Обучение без учителя: кластеризация. Выявление аномалий в данных (Anomaly detection).

Екатерина Кондратьева

Обучение без учителя (unsupervised learning):

Или анализ данных без разметки. Можно условно разделить на три больших направления:

1. кластерный анализ (кластеризация), обнаружение аномалий (anomaly detection);
2. методы снижения размерности (dimensionality reduction), оценка внутренней размерности выборки (component analysis), генерация признаков пониженной размерности (feature engineering);
3. *обучение с подкреплением (reinforcement learning) чаще deep learning, поэтому в этом курсе не рассматривается.

1. Кластерный анализ

Кластеризация

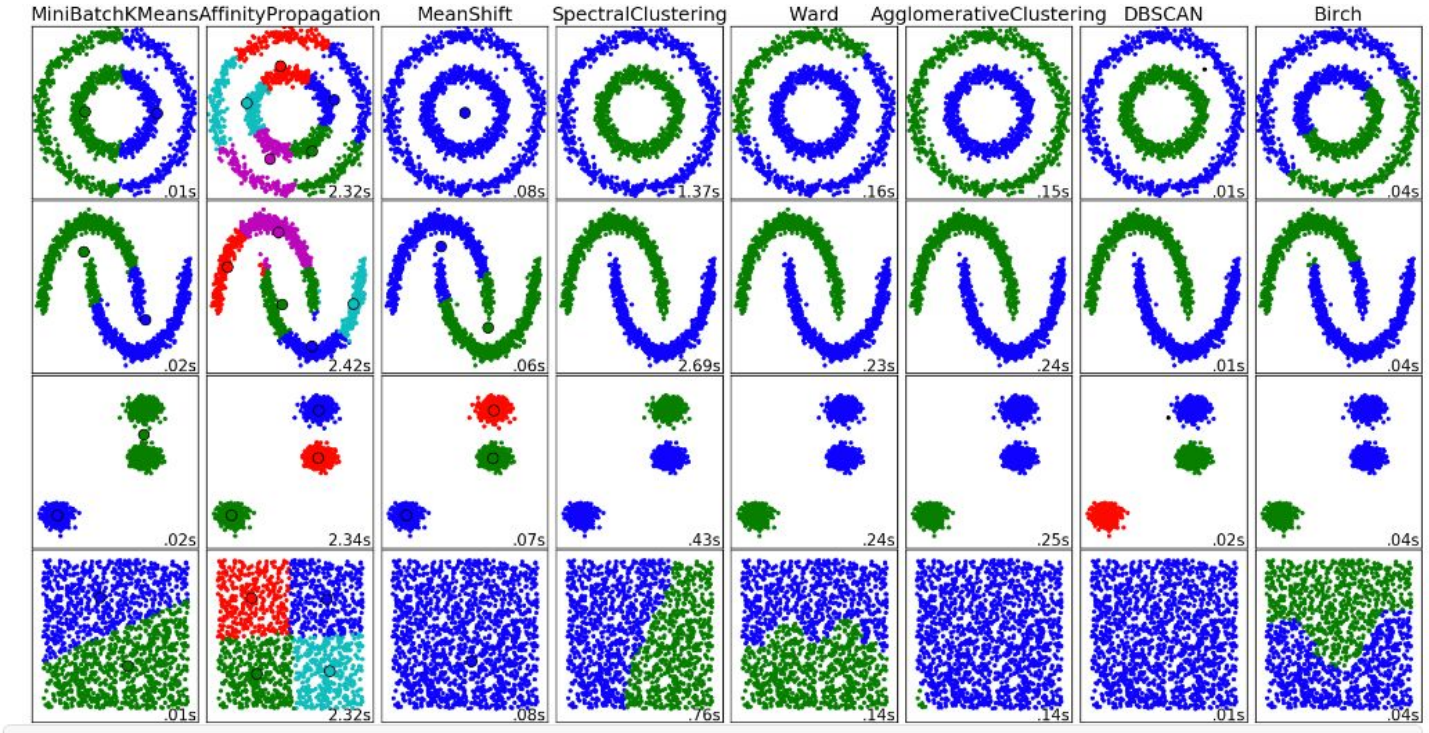
Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу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

