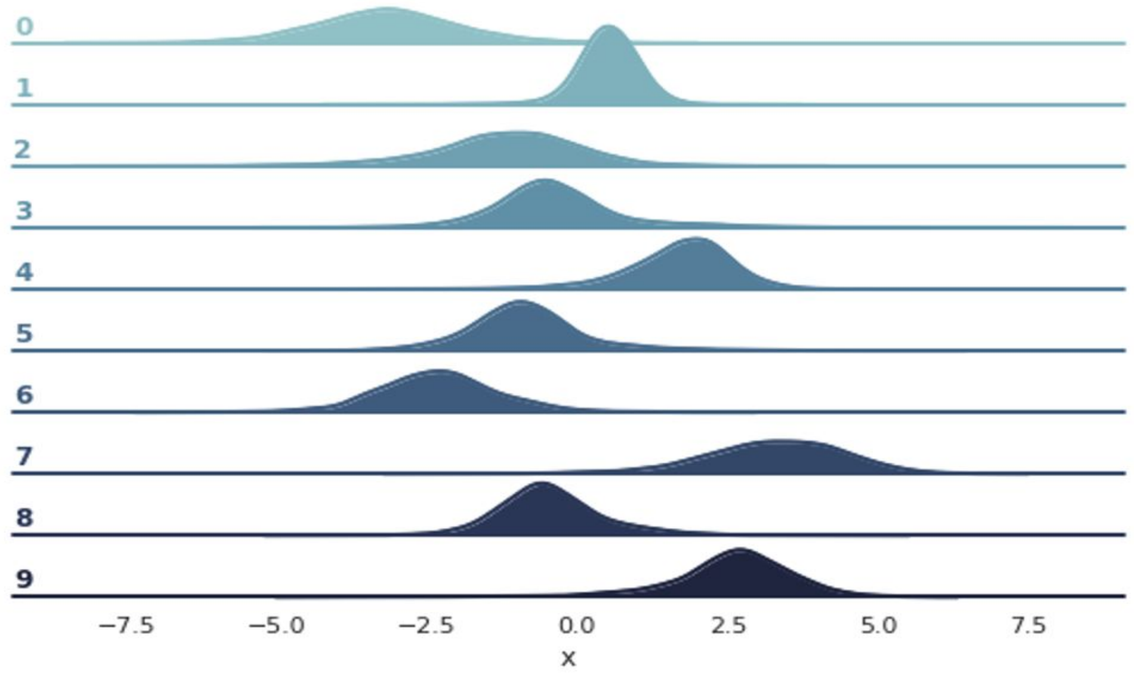




Two-Dimensional Representation



One-Dimensional Representation



Отображение картинок
MNIST
в 2- и 1- мерное
пространство

NCA Анализ компонентов соседства

Метод уменьшения размерности, используемый в качестве этапа предварительной обработки в приложениях машинного обучения и классификации. Решаемая задача максимизации качества классификации методом К-средних, с использованием концепции *стохастических ближайших соседей*

$$\arg\max_{L} \sum_{i=0}^{N-1} p_i$$

Используется расстояние Махаланобиса

$$d(x,y) = ||L\,(x-y)||^2 = (x-y)^\top Q(x-y) = (Ax-Ay)^\top (Ax-Ay)$$

Класс точки определяется взвешенным объединением
классов всех остальных точек

$$p_i = \sum_{j \in C_i} p_{ij}$$

$$p_{ii} = 0 \qquad p_{ij} = \frac{\exp(-\|Ax_i - Ax_j\|^2)}{\sum_{k \neq i} \exp(-\|Ax_i - Ax_k\|^2)}$$

