Урок 5

Поиск аномалий

5.1. Задача обнаружения аномалий

5.1.1. Связь обнаружения аномалий с другими задачами машинного обучения

В курсе уже встречалось несколько примеров задач поиска структуры данных (обучения без учителя). Например, задача кластеризации, в которой требуется найти такие группы объектов, что объекты внутри внутри каждой из них были похожи друг на друга (рисунок 5.1).

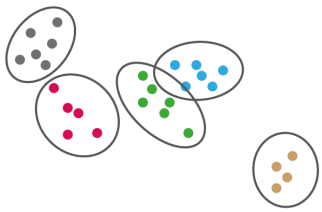

Рис. 5.1: Пример задачи кластеризации

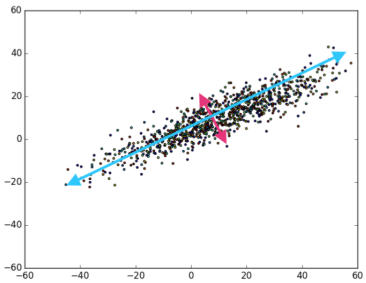

Рис. 5.2: Пример задачи понижения размерности

Другой пример задачи поиска структуры в данных — это задача понижения размерности, когда имеется некоторая выборка, и требуется её спроецировать в пространство меньшей размерности так, чтобы сохранить как можно больше информации (рисунок 5.2), то есть найти какие-то направления, которые наиболее информативны для данной выборки.

Задача поиска или обнаружения аномалии немного отличается от вышеуказанных. В ней нужно найти в выборке объекты, которые не похожи на большинство объектов, которые выделяются, являются аномальными (рисунок 5.3). При этом примеров аномалий либо нет вообще, либо их очень мало. Именно поэтому эта задача относится к обучению без учителя: данные не размечены. Итак, необходимо научиться понимать, похож ли новый объект на остальные, те, которые были известны до этого.

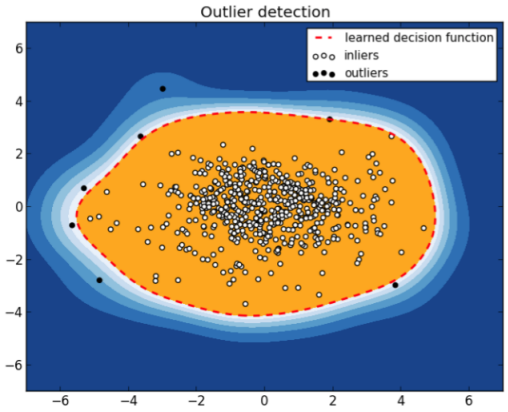

Рис. 5.3: Пример задачи поиска аномалий

5.1.2. Применение обнаружения аномалий на практике

Полезно рассмотреть несколько примеров применения задачи обнаружения аномалий. **Первый пример** касается изучения клиентов банка. Пусть каждый объект — это клиент банка в определенный момент времени. Его можно описать характеристиками транзакций в данный момент, поведением в интернет-банке и т.д. Вопрос, на который требуется ответить: не является ли его поведение необычным? Если оно выбивается из среднего по всем клиентам, то это повод заподозрить, что клиент делает что-то не так. Возможно, это мошенник, который украл карту и пытается вывести с нее деньги — будет повод заблокировать карту и позвонить клиенту банка.

Другой пример задачи обнаружения аномалий — это мониторинг сложной компьютерной системы, которая состоит из большого количества взаимосвязанных машин. Можно отслеживать много показателей: загрузку процессоров, использование памяти на каждой машине, нагрузку на сеть и т.д. Вопрос, на которой требуется ответить: отличается ли текущее состояние системы от характеристик тех состояний, про которые известно, что они нормальные. Если отличается, и система ведет себя как-то иначе, то это повод задуматься, не случилась ли какая-то поломка, нужно ли продиагностировать систему и что-то починить в ней.

Наконец, **третий пример**. Пусть имеется модель, которая по отзыву о банке определяет его тональность: позитивную или негативную. Процедура следующая: клиент заходит на сайт банка, в специальную форму вводит некоторый отзыв, дальше этот отзыв приходит на вход модели, которая определяет, позитивный он или негативный. Если он негативный, то нужно сообщить об этом сотрудникам банка, чтобы они решили возникшую проблему. Но помимо этого хотелось бы понимать, когда приходит новый отзыв, применима ли к нему имеющаяся модель машинного обучения, возможно ли его классифицировать этой же моделью. Дело в том, что распределение признаков этого объекта могло измениться. Например, банк мог поменять название продуктов, и поэтому теперь слова, встречающиеся в отзыве, совершенно другие, — модель к ним не готова. Или банк мог изменить ограничение на длину отзыва. До клиенты писали довольно длинные тексты, а теперь длину ограничили, из-за этого отзыв должен быть очень коротким. В этом случае объекты станут совершенно другими: клиенты будут стараться максимально сжато объяснить свою проблему. Из-за этого модель может стать непригодной для решения задачи. Если стало известно, что объекты стали другими, аномальными, по сравнению с теми, на которых обучалась модель, то это повод её обучить на новых размеченных данных.

