Feature Selection

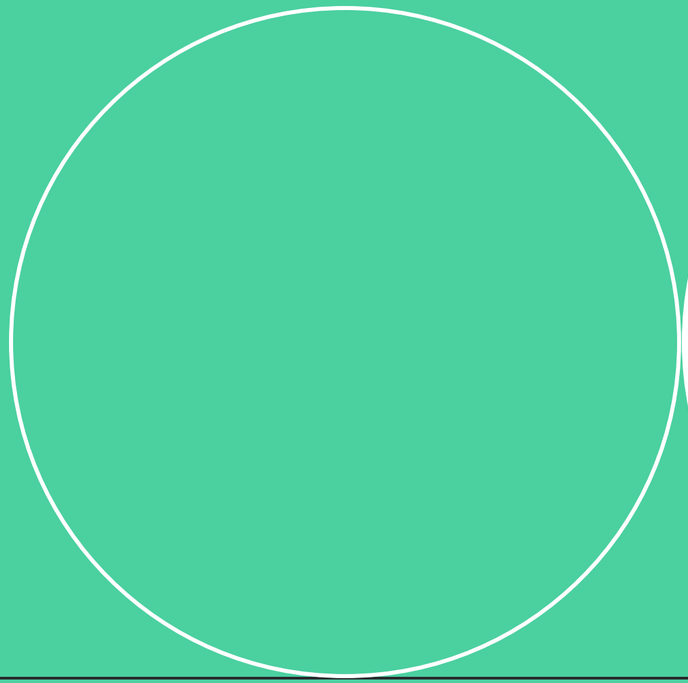
Юлия Пономарева

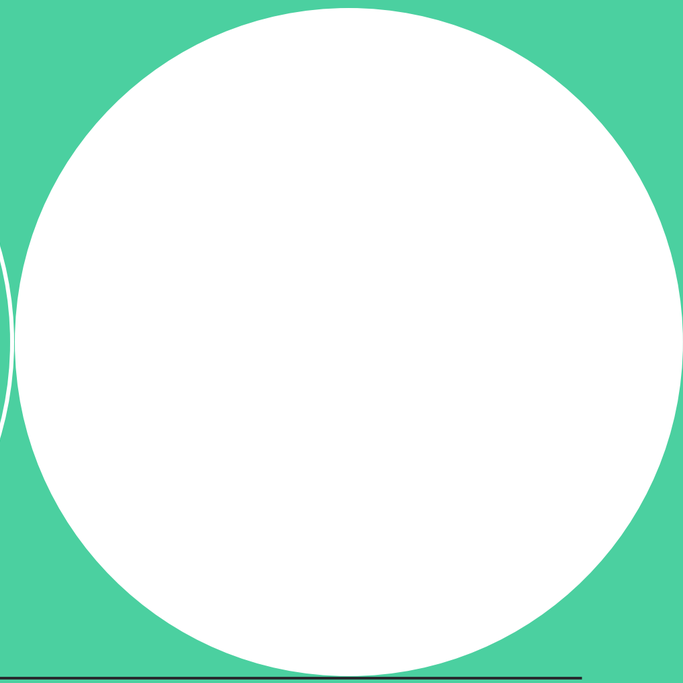
О спикере:

- Data Scientist
- автор YouTube канала «machine learrrning»
- работала в ITMO и Napoleon IT

Содержание

Введение





нетология

Введение. Зачем всё это?

Проклятье размерности - проблема, связанная с экспоненциальным возрастанием количества данных необходимых для обучения с увеличением размерности пространства

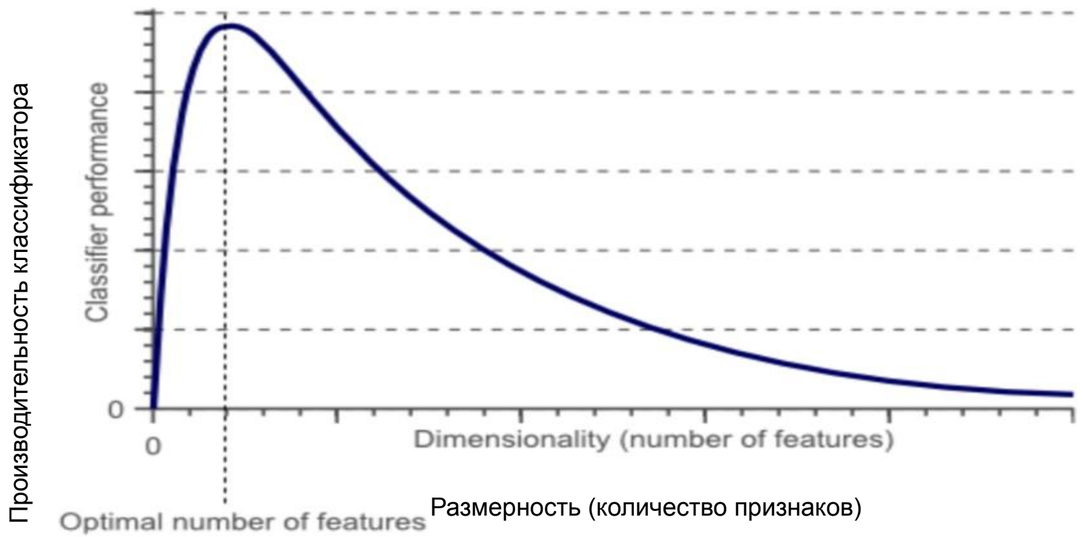
Одно измерение - 5 точек

Два измерение - 25 точек

Три измерения - 125 точек

Введение. Зачем всё это?

С увеличением размерности пространства, некоторые алгоритмы начинают хуже работать



Оптимальное количество признаков

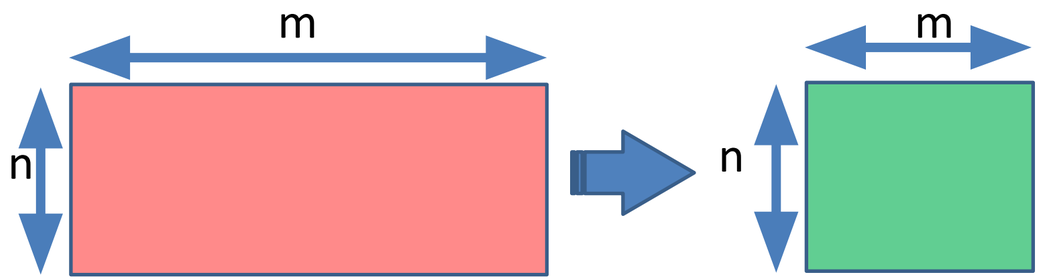
Методы отбора признаков

Позволяют получить:

- упрощение моделей для повышения возможности интерпретации
- сокращение времени тренировки
- уменьшения влияния проклятия размерности
- улучшение обобщения путём сокращения переобучения
- фильтрацию шумных признаков

