Поиск выбросов и аномалий

Лекция №3

Лектор: Шевляков Артём

Напоминаю определения

1. Поиск выбросов (outlier detection). Есть множество объектов M. Найти в нем все аномальные объекты.

2. Поиск новизны (novelty detection). Есть множество объектов M. Определить, является ли объект $A \notin M$ похожим на объекты из M или нет?

Отличие выбросов от новизны и пропусков

выброс VS пропуск в данных

Выброс – это часто реально существующий объект, но обладающий аномальными свойствами, он сильно отличается от других объектов выборки.

выброс VS новизна

Новизна считается по отношению к старой выборке объектов. А выброс является аномальным уже для своих «соседей» по выборке.

Примеры выбросов

1. (из Википедии) если наугад измерять температуру предметов в комнате, получим цифры от 18 до 22 °C, но радиатор отопления будет иметь температуру в 70° - и это выброс, не типичное значение!

Примеры выбросов

2. На матфаке ОмГУ я проводил анкетирование порядка 60 студентов: просил их написать средние баллы за все сессии. Выбросом оказался…

Примеры выбросов

2. На матфаке ОмГУ я проводил анкетирование порядка 60 студентов: просил их написать средние баллы за все сессии. Выбросом оказался КРУГЛЫЙ ОТЛИЧНИК (что было установлено с помощью алгоритма поиска выбросов)

Зачем нужно искать и уничтожать выбросы?

1. Если данные будут использоваться при решении задачи предсказания, то удаление выбросов, как правило, повышает точность предсказания (ибо правило «мусор на входе – мусор на выходе» никто не отменял).
2. Удаление выбросов позволяет получить нормальные (типичные, эталонные) объекты.
3. Многие характеристики (например, среднее значение) очень чувствительны к наличию выбросов.

Идеальных методов обнаружения выбросов не бывает потому, что

1. … не существует формального определения выброса.
2. … алгоритм, беспощадный к выбросам, будет удалять и часть «нормальных» объектов.
3. … алгоритм, гуманный к «нормальным» объектам, будет пропускать часть выбросов.

Построить идеальный детектор выбросов – это всё равно что предложить мед. анализ без ложноположительных и ложноотрицательных результатов.

Методы обнаружения выбросов

I. Поиск аномальных объектов с помощью здравого смысла. Например, если известен нормальный диапазон для значений признака.

Пример: человек с ростом более 200см (такие люди могут в реальности существовать, но их очень мало). Таких людей лучше объявить выбросами.

Методы обнаружения выбросов

II. Методы, основанные на анализе одного признака (каждый признак берётся отдельно и ищутся объекты аномальными значениями этого признака).

III. Методы, основанные на одновременном анализе нескольких признаков.

Методы, анализирующие признаки по отдельности

Методы, анализирующие один признак

Вот у вас есть значения признака

$$P = (p_1, p_2, \dots, p_n)$$

Основная идея поиска аномалий:
найти значения p_i, расположенные вдали от среднего значения (медианы).
Далее используем обозначения:
$\bar{p}$ - среднее значение, n - объем выборки,
s_P - отклонение

Простейшие методы

1. Удалить все объекты, у которых величина $|p_i - \bar{p}|$ слишком велика.
2. Удалить все объекты, у которых величина $\frac{|p_i - \bar{p}|}{S_P}$ слишком велика.
   (см. след. слайд)
3. Более продвинутый критерий Шовене (см. ниже)
4. Определение выбросов без использования среднего и отклонения.

Пожалуй, самое простое правило для определения выброса

1. Пусть *x* – подозрительное значение.
2. Исключим *x* из выборки и по оставшимся элементам вычислим среднее $\bar{p}$ и отклонение s_P.
3. Если выборка симметричная (см. лекцию 1), то *x* будет выбросом, если он не принадлежит интервалу

$$(\bar{p} - 3s_P, \bar{p} + 3s_P)$$

4. Если выборка не симметричная, то *x* будет выбросом, если он не принадлежит интервалу

$$(\bar{p} - 5s_P, \bar{p} + 5s_P)$$

Критерий Шовене (Chauvenet)

Критерий Шовене (Chauvenet)

Значение p_i является выбросом, если выполнено неравенство

$$erfc\left(\frac{|p_i - \bar{p}|}{s_P}\right) < \frac{1}{2n}$$

где

$$erfc(x) =$$ 18+ censored

дополнение к функции ошибок.