Далее будут описаны два подхода к обнаружению аномалий. Первый основан на восстановлении плотности распределения, второй подход использует методы классификации.

5.2. Параметрическое восстановление плотности

5.2.1. Вероятностный подход к обнаружению аномалий

В вероятностном подходе к обнаружению аномалий считается, что аномалия — это объект, который был получен из распределения, отличного от того, с помощью которого сгенерирована обучающая выборка. Возникает вопрос: как найти распределение, из которого была получена выборка? Если найти это распределение, то можно оценить вероятность принадлежности нового объекта этому распределению. Если вероятность получить новый объект из этого распределения очень мала, то это, скорее всего, аномалия.

Существует три основных подхода к восстановлению вероятностных плотностей:

- параметрический подход,
- непараметрический подход,
- восстановление смесей.

5.2.2. Параметрический подход

Итак, существует некоторое вероятностное распределение $p(x)$ на всех объектах, которые можно получить. Считается, что это распределение является параметрическим:

$$p(x) = \phi(x|\theta),$$

где θ задаёт параметры распределения.

Один из самых известных примеров параметрического семейства распределений — это нормальное распределение (рисунок 5.4):

$$\phi(x|\theta) = \mathcal{N}(\mu, \Sigma)$$

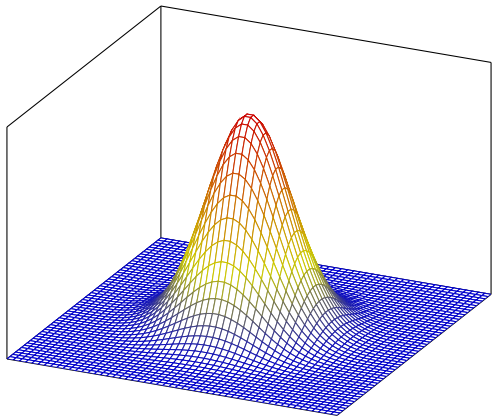

Рис. 5.4: Нормальное распределение

Нормальное распределение задаётся параметрами μ, Σ (центр и ковариационная матрица). По форме оно похоже на шляпу, при этом параметр μ определяет, где находится центр этой шляпы, а параметр Σ — то, насколько она сосредоточена вокруг центра или размазана по всему пространству. Таким образом, если пытаться моделировать выборку с помощью нормального распределения, то необходимо определить параметры μ и Σ. Как это делать?

5.2.3. Метод максимального правдоподобия

Логично искать параметр распределения так, чтобы максимизировать вероятность объектов обучающей выборки быть порождёнными этим распределением. В этом случае объекты, которые не похожи на эту выборку будут получать низкие вероятности. Именно так работает метод максимального правдоподобия, который старается подобрать такое распределение из параметрического семейства, что с его точки зрения объекты

обучающей выборки будут как можно более вероятны. Работать с самим правдоподобием неудобно, поскольку это — произведение значений плотности во всех точках обучающей выборки. Вместо можно взять его логарифм и пытаться максимизировать полученную сумму:

$$\sum_{i=1}^{\ell} \log\phi(x_i|\theta) \to \max_{\theta}$$

Для некоторых распределений эту задачу можно решить аналитически, если посчитать частные производные и приравнять их к 0. Например, для нормального распределения такие решения существуют:

$$\mu = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i$$

$$\Sigma = \frac{1}{\ell}\sum_{i=1}^{\ell} (x_i - \mu)(x_i - \mu)^T$$

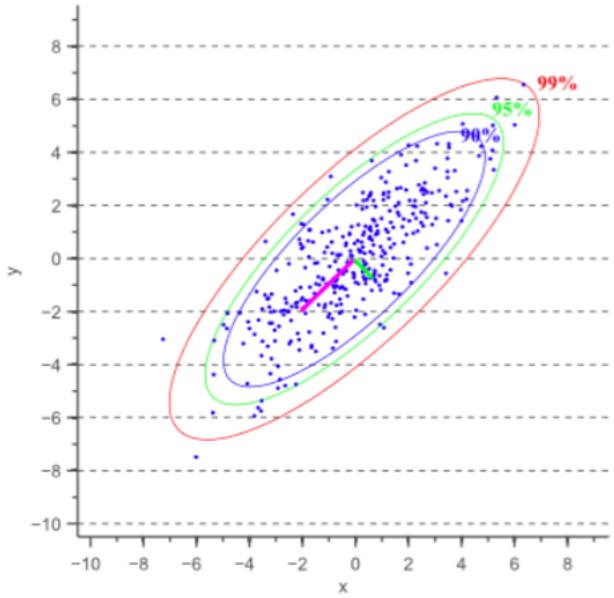

Рис. 5.5: Выборка, порождённая нормальным распределением

Пусть выборка, изображённая на рисунке 5.5, порождена нормальным распределением. Если найти параметры этого распределения методом максимального правдоподобия, то оно будет выглядеть так, как задают линии уровная на рисунке: внутри синей линии уровня находится 90 % всей вероятности, внутри зеленой — 95 %, внутри красной — 99 %. Таким образом, вероятность получить объект вне красного эллипса очень мала. При этом в выборке присутствуют две точки, которые находятся вне красного эллипса. Поскольку основная выборка очень хорошо описывается этим нормальным распределением, а те две синие точки им описываются плохо, то можно предположить, что они пришли из другого распределения, что это аномалия.

Итак, если уже найдено некоторое распределение $p(x)$, и приходит новый объект x, необходимо вычислить вероятность порождения этого объекта данным распределением и сравнить её с некоторым порогом t. Если вероятность меньше этого порога, объект объявляется аномалией.

Возникает вопрос: как выбирать порог t? На этот счёт не существует однозначных рекомендаций. Можно, например, выбирать его из априорных соображений (как было в примере, объявлять аномалиями все объекты, находящиеся вне линии уровня 99 %). Или, если имеются объекты, про которые точно известно, что это — аномалии, то можно подобрать порог t так, чтобы эти объекты были объявлены аномальными, а все остальные — сгенерированными из распределения $p(x)$.

5.2.4. Модель смеси распределений, EM-алгоритм

В некоторых случаях параметрического подхода оказывается недостаточно. Например, на рисунке 5.6 изображена выборка, которая сгенерирована из двух номральных распределений с одинаковыми матрицами ковариаций, но разными центрами, таким образом, получается два облака точек. Описать эту выборку одним нормальным распределением будет невозможно, зато для этого отлично подходит модель смеси распределений.

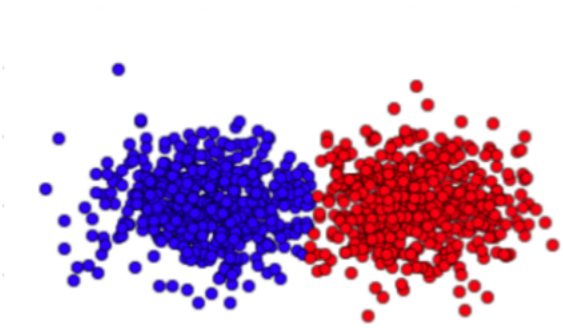

Рис. 5.6: Выборка, порождённая двумя нормальными распределениями

Смесью называется такое распределение $p(x)$, которое представляется в виде взвешенной суммы других распределений:

$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \phi(x|\theta_j)$$

Распределения $p_j(x)$ называются компонентами смеси, и, как правило, они являются параметрическими распределениями. Собственно, каждая компонента p_j является членом параметрического семейства $\phi(x)$ со своим параметром θ_j.

Смеси распределений уже упоминались в уроке, когда рассказывалось про кластеризацию с помощью EM-алгоритма, который можно использовать и для решения этой задачи. Этот алгоритм состоит из повторения E-шага и M-шага до тех пор, пока не будет достигнута сходимость. На E-шаге вычисляются апостериорные вероятности того, что объект i принадлежит компоненте j смеси:

$$g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

На M-шаге апостериорные вероятности используются, чтобы обновить оценки на параметров θ. Эти оценки вычисляются путем решения задачи максимизации взвешенного правдоподобия:

$$w_j = \frac{1}{N} \sum_{j=1}^{N} g_{ji}$$

$$\theta_j = \text{argmax}_\theta \sum_{j=1}^{N} g_{ji} \ln \phi(\theta, x).$$

Благодаря этому алгоритму можно определить, какая именно смесь из K распределений порождает выборку.

5.3. Непараметрическое восстановление плотности

5.3.1. Формула Парзена-Розенблатта и его параметры

Непараметрический подход к восстановлению плотности состоит в том, что вид распределения пытаются восстановить, не вводя никаких семейств распределений, используя только сами данные.