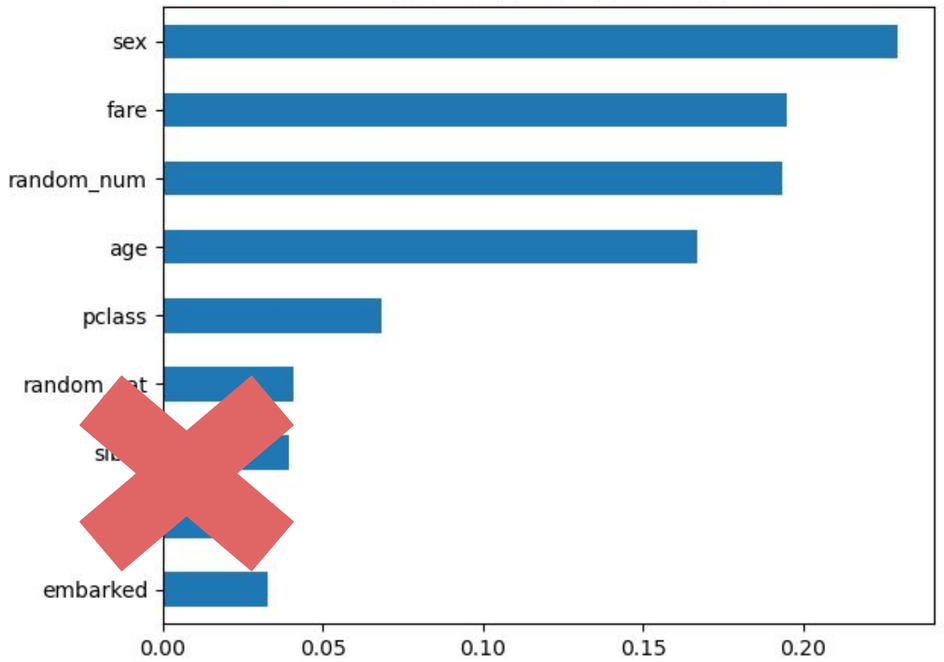
Как можно избежать проклятия размерности?

Преобразовать признаки

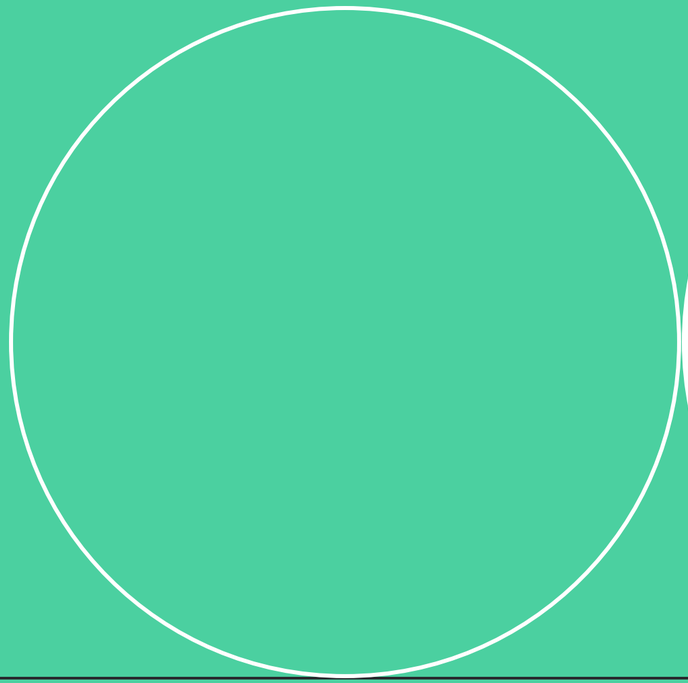


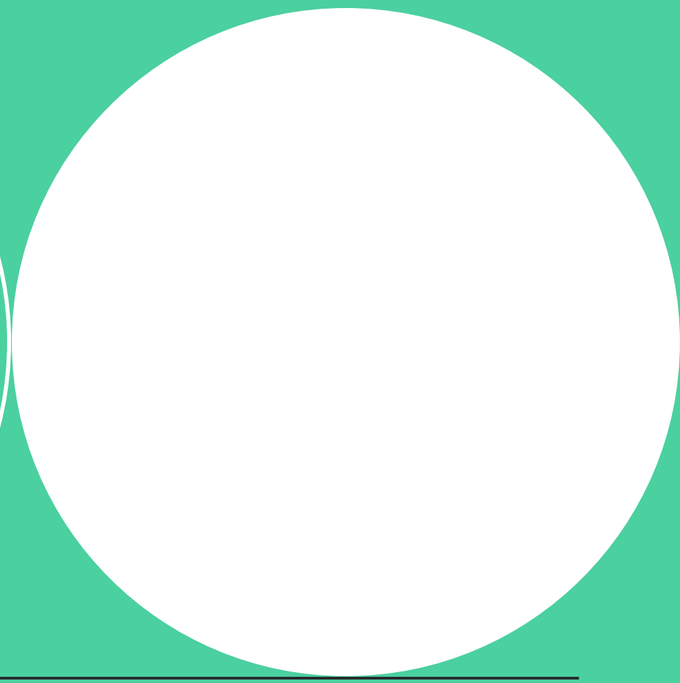
n – количество примеров
m – количество признаков

Отобрать признаки



Методы отбора признаков





нетология

Методы отбора признаков

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

Фильтры (одномерный отбор)
основаны на некоторых показателях, которые не зависит от метода классификации
(коэффициент корреляции, взаимная информация, F-тест, Хи-квадрат)

Обертки
опираются на информацию о метрике качества, полученную от 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, перемешивания и др.)

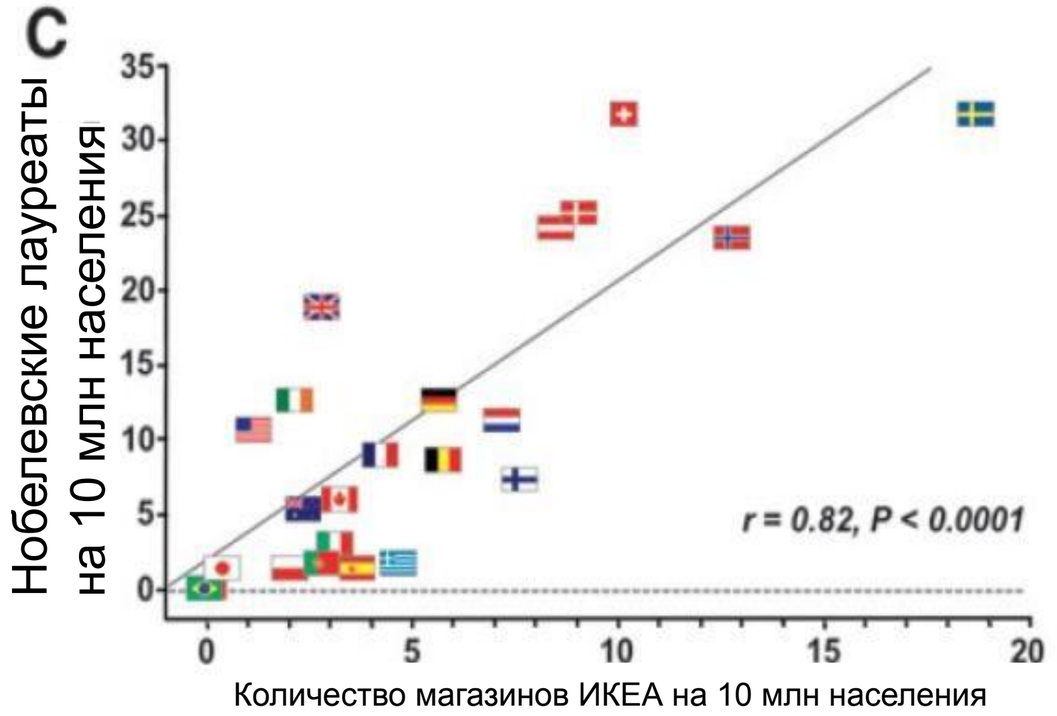
Фильтры
Одномерный отбор

Корреляция

Корреля́ция — статистическая взаимосвязь двух величин. При этом изменения значений одной величин сопутствуют систематическому изменению значений другой величины

Коэффициент корреляции Пирсона

$$R_{k_i,p} = \frac{\sum_{i=1}^{n}(k_i - \hat{k}) \cdot (p_i - \hat{p})}{\sqrt{\sum_{i=1}^{n}(k_i - \hat{k})^2 \cdot \sum_{i=1}^{n}(p_i - \hat{p})^2}}$$