Критерий Шовене (Chauvenet)

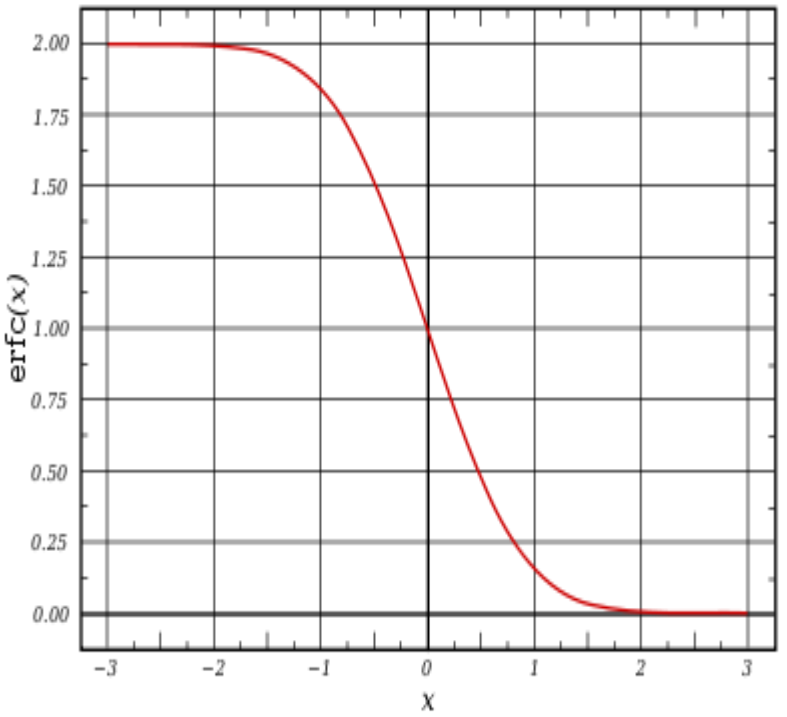
Значение p_i является выбросом, если выполнено неравенство

$$erfc\left(\frac{|p_i - \bar{p}|}{s_P}\right) < \frac{1}{2n}$$

где

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t} dt$$

дополнение к функции ошибок.



Пример

Есть *n=14* объектов со следующими значениями признака *P*

8.02 8.16 3.97 8.64 0.84 4.46 0.81 7.74 8.78 9.26 20.46 29.87 10.38 25.71

Вычисляем: $\bar{p}$ =10.51, S_P =8.77.

Проверка для 25.71:

$$erfc\left(\frac{|25.71-10.51|}{8.77}\right)=0.014<0.036=\frac{1}{2*14}$$

то есть это выброс!!!

Пример

Есть *n=14* объектов со следующими значениями признака *P*

8.02 8.16 3.97 8.64 0.84 4.46 0.81 7.74 8.78 9.26 20.46 29.87 10.38 25.71

Вычисляем: $\bar{p}$ =10.51, S_P =8.77.

Проверка для 29.87 также говорит: «выброс». А вот 20.46 уже не выброс:

$$erfc\left(\frac{|20.46-10.51|}{8.77}\right)=0.11>0.036=\frac{1}{2*14}$$

Пример (вторая итерация)

Осталось *n=12* объектов со следующими значениями признака *P*

8.02 8.16 3.97 8.64 0.84 4.46 0.81 7.74 8.78 9.26 20.46 10.38

Вычисляем: $\bar{p}$ =7.63, S_P =5.17.

Проверка для 20.46:

$$erfc\left(\frac{|20.46-7.63|}{5.17}\right)=0.00045<0.042=\frac{1}{2*12}$$

то есть теперь 20.46 стало выбросом!!!

Пример (вторая итерация)

Осталось *n=12* объектов со следующими значениями признака *P*

8.02 8.16 3.97 8.64 0.84 4.46 0.81 7.74 8.78 9.26 20.46 10.38

Вычисляем: $\bar{p}$ =7.63, S_P=5.17.

Проверка для 10.38:

$$erfc\left(\frac{|10.38-7.63|}{5.17}\right)=0.452>0.042=\frac{1}{2*12}$$

то есть 10.38 не выброс. Остальные числа также проходят проверку.

Пример (третья итерация)

Осталось *n=11* объектов со следующими значениями признака *P*

8.02 8.16 3.97 8.64 0.84 4.46 0.81 7.74 8.78 9.26 10.38

Вычисляем: $\bar{p}$ =6.46, S_P=3.38.

Проверка для 10.38:

$$erfc\left(\frac{|10.38-6.46|}{3.38}\right)=0.1>0.045=\frac{1}{2*11}$$

то есть 10.38 не выброс. Остальные числа также проходят проверку. КОНЕЦ работы, так как новые выбросы не появляются.

Настройка критерия Шовене

Значение p_i является выбросом, если выполнено неравенство

$$erfc\left(\frac{|p_i - \bar{p}|}{s_P}\right) < \frac{1}{2n}$$

Константу 2 (в формуле) можно заменить на любую другую константу. Это сделает критерий более беспощадным (либо более лояльным) к выбросам.

Поиск выбросов без использования среднего и отклонения

Философское затруднение

Значения среднего и отклонения сильно чувствительны к наличию выбросов.

Таким образом, возникает замкнутый круг: мы ищем выбросы с помощью среднего и отклонения, чьи значения как раз и обусловлены наличием выбросов.

Возникает идея искать выбросы без использования величин, которые сильно зависят от выбросов.

Критерий, основанный на квартилях выборки

1-ая квартиль Q_{25}: это такое число, что ровно 25% выборки меньше его.

2-ая квартиль Q_{50}: это такое число, что ровно 50% выборки меньше его (фактически это - …)

Критерий, основанный на квартилях выборки

1-ая квартиль Q_{25}: это такое число, что ровно 25% выборки меньше его.

