

Обучение без учителя: кластеризация.

Снижение размерности данных PCA.

Екатерина Кондратьева

# Кластерный анализ

# Кластеризация

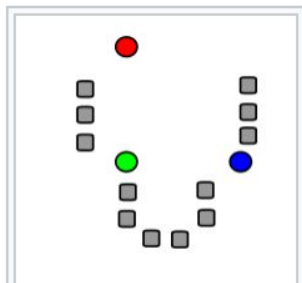
Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Реализации алгоритмов: [https://scikit-learn.org/0.18/auto\\_examples/cluster/plot\\_cluster\\_comparison.html](https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html),  
<https://scikit-learn.org/stable/modules/clustering.html#k-means>

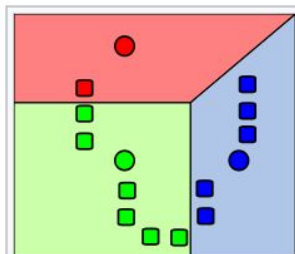
Лекция: <https://ru.coursera.org/lecture/unsupervised-learning/vybor-mietoda-klastierizatsii-RZSVo>

Unsupervised learning: <https://ru.coursera.org/learn/unsupervised-learning>

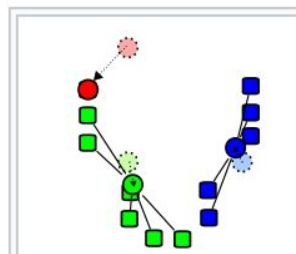
# Метод k- средних



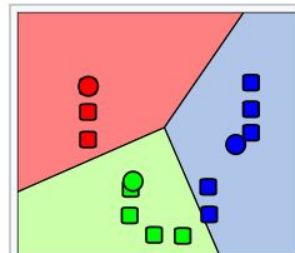
Исходные точки и  
случайно выбранные  
начальные точки.



Точки, отнесённые к  
начальным центрам.  
Разбиение на  
плоскости —  
**диаграмма Вороного**  
относительно  
начальных центров.



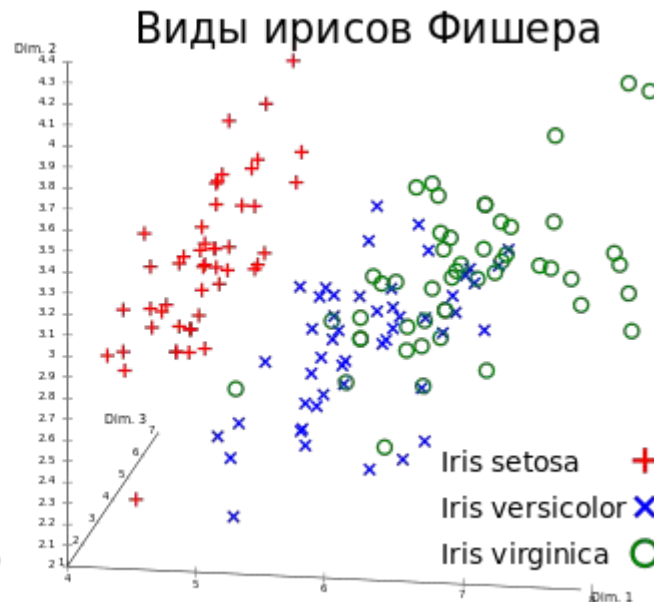
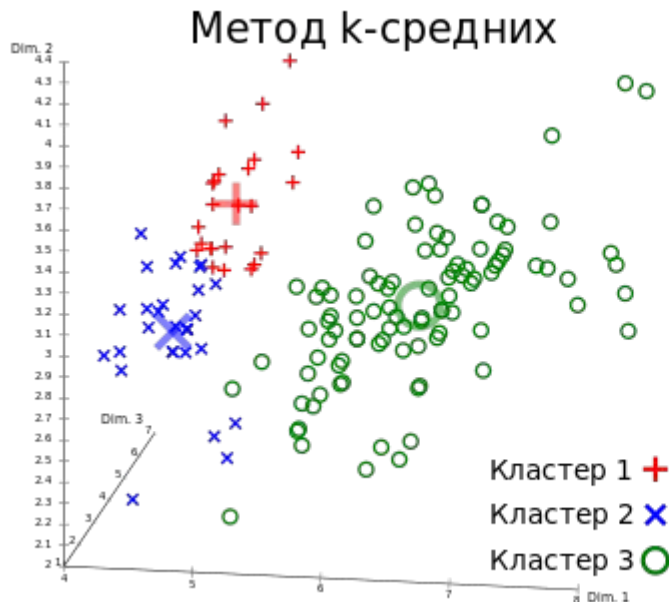
Вычисление новых  
центров кластеров  
(Ищется **центр масс**).



Предыдущие шаги,  
за исключением  
первого, повторяются,  
пока алгоритм не  
сойдётся.

# Минусы метода k-средних

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения  $V$ , а только одного из локальных минимумов.
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.



# Методы снижения размерности

Зачем нужно снижать размерность выборки?

# Методы снижения размерности

Как уменьшить размерность выборки?

# Методы снижения размерности

Как уменьшить размерность выборки?

- удалить неинформативные характеристики объектов (т.е. те, которые вносят наименьший вклад в формирование решающего правила)
- преобразовать имеющиеся характеристики новые, количество которых уменьшит размерность выборки, без потери информации.