ForexAW.com

Взаимная информация (Mutual Information)

MI(переменная x1 ; таргет) = Entropy(переменная x1) Entropy(переменная x1 |таргет)

$$Entropy = -\sum p(X) \log p(X)$$

here p(x) is a fraction of examples in a given class

Зависимость между полом и пользованием страховыми услугами

Пол	Пользуетесь ли Вы услугами страхования жизни?	
	Да	Нет
Мужской	39%	54%
Женский	61%	46%
Итого по столбцу	100%	100%

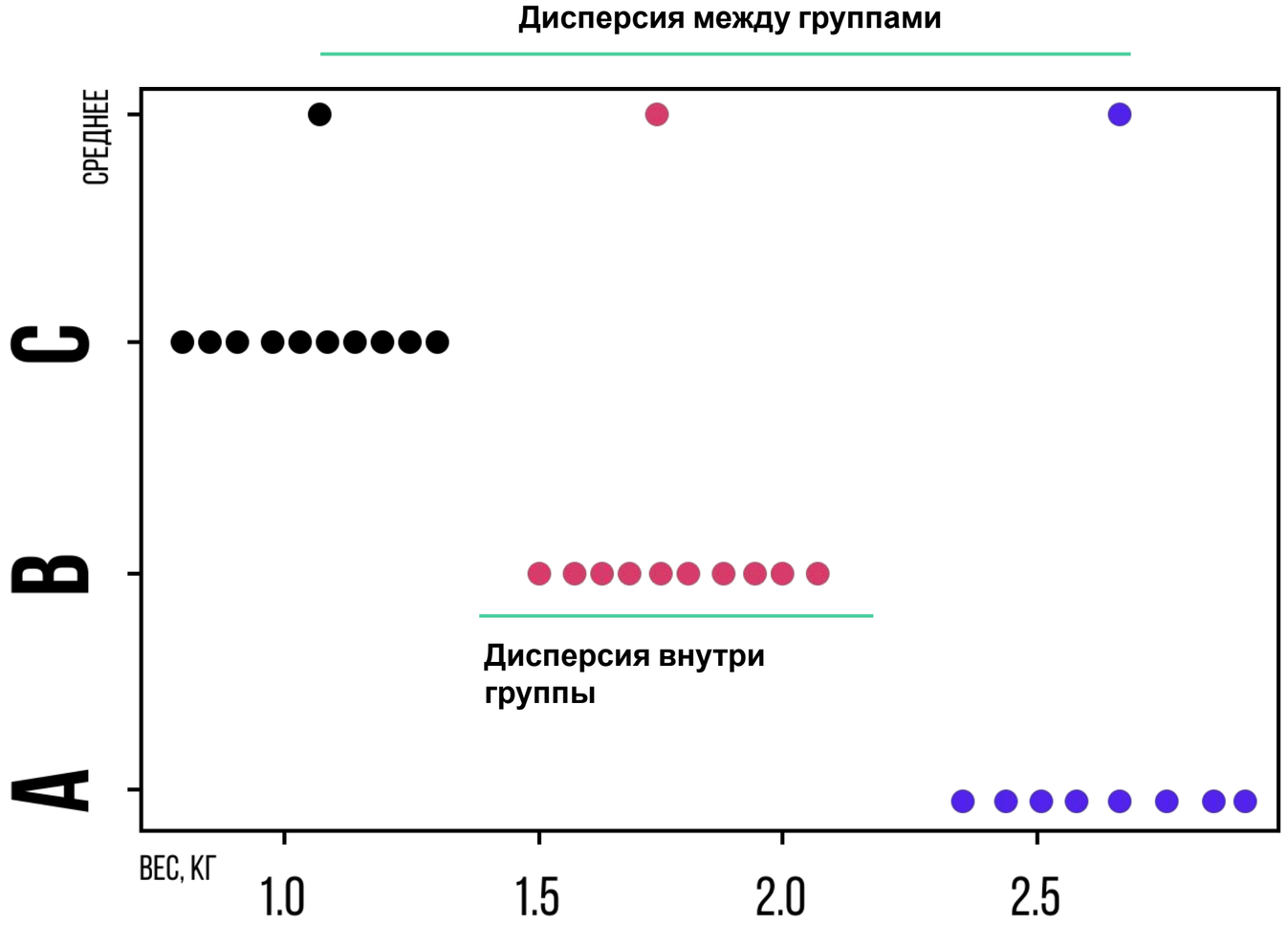
Чем выше значение MI, тем сильнее связь между этой переменной и таргетом, что говорит о том, что мы должны поместить эту переменную в набор данных для обучения

F-тест (критерий Фишера)

Очевидно, что при равенстве дисперсий величина критерия будет равна единице

$$F = \frac{\text{Дисперсия между группами}}{\text{Дисперсия внутри группы}}$$

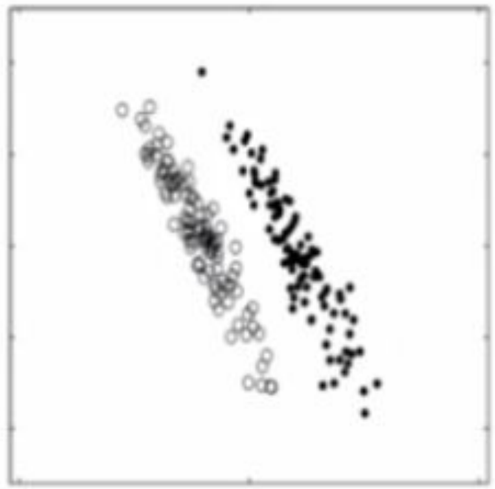
Чем больше F, тем проще различить выборки

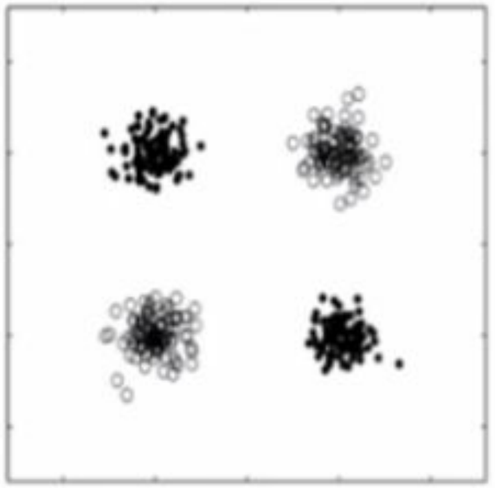


*https://ru.wikipedia.org/wiki/F-%D1%82%D0%B5%D1%81%D1%82

Одномерный отбор

У одномерного отбора признаков есть проблема - они не учитывают взаимосвязь признаков, зависимость целевой переменной от сложной комбинации признаков.