2-ая квартиль Q_{50}: это такое число, что ровно 50% выборки меньше его (фактически это медиана).

3-ья квартиль Q_{75}: это такое число, что ровно 75% выборки меньше его.

Например, для выборки

(-10,0,1,2,4,5,5,6,7,100)

имеем $Q_{25}=1$ $Q_{50}=4.5$ (медиана) $Q_{75}=6$

Как быстро найти Q_{25} (Q_{75})? Нужно вычислить медиану и разбить выборку на две части: A={что меньше медианы}, B={что больше медианы}. Тогда Q_{25} (Q_{75}) будет медианой в A (B).

Факт: 50% элементов выборки принадлежат интервалу $[Q_{25}, Q_{75}]$.

Идея: элементы, которые сильно далеки от интервала $[Q_{25}, Q_{75}]$ можно объявить выбросами.

Поиск выбросов с помощью квартилей

Правило: если элемент не попадает в интервал

$$(Q_{25}-1.5*(Q_{75}-Q_{25}),\ Q_{75}+1.5*(Q_{75}-Q_{25})),$$

то он объявляется выбросом.

В примере (-5,0,1,2,4,5,5,6,7,100) получаем $Q_{25}=1$, $Q_{75}=6$

Интервал будет такой: (1-1.5*5, 6+1.5*5)=

(-6.5,13.5).

Таким образом, элемент 100 – выброс.

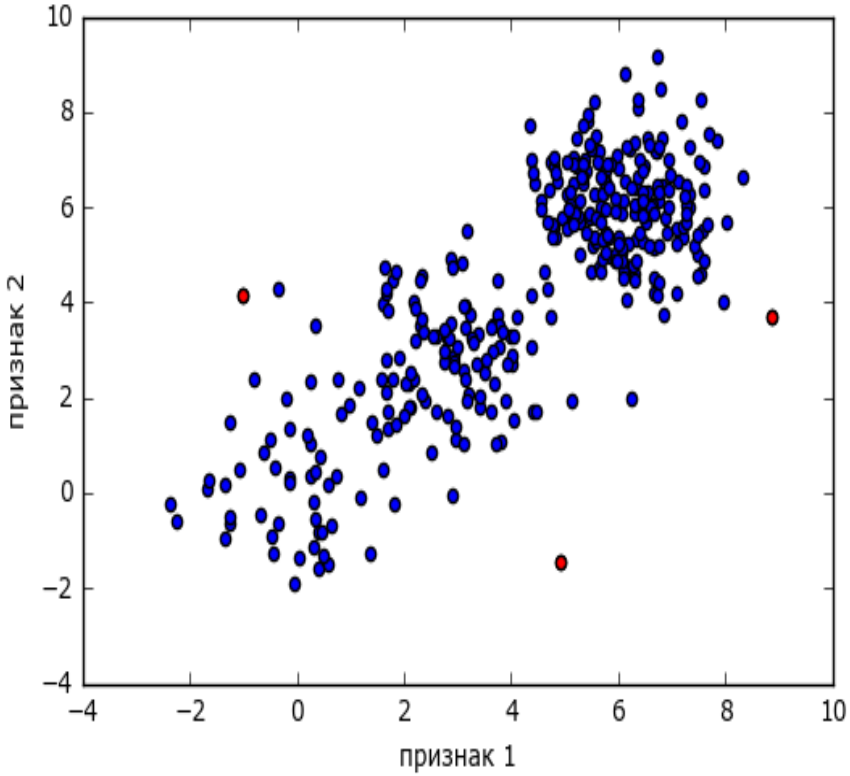
Методы, анализирующие несколько признаков

Недостатки методов, которые анализируют 1 признак (1-й недостаток)

В выборке (1, 50, 2, 1, 101, 2, 1, 100, 1, 3, 101, 1, 3, 100, 101, 100, 100) число 50, очевидно, аномально. Но поскольку 50 очень близко к среднему значению, то все описанные выше методы его не заметят.

Недостатки методов, которые анализируют 1 признак (2-й недостаток)