Реализации алгоритмов: https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html, https://scikit-learn.org/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

Кластерный анализ



https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html

Суровая реальность:

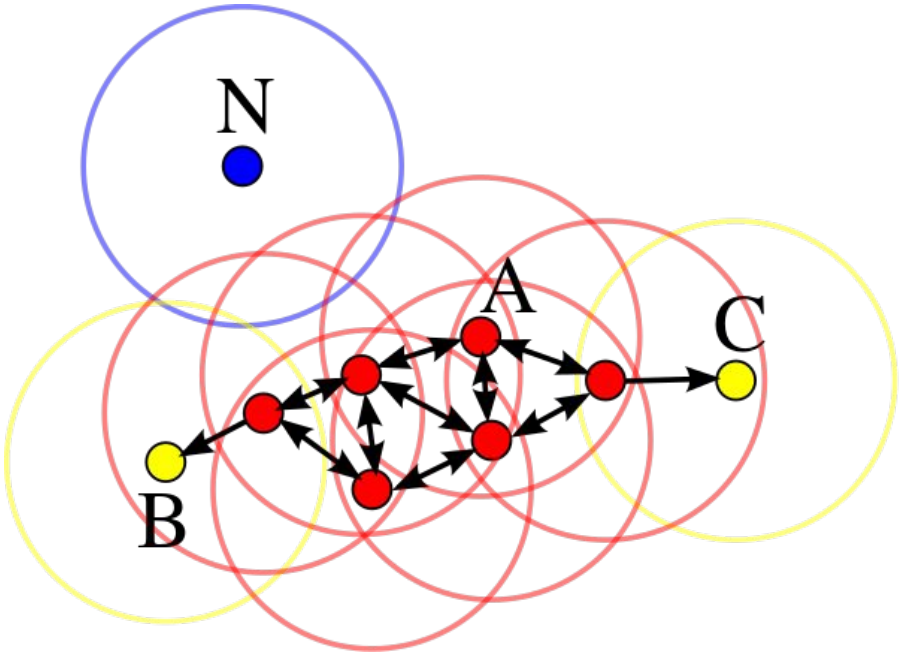
Универсального алгоритма кластеризации нет, но можно подбирать алгоритм под тип данных.

Кластеризация часто подразумевает предположение о количестве классов.

DBSCAN

Основанная на плотности пространственная кластеризация для приложений с шумами (англ. Density-based spatial clustering of applications with noise, DBSCAN).

DBSCAN требует задания двух параметров: радиуса окружности *epsilon* и минимального числа точек, которые должны образовывать плотную область