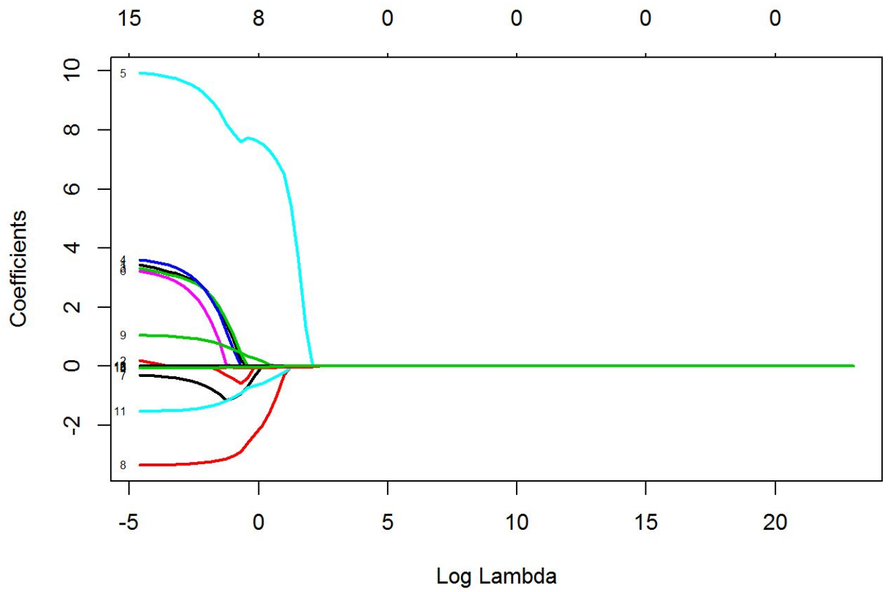




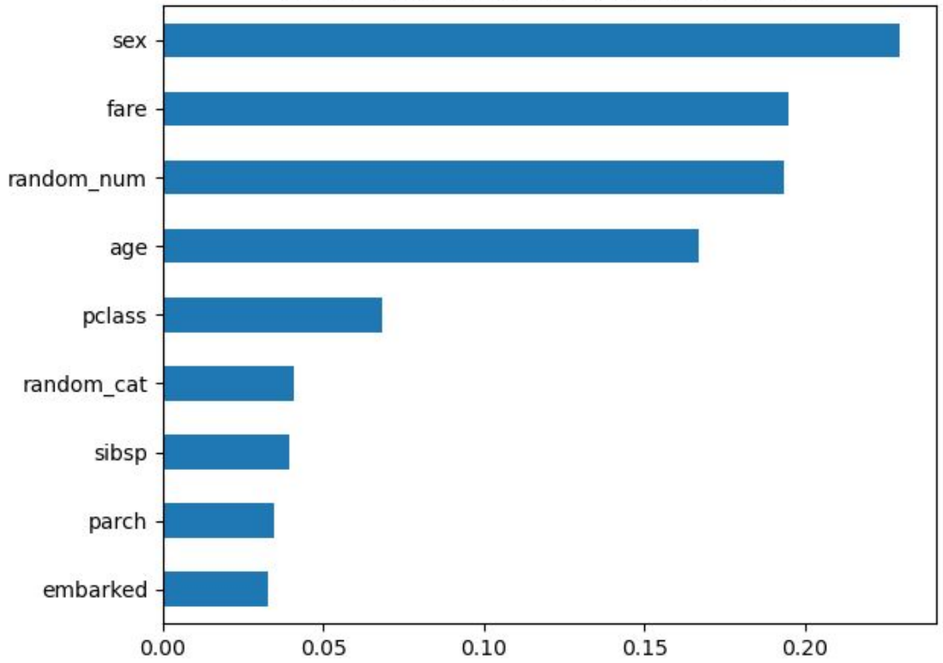
Встроенные в алгоритмы

Методы встроенные в алгоритмы

L1 регуляризация

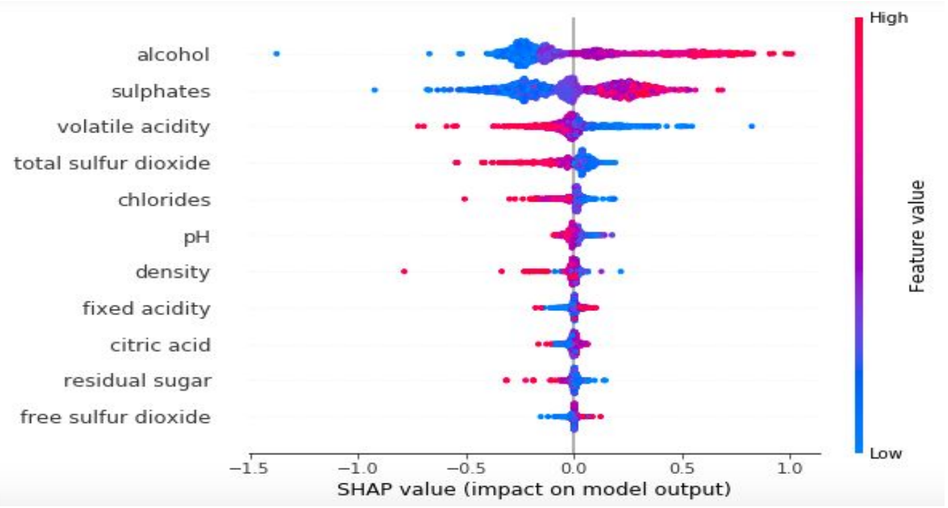


Feature Importance



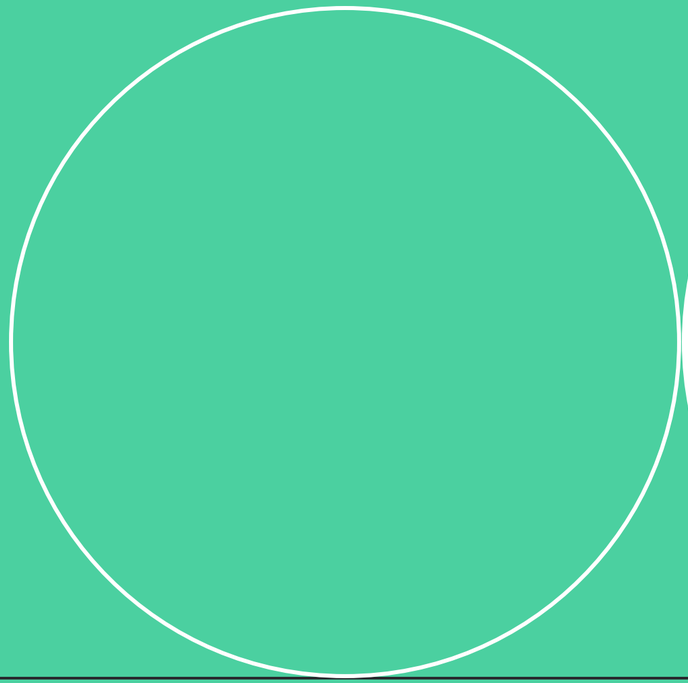
Методы встроенные в алгоритмы (отдельная библиотека - SHAP*)

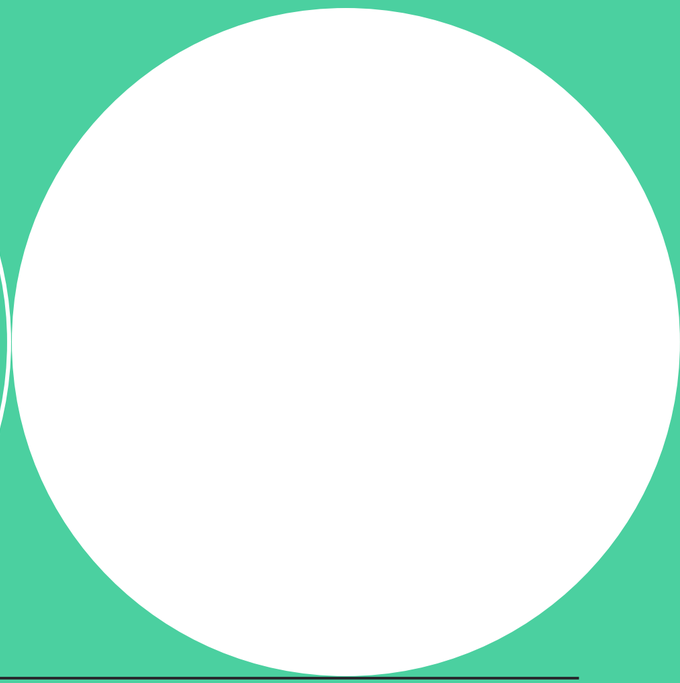
SHAP – значения показывают, насколько данный конкретный признак изменил наше предсказание (по сравнению с тем, как мы сделали бы это предсказание при некотором базовом значении этого признака)
Метод подходит для большинства моделей МЛ