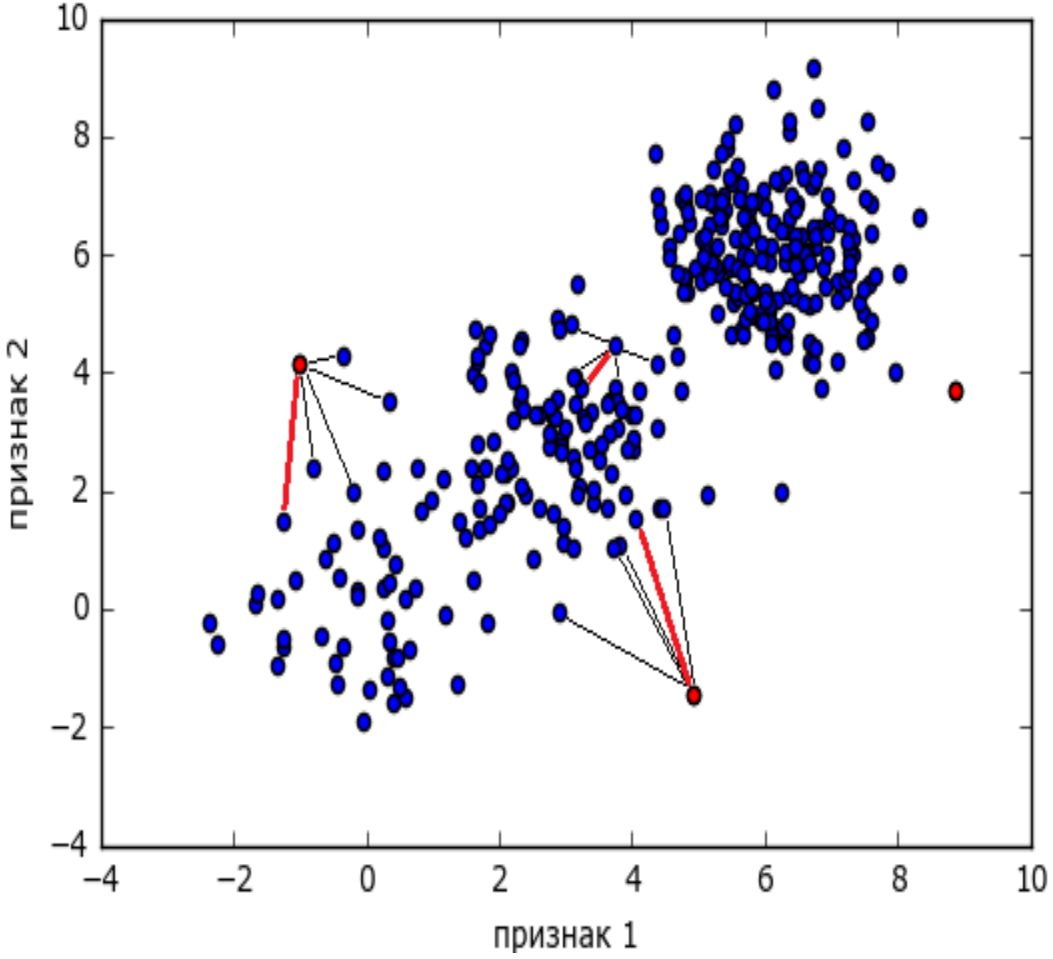
Аномалия часто характеризуется не только экстремальными значениями отдельных признаков. На картинке каждый выброс имеет «нормальное» значения каждого признака, но их комбинация приводит к аномалии.



Метрические методы

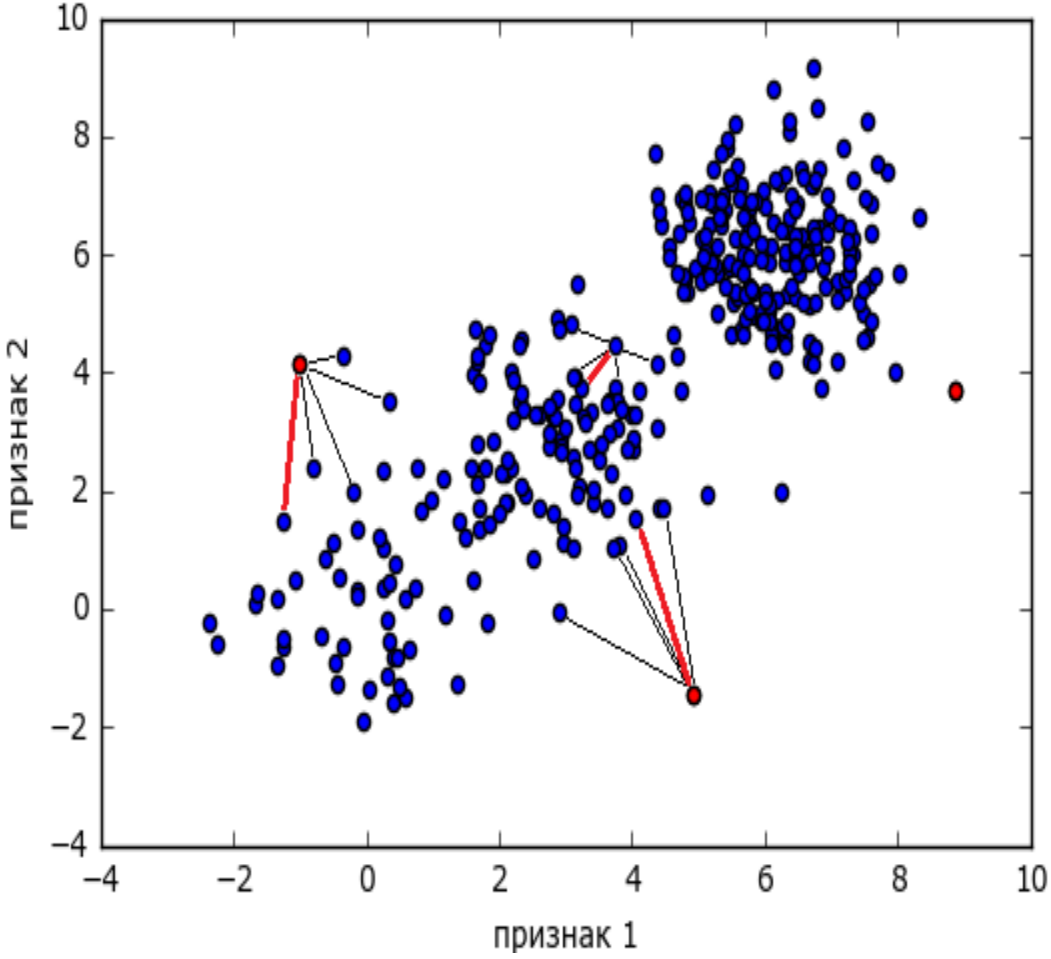
Представим объекты с m признаками с помощью точек в пространстве R^m.