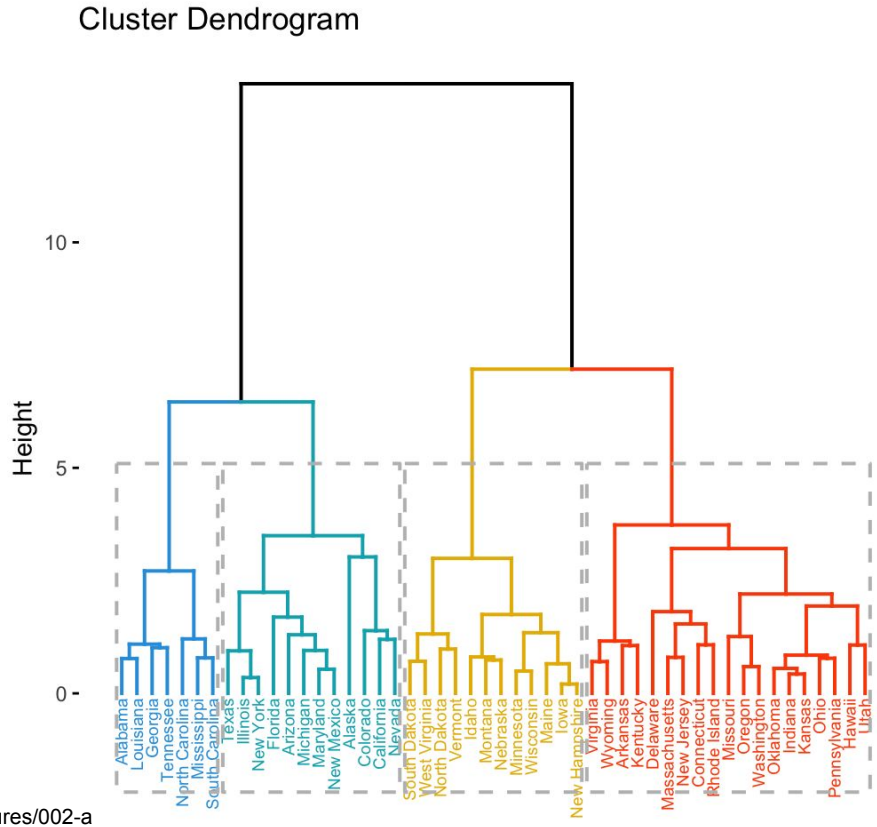


https://upload.wikimedia.org/wikipedia/commons/thumb/a/af/DBSCAN-Illustration.svg/600px-DBSCAN-Illustration.svg.png

Agglomerative clustering

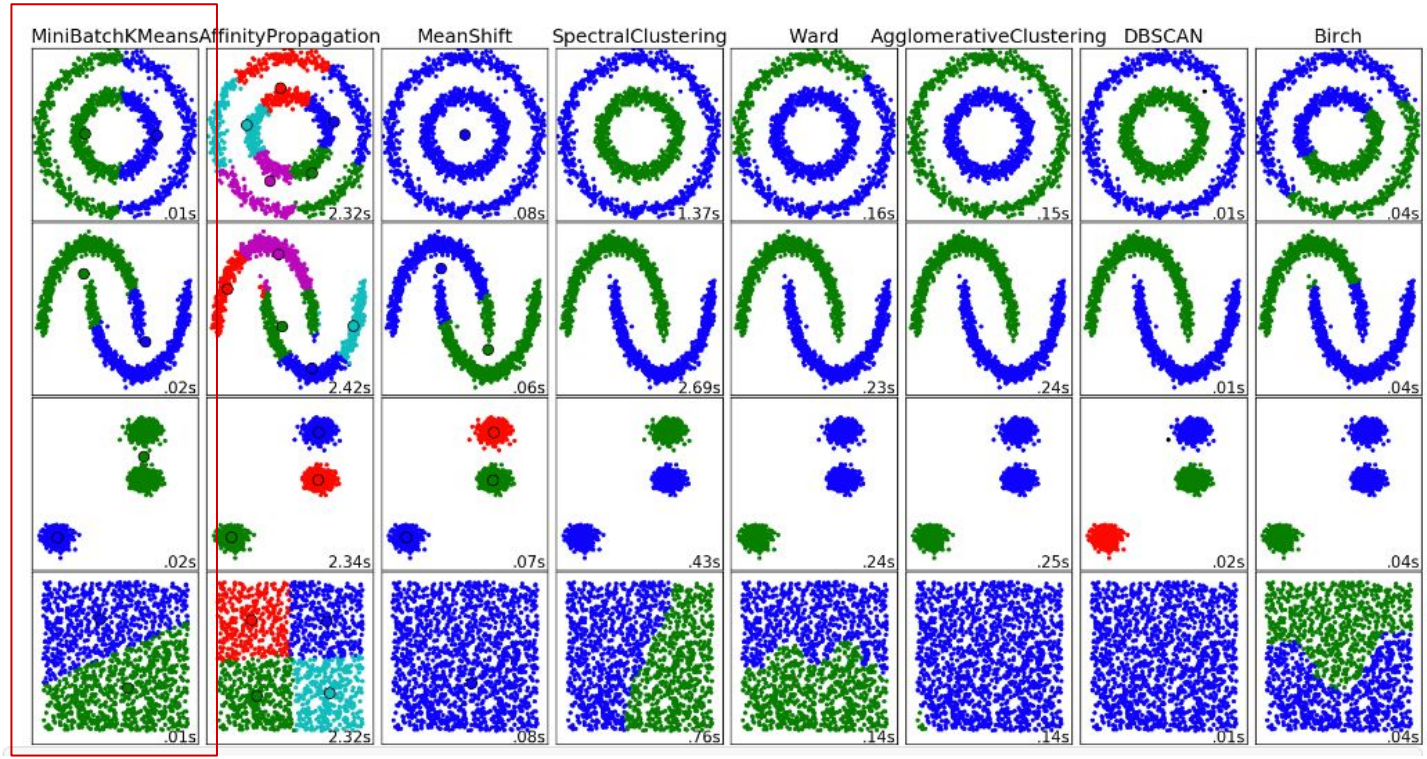
Итеративно соединяет пары классов (изначальной каждый элемент это отдельный класс) в соответствии с расстоянием на выбранной метрикой.

```
sklearn.cluster.AgglomerativeCluster
ing(n_clusters=2,
affinity='euclidean',
memory=None,...)
```