**как это сделать?**



# Методы снижения размерности

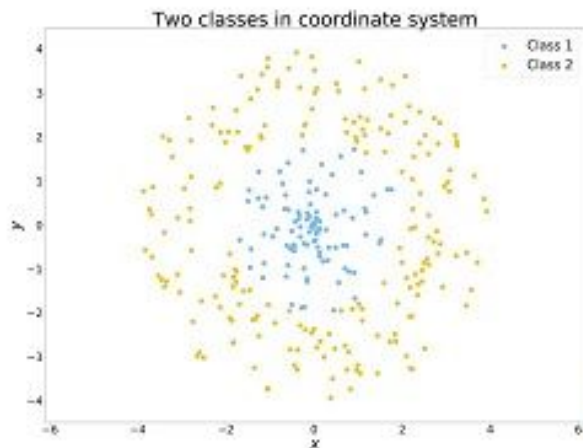
Как уменьшить размерность выборки?

- удалить неинформативные характеристики объектов (т.е. те, которые вносят наименьший вклад в формирование решающего правила)
- преобразовать имеющиеся характеристики новые, количество которых уменьшит размерность выборки, без потери информации.

**как это сделать?**

- **feature engineering, dimensionality reduction methods** (часто подразумевается manifold learning, или геометрические методы снижения размерности)

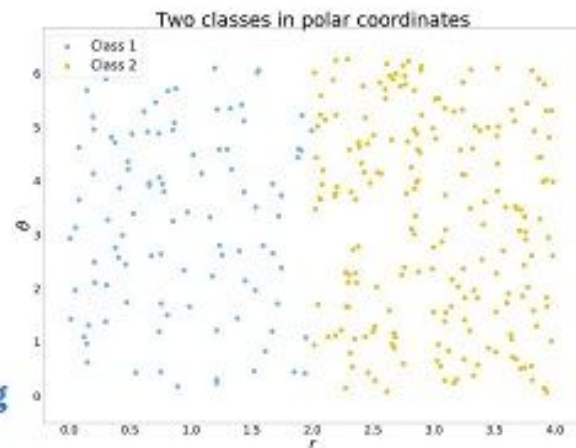
# Feature engineering



Tangled



Feature engineering



Transparent

feature engineering - генерация новых признаков, разделяющих данные

## Raw Data

```
0 : {  
  house_info : {  
    num_rooms: 6  
    num_bedrooms: 3  
    street_name: "Shorebird Way"  
    num_basement_rooms: -1  
    ...  
  }  
}
```

Raw data doesn't come to us as feature vectors.

Feature Engineering

## Feature Vector

```
[  
  6.0,  
  1.0,  
  0.0,  
  0.0,  
  0.0,  
  9.321,  
  -2.20,  
  1.01,  
  0.0,  
  ...,  
]
```

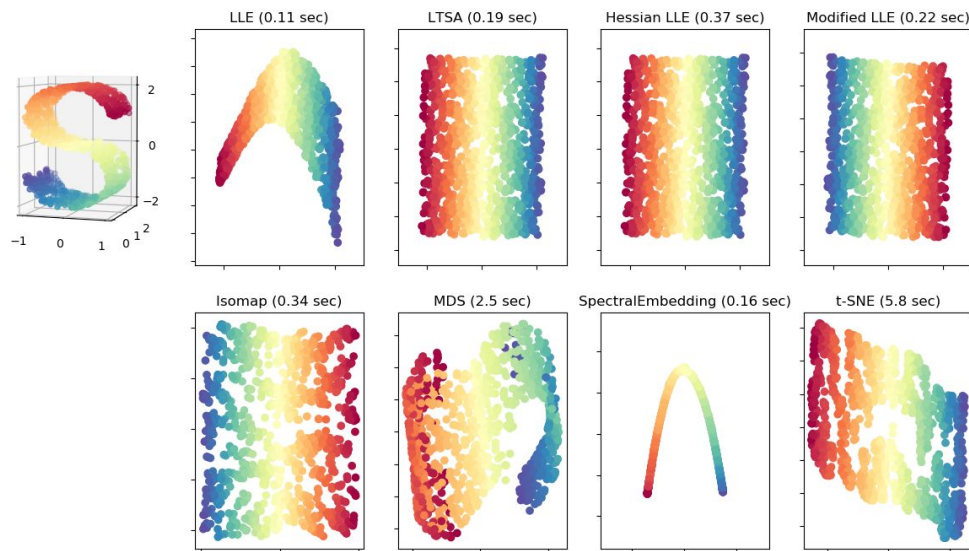
Process of creating features from raw data is **feature engineering**.

feature engineering - как обработка “сырых данных”

<https://developers.google.com/machine-learning/crash-course/representation/feature-engineering>

# Снижение размерности

Manifold Learning with 1000 points, 10 neighbors



<https://scikit-learn.org/stable/modules/manifold.html>

# Снижение размерности

- Линейные (PCA, SVD и др.)
- Нелинейные (Isomap, tSNE и др.)

Линейные <https://ru.coursera.org/lecture/unsupervised-learning/mietod-ghlavnykh-komponent-rieshieniie-e72bH>  
<https://ru.coursera.org/lecture/python-for-data-science/mietod-glavnykh-komponent-principal-component-analysis-X8bem>

Нелинейные  
<https://ru.coursera.org/lecture/vvedenie-mashinnoe-obuchenie/nielinieinyie-mietody-ponizhieniia-razmiernosti-QloeT>