Идея: у выброса мало соседей, а у типичного объекта много.



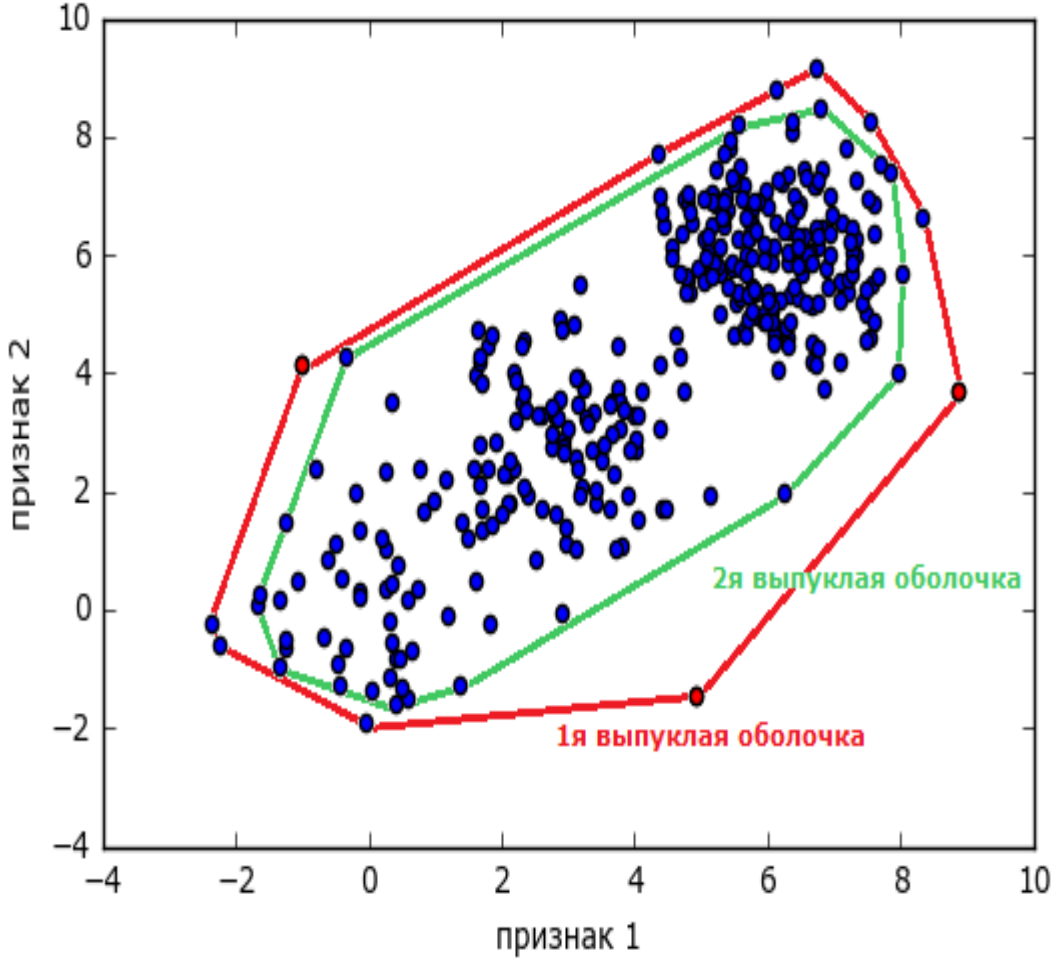
Метрические методы

Можно найти расстояние от каждого объекта до его ближайшего соседа. У выбросов такое расстояние будет большим (можно предложить и другие варианты алгоритма).