* Лучше изучить все способы применения этого пакета
https://shap.readthedocs.io/en/latest/index.html

Преобразование признаков





нетология

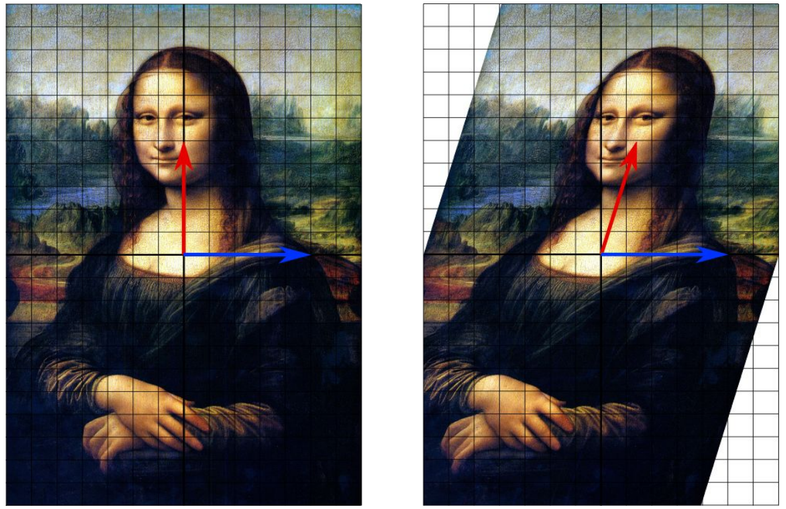
Преобразование признаков

Преобразование признаков может проводиться с помощью:

- PCA - Principal Component analysis (Метода главных компонент)
- LDA - Linear discriminant analysis (Линейного дискриминантного анализа)
- NCA - Neighbourhood components analysis (Анализа компонентов соседств)

Собственные векторы

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



Собственный вектор - вектор, который при умножении на матрицу дает точно такой же вектор, но измененный в масштабе

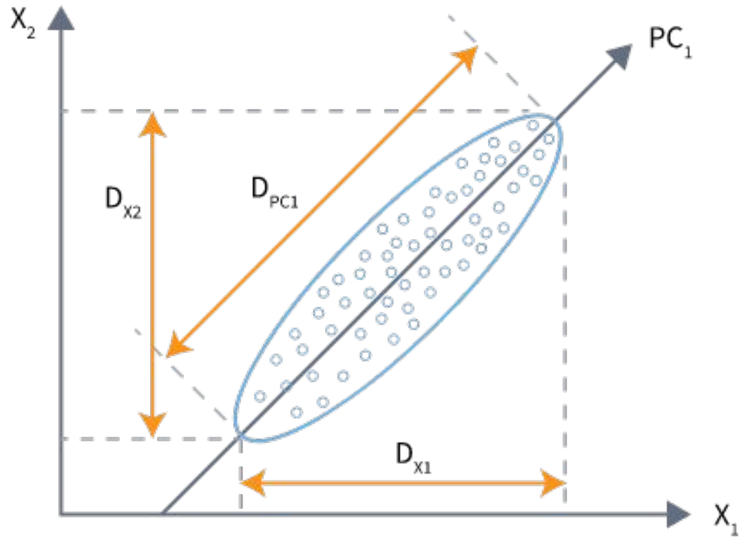
PCA

PCA - Principal component analysis

Позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры

- Первая ось новой системы координат строится таким образом, чтобы дисперсия данных вдоль неё была бы максимальна

Первая ось называется первой главной компонентой

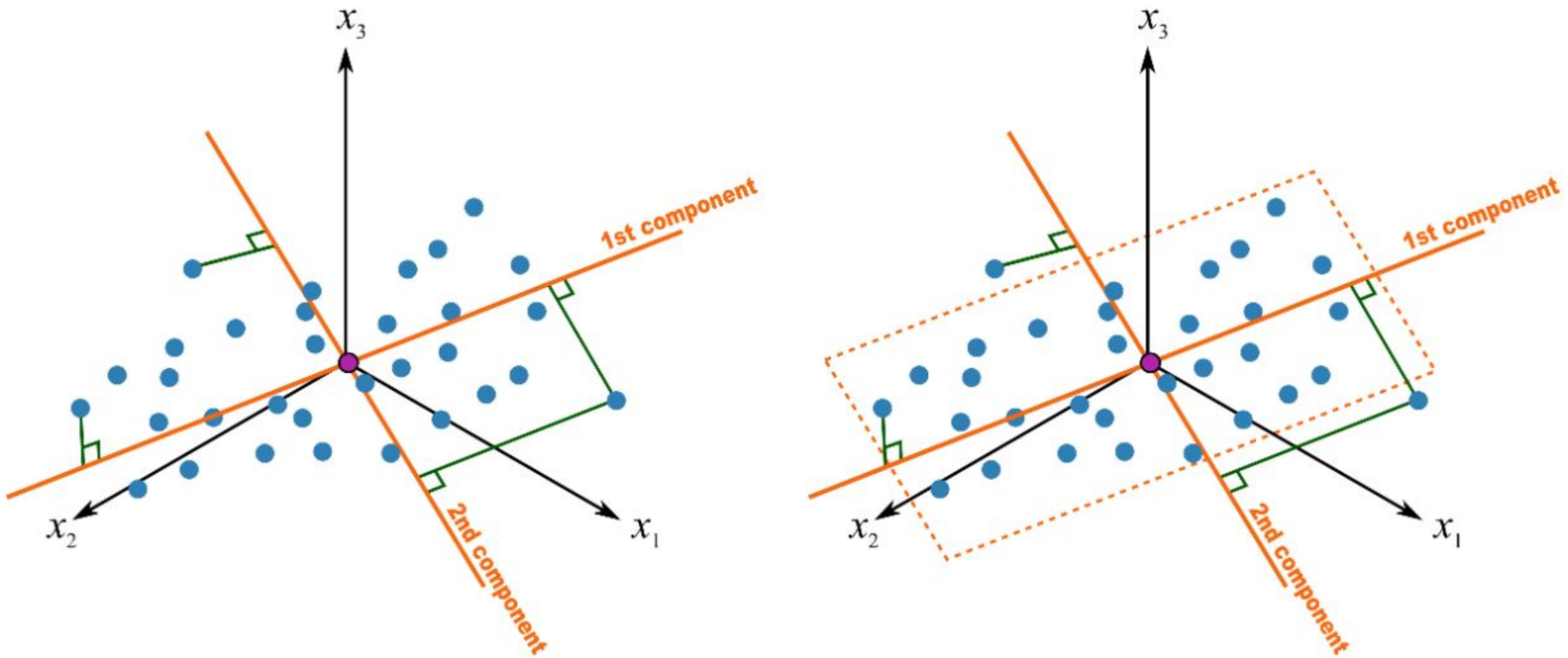


PCA - Principal component analysis

Позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры

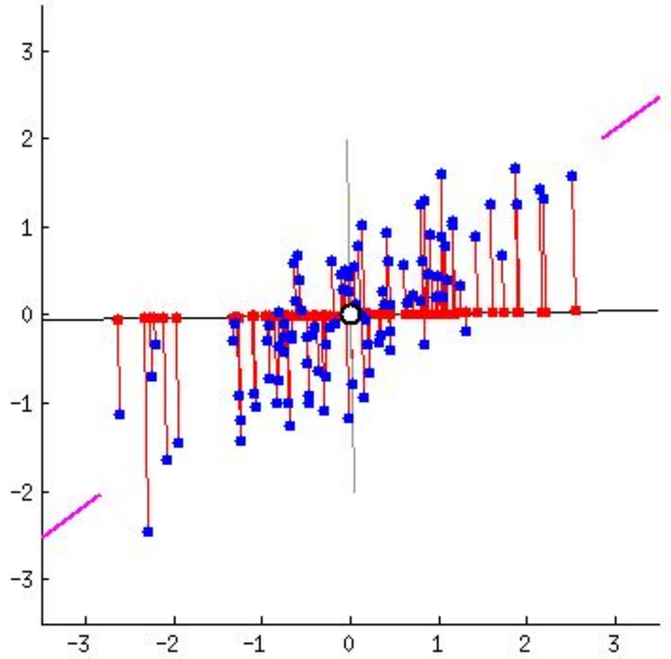
- Первая ось новой системы координат строится таким образом, чтобы дисперсия данных вдоль неё была бы максимальна
- Вторая ось строится перпендикулярно первой так, чтобы дисперсия данных вдоль неё, была бы максимальной их оставшихся возможных и т.д.