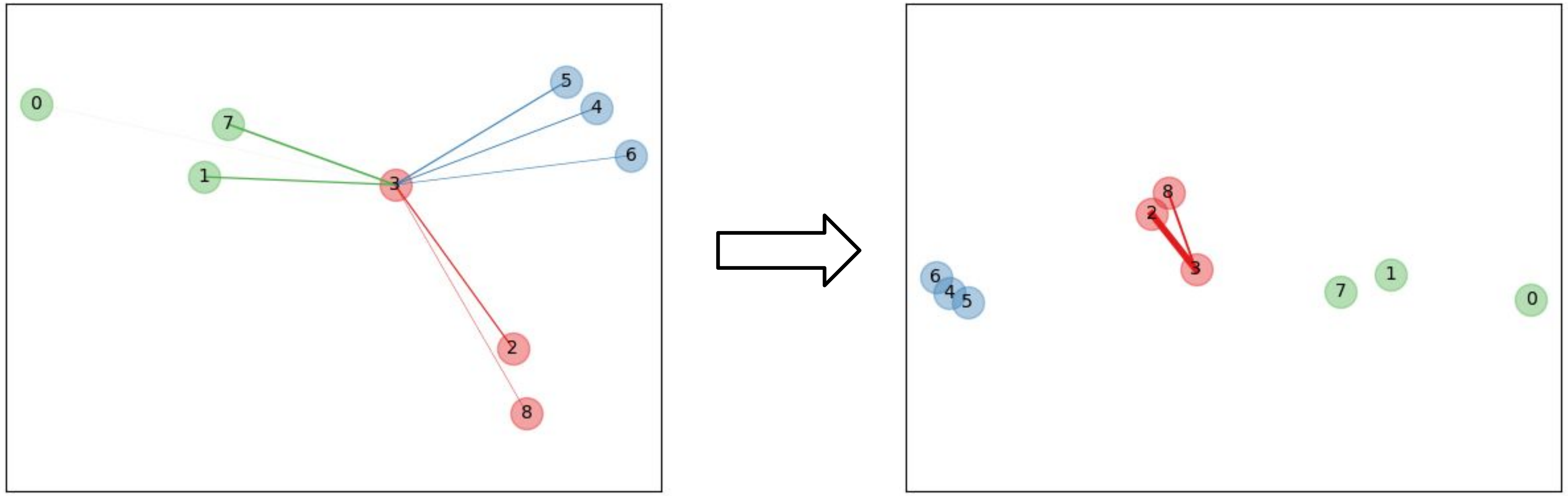
Приближение матрицы преобразования происходит градиентными итеративными методами

$$\frac{\partial f}{\partial A} = 2A \sum_i \left(p_i \sum_k p_{ik} x_{ik} x_{ik}^\top - \sum_{j \in C_i} p_{ij} x_{ij} x_{ij}^\top \right)$$

NCA

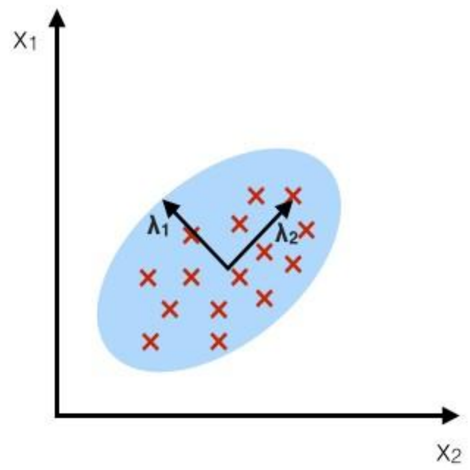
Анализ компонентов соседства

Метод уменьшения размерности, используемый в качестве этапа предварительной обработки в приложениях машинного обучения и классификации. Решаемая задача максимизации качества классификации методом K-средних, с использованием концепции *стохастических ближайших соседей*