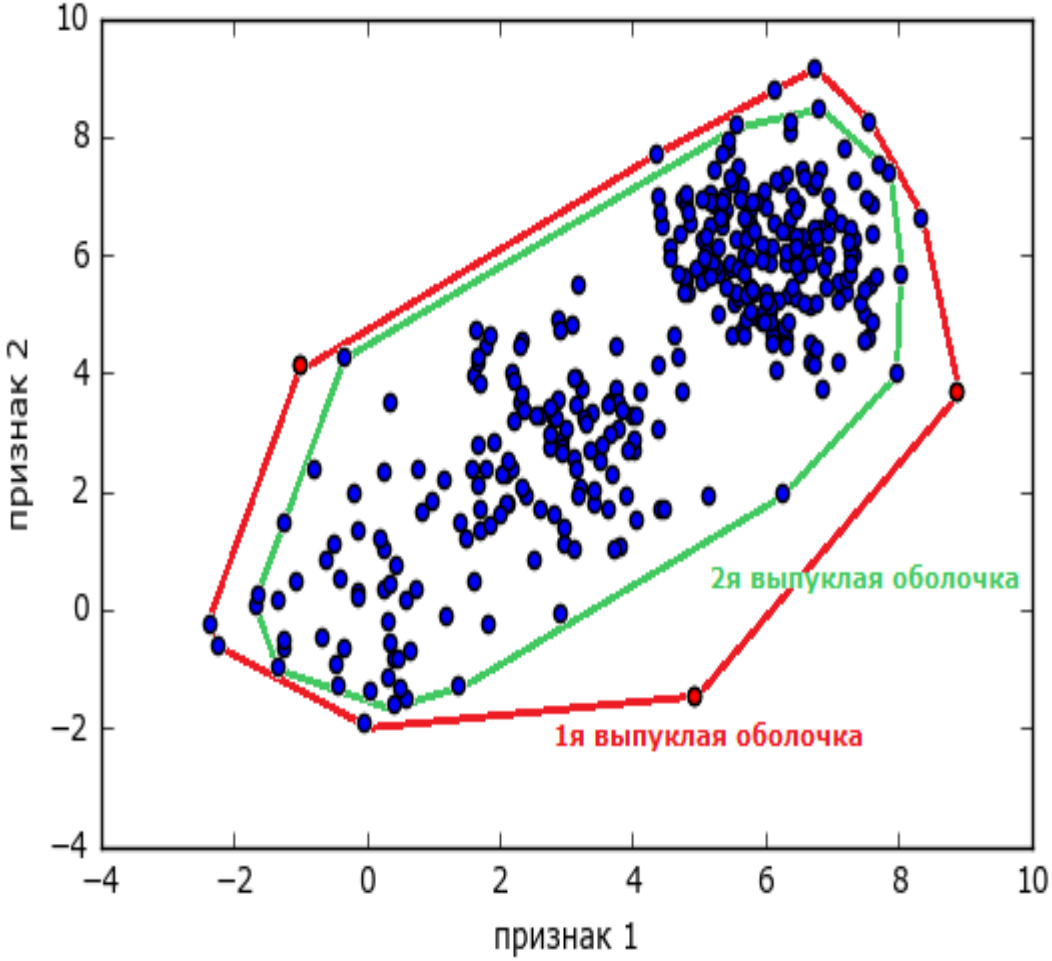
Геометрические методы

Можно вычислить выпуклую оболочку объектов (как точек в m-мерном пространстве). Выбросами будут …