https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/003-hierarchical-clustering-in-r/figures/002-agglomerative-clustering-cutree-cut-dendrogram-1.png

Кластерный анализ

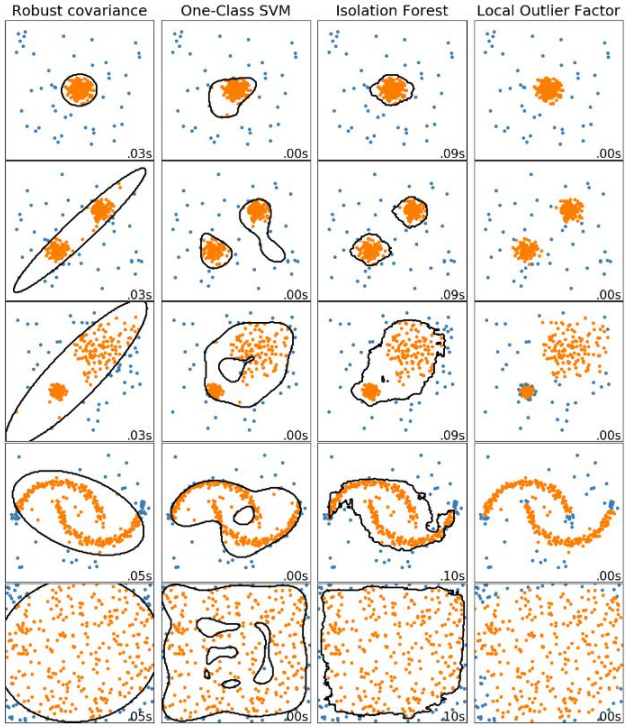


https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html

Обнаружение аномалий

часто предполагается предположение о количестве “примеси” в выборке

Обнаружение аномалий (anomaly detection)



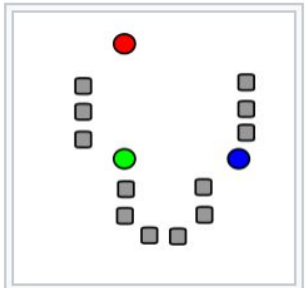
https://scikit-learn.org/stable/modules/outlier_detection.html

Аналогия методов классификации (регрессии)

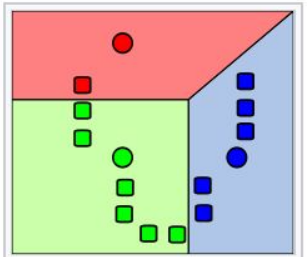
Знакомые нам методы машинного обучения для классификации (регрессии) имеют аналоги (схожие с ними методы) для кластеризации:

- Random Forest Classifier - Isolation Forest -* Agglomerative clustering
- KNN Classifier - KMeans - Local Outlier Factor
- SVC - One-class SVM

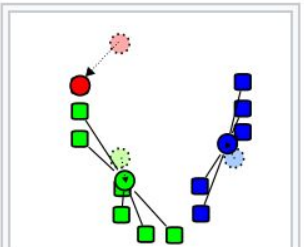
Пример: Метод k- средних



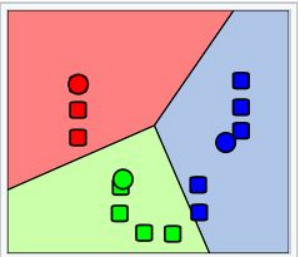
Исходные точки и случайно выбранные начальные точки.



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



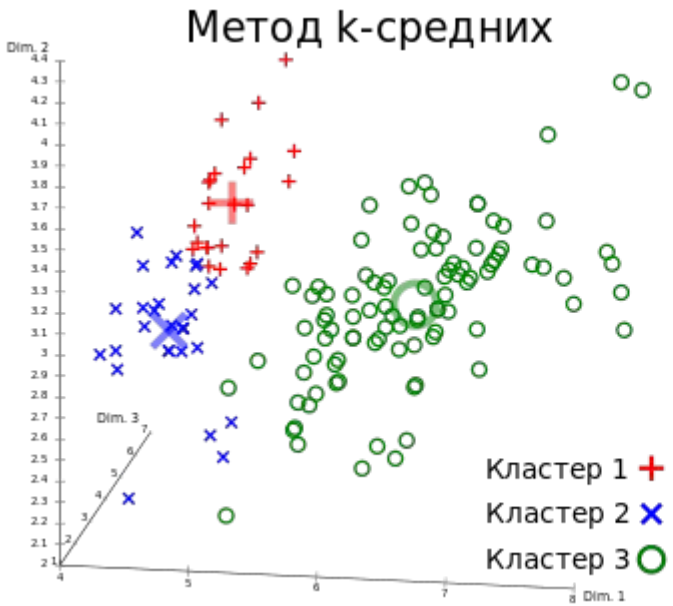
Вычисление новых центров кластеров (Ищется центр масс).

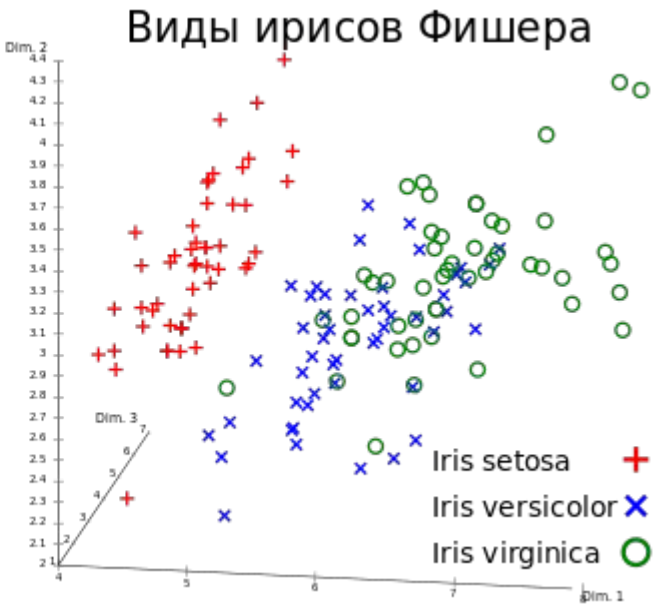


Предыдущие шаги, за исключением первого, повторяются, пока алгоритм не сойдётся.

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.