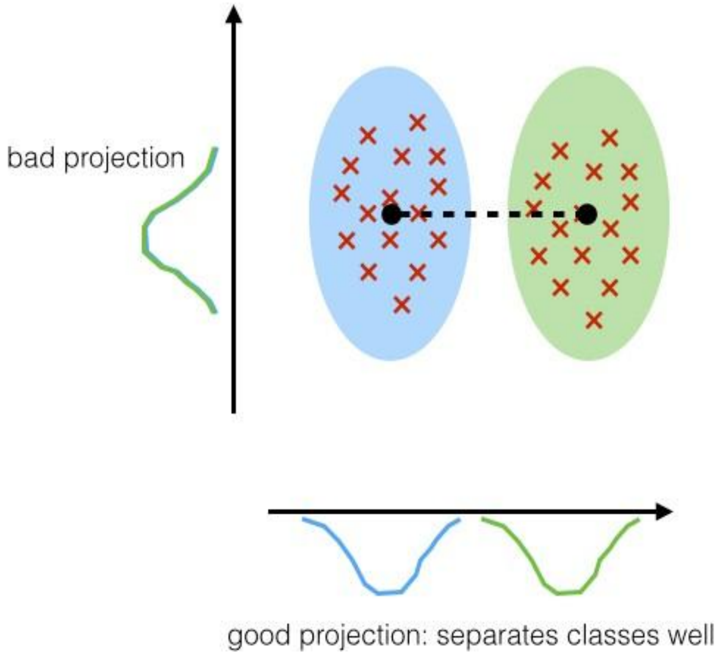


Сравнение LDA, PCA и NCA

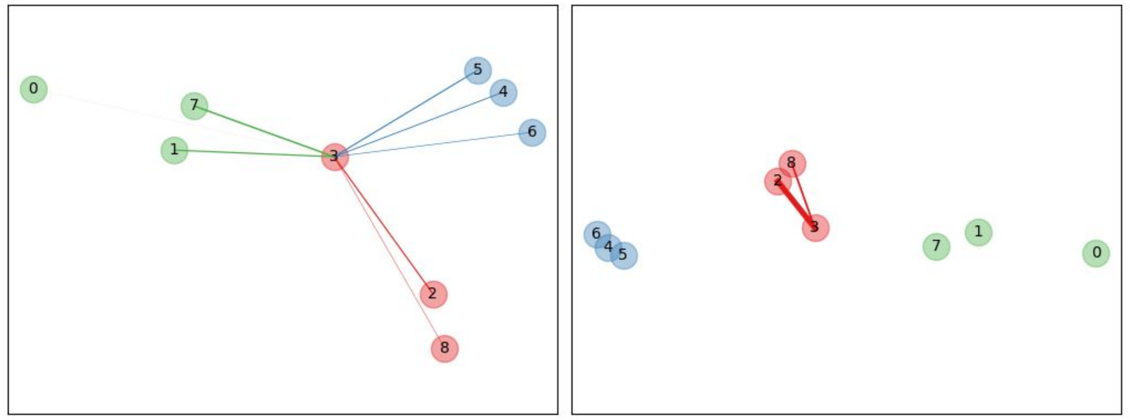
PCA:
component axes that
maximize the variance