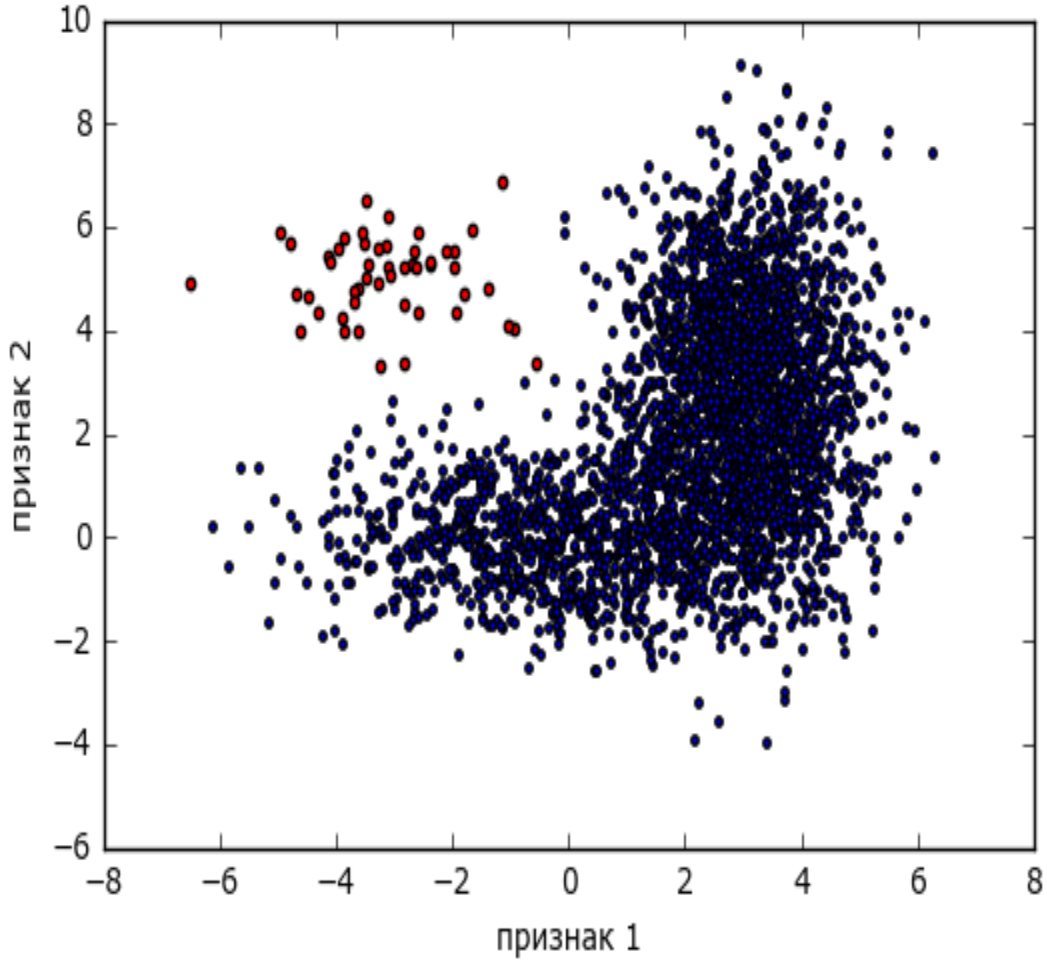
Геометрические методы

Можно вычислить выпуклую оболочку объектов (как точек в m-мерном пространстве). Выбросами будут объекты на границе. (Их можно удалить, а процедуру повторить еще несколько раз).



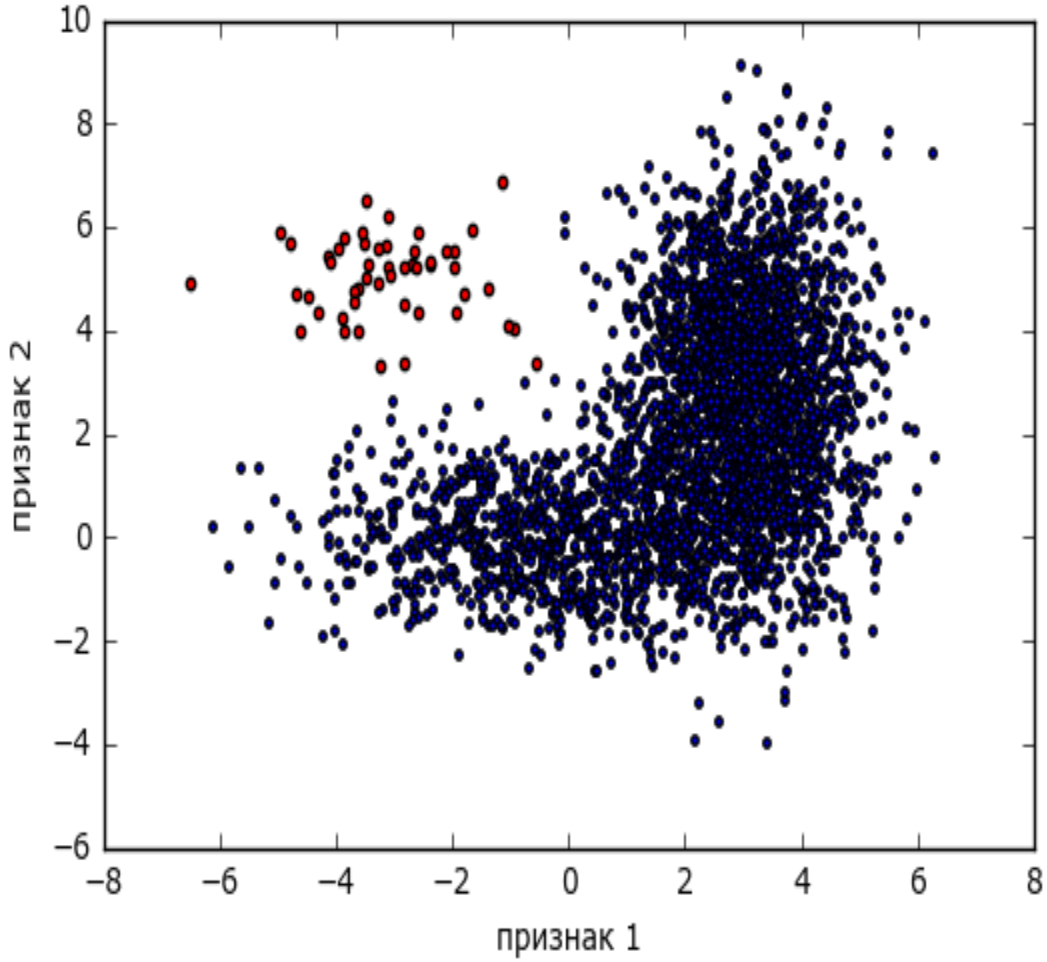
Поиск выбросов с помощью кластеризации

Можно запустить алгоритм кластеризации (подробности в след. лекции). Он разобьет объекты на группы. Выбросы – это ...



Поиск выбросов с помощью кластеризации

Можно запустить алгоритм кластеризации (подробности в след. лекции). Он разобьет объекты на группы. Выбросы – это элементы малых (в том числе и одноэлементных) групп.