Первая ось называется первой главной компонентой, вторая — второй и т.д.



PCA - Principal component analysis

Позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры



PCA - Principal component analysis

Таким образом, для реализации метода главных компонент нужно :

- найти собственные значения матрицы $X^T X$;
- отобрать d максимальных;
- составить матрицу W^T, столбцы которой будут являться собственными векторами, соответствующими отобранным собственным значениям, расположенным в порядке убывания;
- получить новую матрицу "объекты-признаки", умножив исходную матрицу X на матрицу весов W :

$$Z = XW.$$

LDA

LDA - Linear discriminant analysis

Линейный дискриминантный анализ - метод уменьшения размерности, используемый в качестве этапа предварительной обработки в приложениях машинного обучения и классификации

- Первый этап - вычислить разделимость между разными классами (расстоение 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 этап - вычислить расстояние между средним значением и выборкой каждого класса, которое называется внутриклассовой дисперсией.

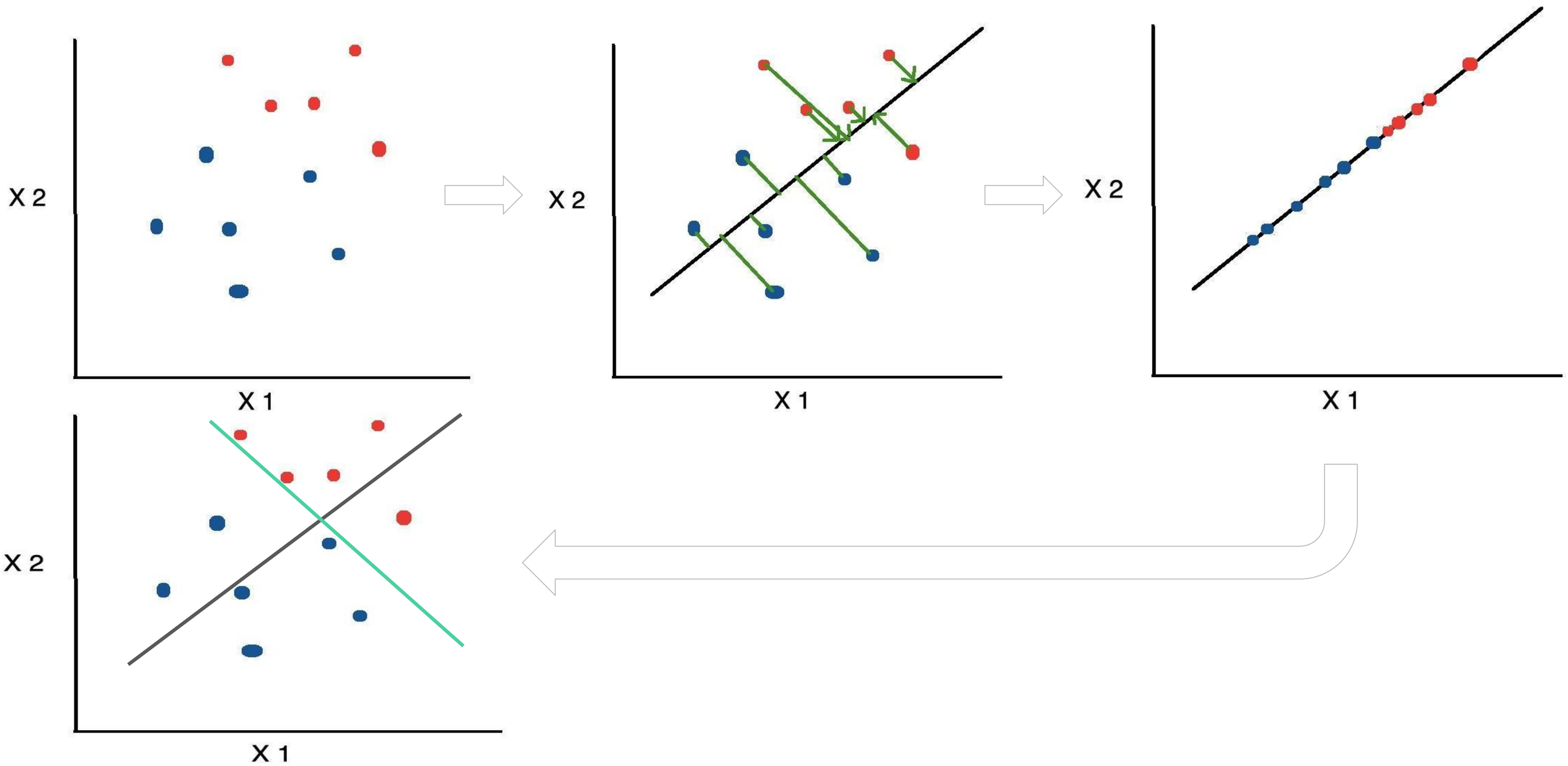
$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (\overline{x_{i,j}} - \overline{x_i})(\overline{x_{i,j}} - \overline{x_i})^T$$

- Третий этап - построить пространство более низкой размерности, которое максимизирует дисперсию между классами и минимизирует дисперсию внутри класса.

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

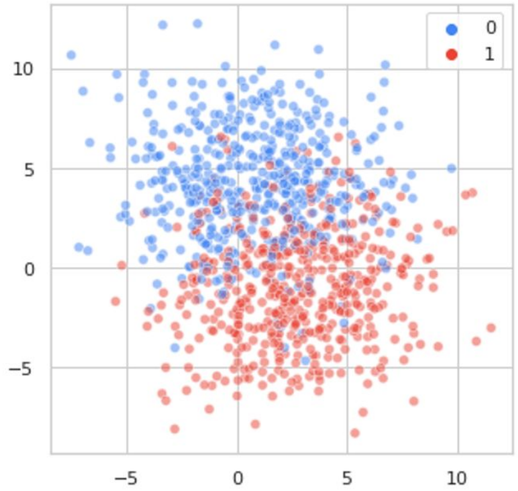
https://cmi.to/методы/линейный-дискриминантный-анализ/