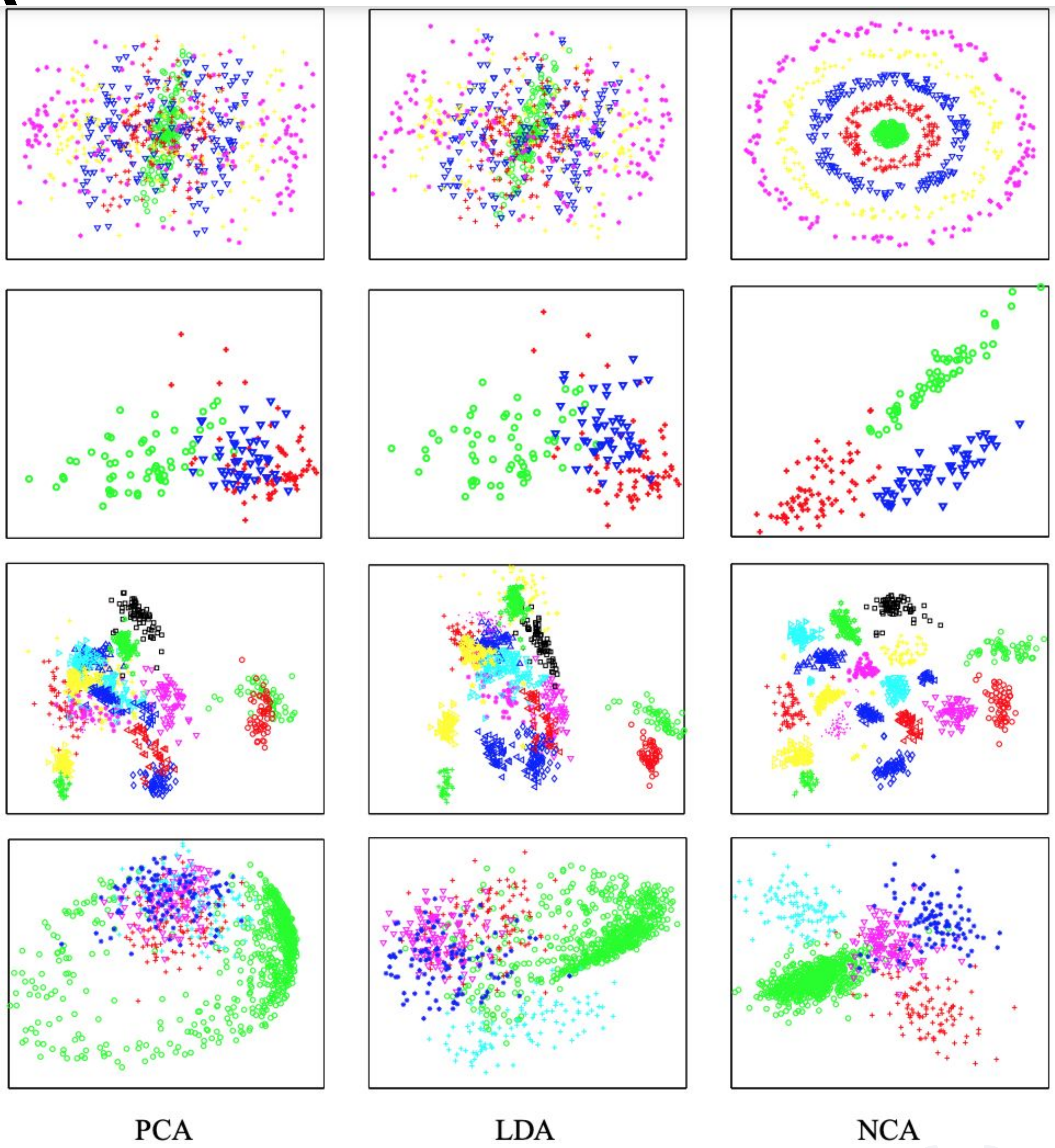


LDA:
maximizing the component
axes for class-separation



NCA:
maximizing component axes
for better KNN performance - classifier





Метод случайных проекции (RP)

При большом количество объектов $N > 10^6$ и признаков $d > 10^4$ преобразования для PCA, LDA, NCA становятся достаточно сложными.

Поэтому в задачах, связанных с обработкой больших объемов данных, используется метод понижение размерности на основе случайных проекций. (базируется на лемма Джонсона-Линденштрасса). Если точки векторного пространства проецируются на некоторое выбранное случайным образом подпространство достаточно большой размерности, то в среднем расстояния между точками сохраняются.

$$Y = XR$$

Для генерации матрицы случайных чисел $R = [d \times k]$ может использоваться стандартное нормальное распределение. (d и k количество признаков до и после преобразования)

ПРАКТИКА

Спасибо за внимание!