Поиск выбросов с помощью моделей предсказания

1) Некоторые вариации метода опорных векторов (SVM) позволяют находить выбросы.
2) Вариация решающих деревьев (decision trees) под названием «изолирующий лес» может искать выбросы.

Подробнее об этом мы поговорим на соответствующих лекциях.

Поиск выбросов с помощью моделей предсказания

3) Можно запустить модель предсказания признака P по другим признакам таблицы. Объекты, для которых …

Поиск выбросов с помощью моделей предсказания

3) Можно запустить модель предсказания признака P по другим признакам таблицы. Объекты, для которых предсказанное и истинное значение P различаются очень сильно, можно объявить выбросами.

Поиск новизны

Поиск новизны можно свести к поиску выбросов

(Поиск новизны) Есть множество объектов M. Определить, является ли объект $A \notin M$ похожим на объекты из M или нет?

Переход к поиску выбросов: добавляем объект A в множество M и запускаем алгоритм поиска выброса. Если A будет детектирован как выброс, то A являлся новизной.

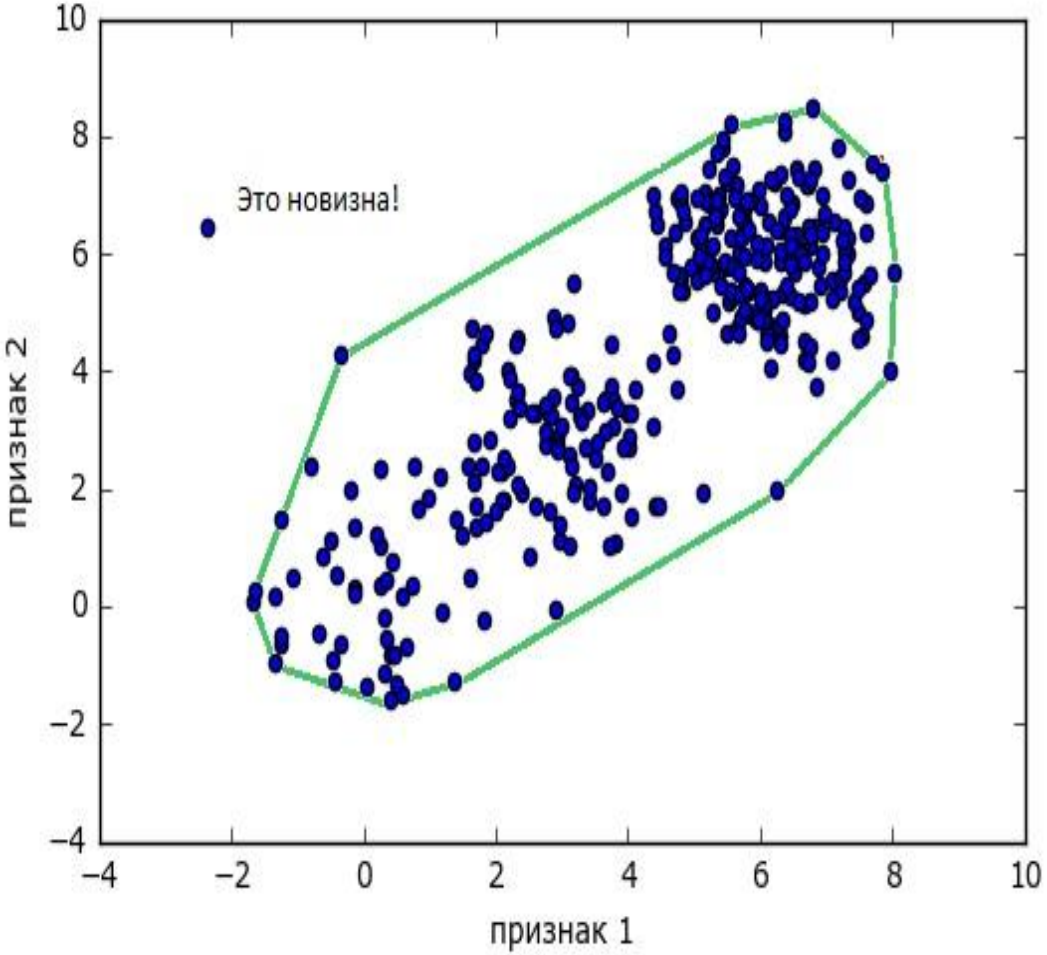
Но есть и алгоритмы, которые ищут новизну, не используя поиск выбросов.

Поиск новизны без поиска выбросов

Строим выпуклую оболочку всех объектов из выборки. Объект А будет считаться новизной, если…

Поиск новизны без поиска выбросов

Строим выпуклую оболочку всех объектов из выборки. Объект А будет считаться новизной, если он НЕ попадает в эту выпуклую оболочку.



Источники картинок и идей

1. https://alexanderdyakonov.wordpress.com/2017/04/19/поиск-аномалий-anomaly-detection/ (отсюда же взяты и картинки)
2. Статья «Cleaning Data the Chauvenet Way» by Lily Lin, Paul Sherman