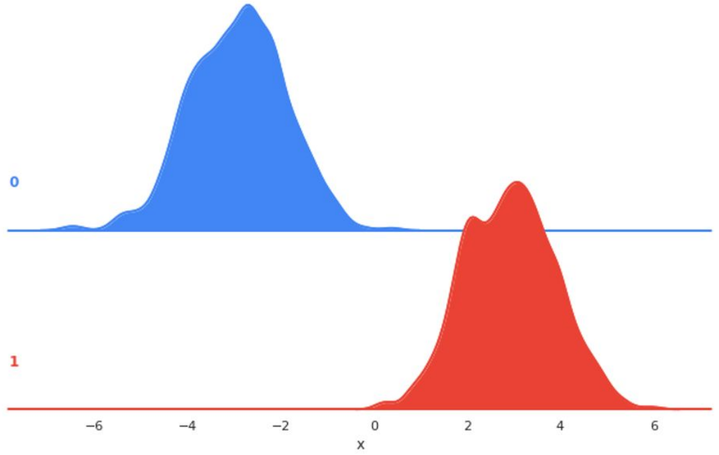
Пример для двух классов



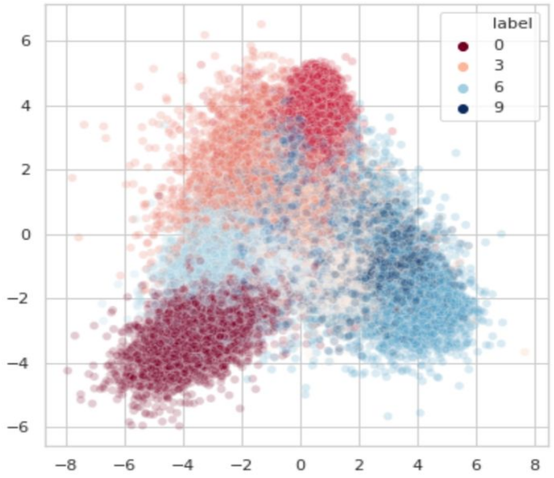
LDA - примеры

Отображение распределение
в 1- мерное пространство

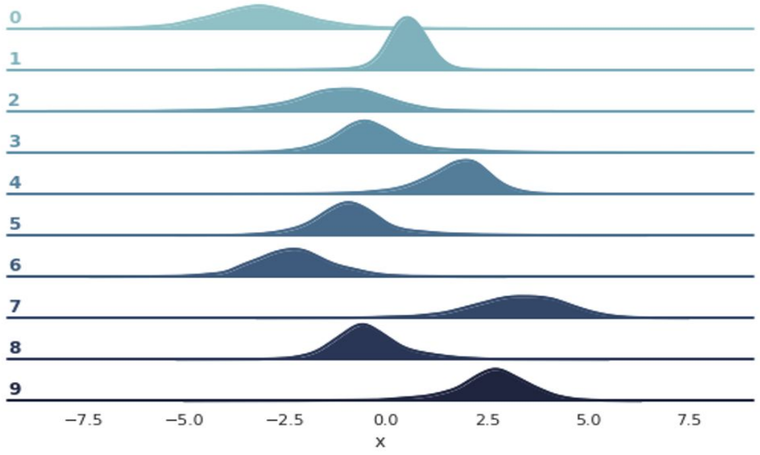




Двумерное представление



Одномерное представление



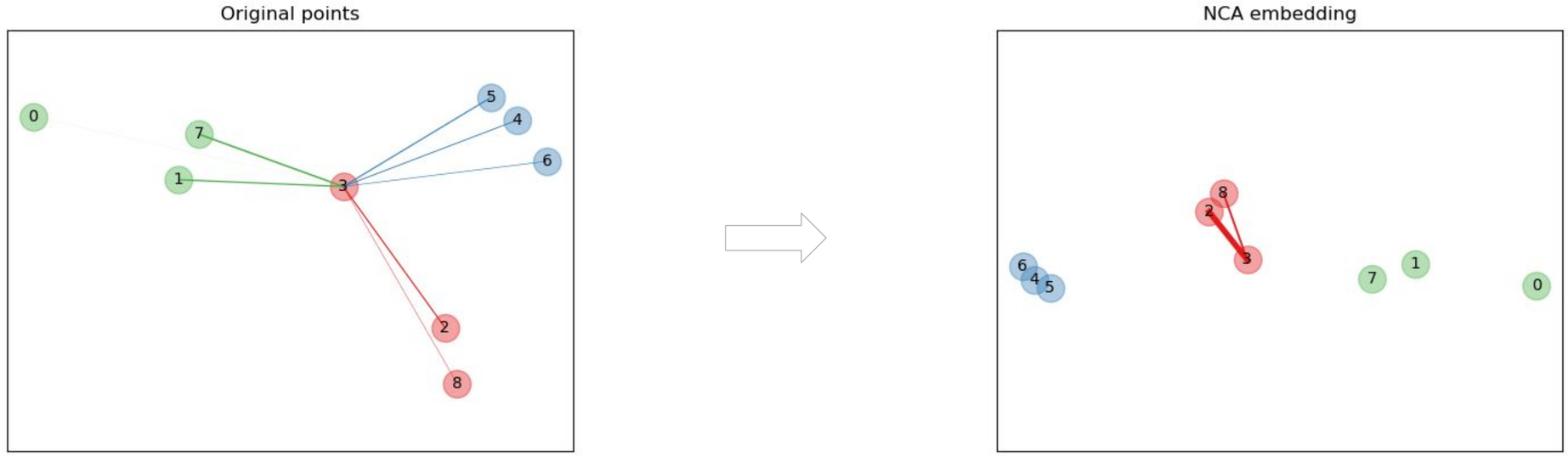
Отображение картинок
MNIST
в 2- и 1- мерное
пространство

NCA

NCA - Neighbourhood components analysis

Анализ компонентов соседств — это алгоритм, использующий метод, аналогичный методу K-ближайших соседей, для нахождения пространства, в котором окрестности точек с одинаковыми метками более плотные, чем точки с разными метками

Мы используем NCA, чтобы изучить эмбеддинги* и построить точки после преобразования. Затем мы берем вложение и находим ближайших соседей



*Термин «эмбеддинг» (от англ. embedding – вложение)

https://scikit-learn.org/stable/auto_examples/neighbors/plot_nca_illustration.html#sphx-glr-auto-examples-neighbors-plot-nca-illustration-py

NCA - Neighbourhood components analysis

Класс точки определяется взвешенным объединением
классов всех остальных точек

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

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

C - множество точек с таким же классом, как у объекта i
A - метрика расстояния

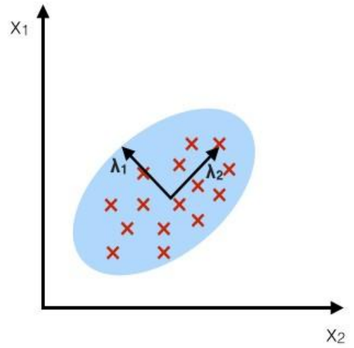
Приближение матрицы преобразования происходит градиентными итеративными методами

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

Сравнение LDA, PCA и NCA

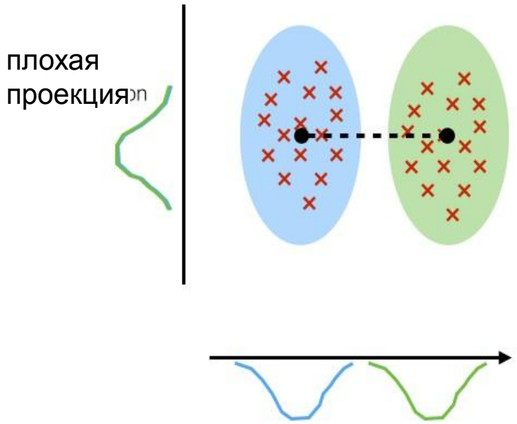
PCA:

Оси компонент максимизируют дисперсию



LDA:

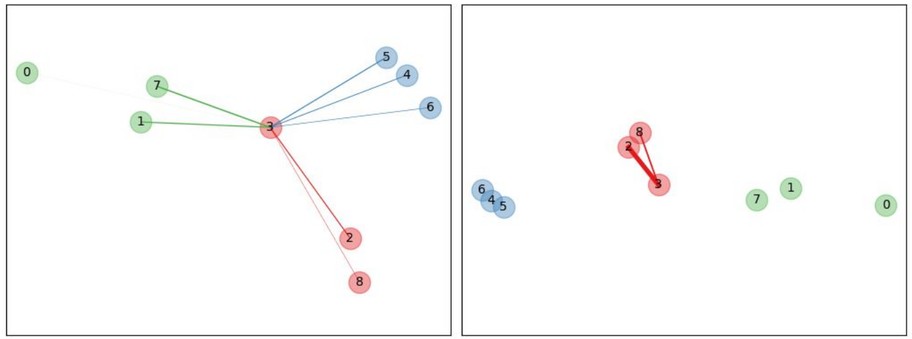
Оси компонент максимизируют разделение классов



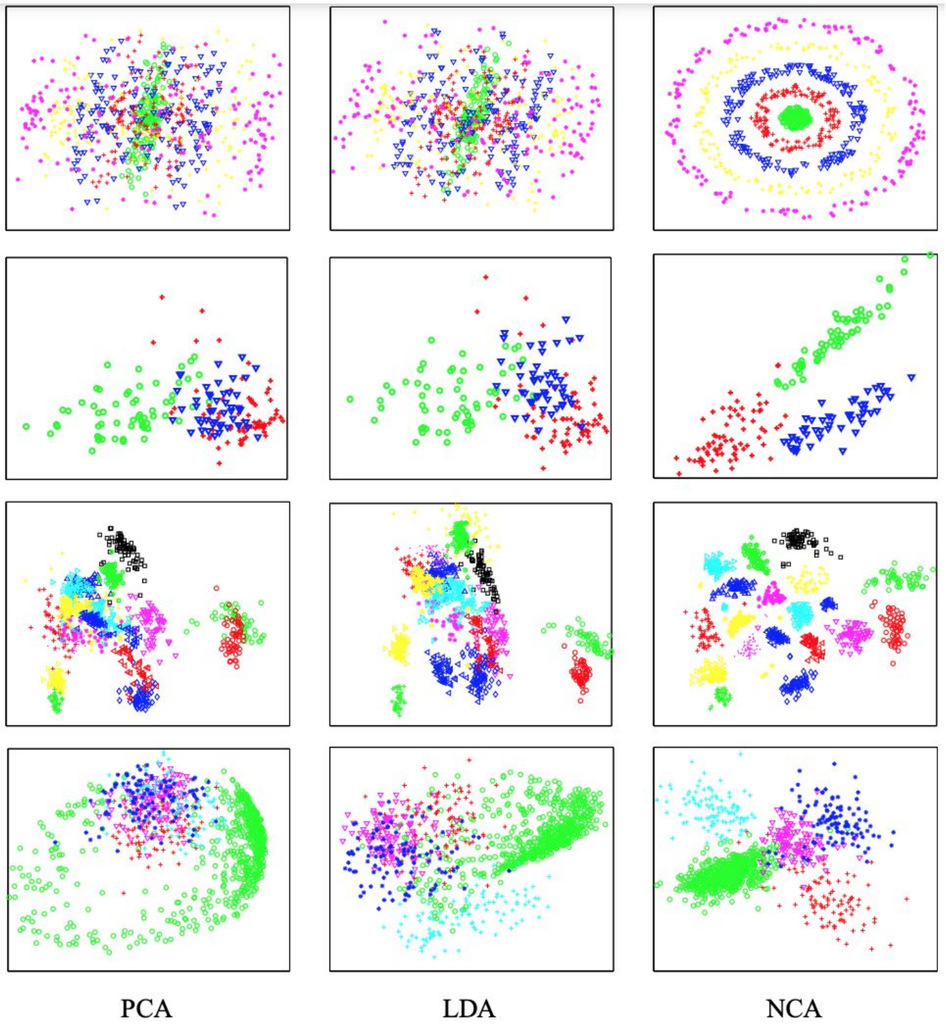
хорошая проекция: хорошо разделяет классы

NCA:

Находит более плотные группы точек с одинаковыми метками, а не разными

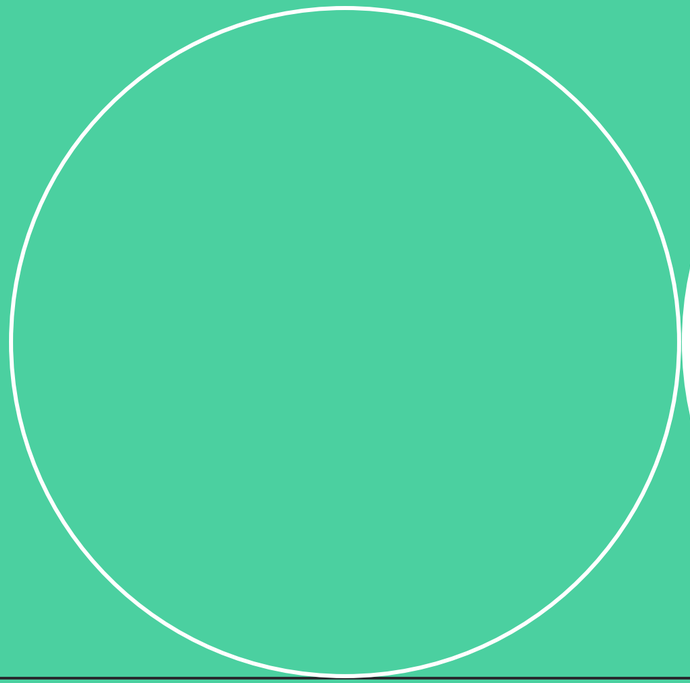


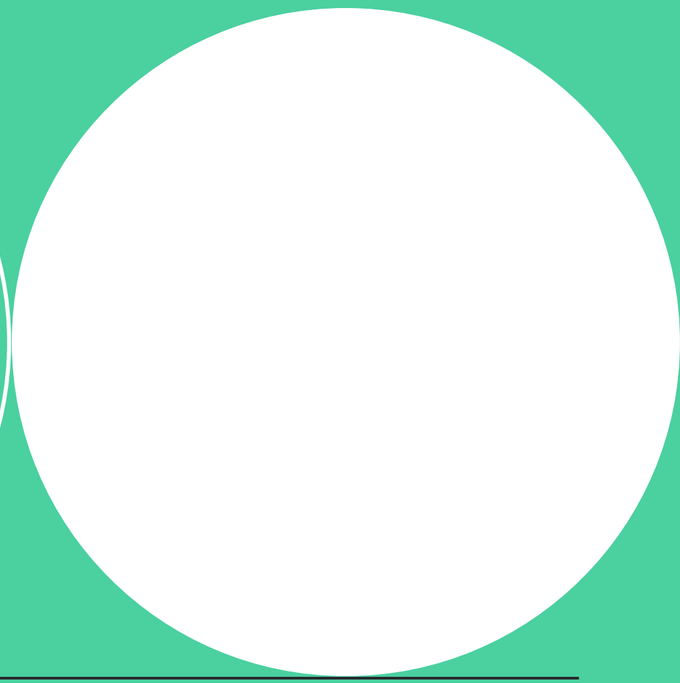
Сравнение LDA, PCA и NCA



Эксперименты проведенные на различных данных с использованием разных алгоритмов показывают, что в зависимости от датасета, алгоритмы работают с разной эффективностью

Практика





нетология

Итоги

Итоги

1. С увеличением размерности пространства некоторые алгоритмы начинают хуже работать и решить эту проблему помогают преобразование и отбор признаков
2. Методы отбора признаков позволяют упростить модели, сократить время тренировки, уменьшить влияние проклятия размерности, сократить переобучение, отфильтровать шумные признаки
3. Методы отбора признаков обычно делят на фильтры, обёртки, встроенные в алгоритмы
4. Преобразование признаков может проводиться с помощью метода главных компонент, линейного дискриминантного анализа, анализа компонентов соседств

Feature Selection