One-class-SVM. SVM

Вспомним функционал SVM:

$$\min_{w,\, b,\, \xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$

$$\text{subject to:}$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \qquad \text{for all } i = 1, \ldots, n$$

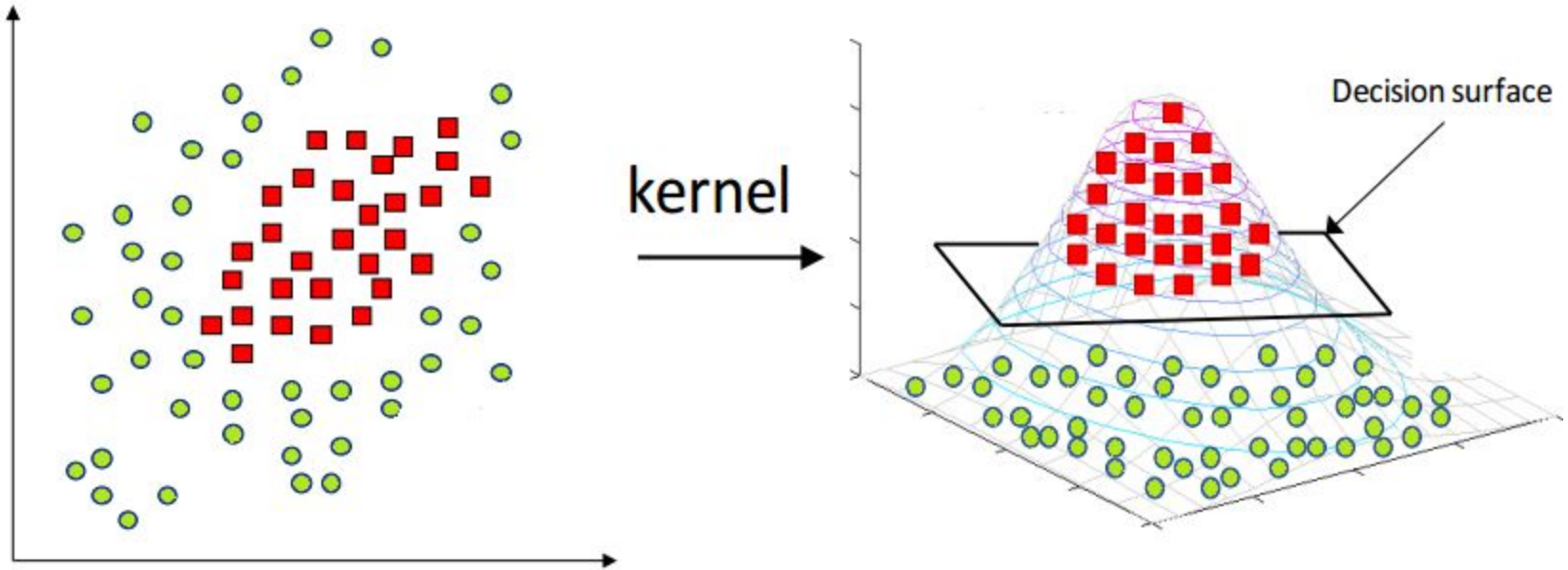
$$\xi_i \geq 0 \qquad \text{for all } i = 1, \ldots, n$$

И гиперплоскость определена векторами $w^T x + b = 0$, with $w \in F$ and $b \in R$.

Таким образом что максимизируется расстояние от гиперплоскости до объектов разных классов.

http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/

SVM kernel trick



https://miro.medium.com/max/1676/1*mCwnu5kXot6buL7jeIafqQ.png

One-Class SVM according to Schölkopf

Объекты класса остаются за гиперплоскостью

$$\min_{w,\,\xi_i,\,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

subject to:

$$(w \cdot \phi(x_i)) \geq \rho - \xi_i \qquad \text{for all } i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad \text{for all } i = 1, \ldots, n$$

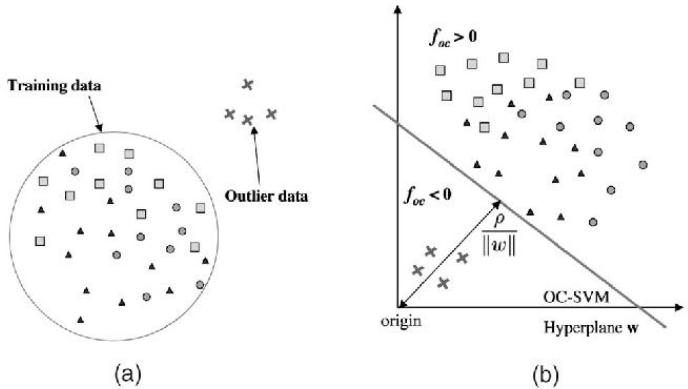


Fig. 2. Using margin as part of GE. (a) Input space showing scatterplot of

http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/
https://d3i71xaburhd42.cloudfront.net/7b00f6a6de6b172903d5f1f27b9418ec9f593885/4-Figure2-1.png

One-Class SVM according to Tax and Duin

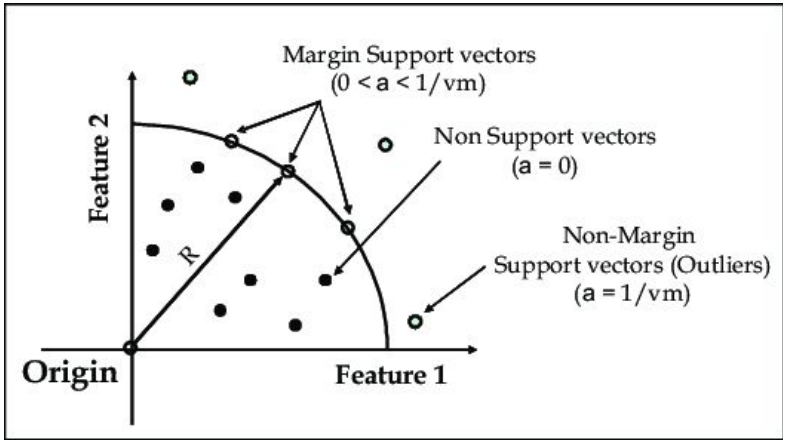
Объекты класса помещаются в окружность (гиперсферу) радиуса R, функционал:

$$\min_{R,\,\mathbf{a}} R^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$\|x_i - \mathbf{a}\|^2 \leq R^2 + \xi_i \qquad \text{for all } i = 1,\ldots,n$$

$$\xi_i \geq 0 \qquad \text{for all } i = 1,\ldots,n$$



http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/
https://www.researchgate.net/profile/Paul_Havinga/publication/221192293/figure/fig1/AS:670040785424388@1536761580202/Geometry-of-the-quarter-sphere-formulation-of-one-class-SVM.png

Метрики оценивания алгоритмов кластеризации?

Почему не подходят метрики точности классификации?

Метрики оценивания алгоритмов кластеризации

- Полнота (completeness)

all members of a given class are assigned to the same cluster.

- Гомогенность (homogeneity)

each cluster contains only members of a single class

- v_ score, silhouette score

v = 2 * (homogeneity * completeness) / (homogeneity + completeness)

https://scikit-learn.org/0.18/modules/generated/sklearn.metrics.completeness_score.html#sklearn.metrics.completeness_score

Метрики оценивания алгоритмов кластеризации

```
>>> from sklearn import metrics
>>> labels true = [0, 0, 0, 1, 1, 1]
>>> labels_pred = [0, 0, 1, 1, 2, 2]

>>> metrics.homogeneity_score(labels_true, labels_pred)
0.66...

>>> metrics.completeness_score(labels_true, labels_pred)
  0.42...
```

https://esigma6.wordpress.com/2018/11/03/2-3-9-3-homogeneity-completeness-and-v-measure/