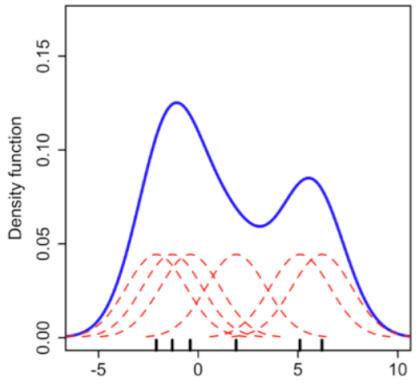

Рис. 5.7: Пример непараметрического восстановления плотности

Пусть имеется одномерная выборка (рисунок 5.7, объекты выборки обозначены засечками по оси абсцисс). В каждой точке обучающей выборки помещают центр небольшой гауссианы (показаны красной линией), таким образом всем точкам на оси присваиваются некоторые вероятности. Далее, в каждой точке эти гауссианы суммируются, и получается итоговое распределение (на рисунке показано синим).

Формально это можно сделать с помощью формулы Парзена-Розенблатта:

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right),$$

где $K(r)$ — ядро, параметр метода, характеризующий вероятность того, что точки x и x_i похожи друг на друга. Ядро — это чётная функция, также для него должно выполняться условие

$$\int K(r)dr = 1.$$

Если это требование не выполнено, плотности будут получаться ненормированными. Всё написанное выше верно для одномерных выборок, x и x_i — это вещественные числа.

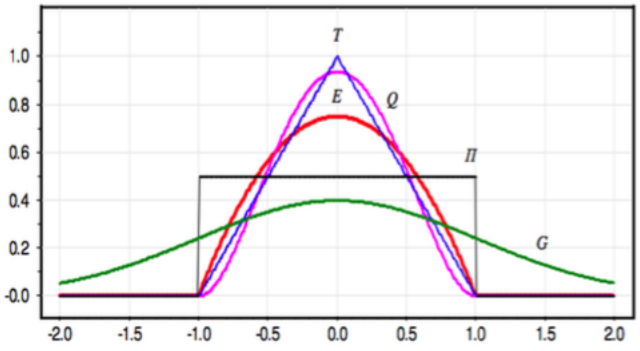

Рис. 5.8: Примеры ядер

Существует много различных ядер. Некоторые из них изображены на рисунке 5.8:

- $E(r) = \frac{3}{4}(1 - r^2)[|r| \leq 1]$ — оптимальное;
- $Q(r) = \frac{15}{16}(1 - r^2)^2[|r| \leq 1]$ — квартическое;
- $T(r) = (1 - |r|)[|r| \leq 1]$ — треугольное;

- $G(r) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}r^2)$ — гауссовское;
- $\Pi(r) = \frac{1}{2}[|r| \leq 1]$ — прямое.

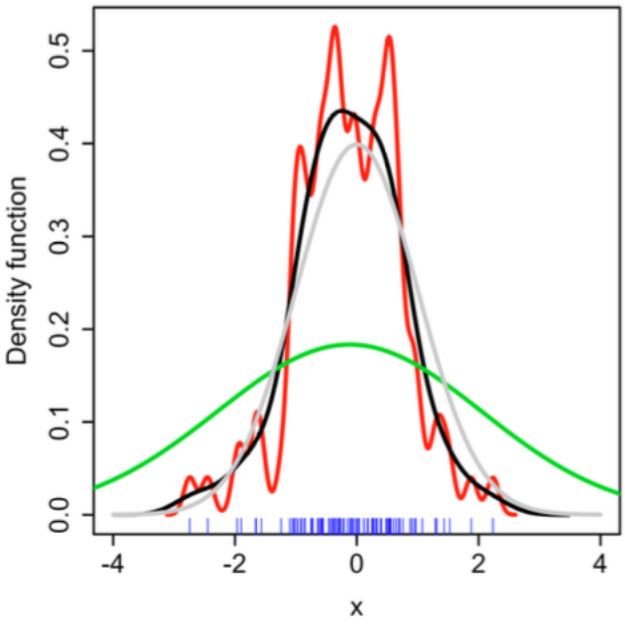

Рис. 5.9: Влияние ширины окна на восстановленную плотность вероятности

Еще один параметр оценки Парзена-Розенблатта — это ширина окна h. Влияние этого параметра на результирующую плотность вероятности можно рассмотреть на примере выборки, изображённой на рисунке 5.9. Элементы выборки на изображены синими засечками на оси абсцисс. Красной, черной и зеленой кривыми показаны непараметрические оценки с гауссовским ядром для разных значений ширины окна. Красная кривая соответствует очень маленькой ширине окна, и результирующая плотность очень чувствительна ко всем точкам. Она получается не очень гладкой и, скорее всего, переобученной. Черная кривая соответствует более высокому значению ширины окна. Эта плотность очень неплохо восстанавливает нормальное распределение, из которого были сгенерированы данные (на рисунке показана серым цветом). Зеленая кривая соответствует оценке плотности с очень большой шириной окна, результирующая плотность характеризуется большой дисперсией. Даже точки на прямой, которые очень далеки от объектов обучающей выборки, получают высокое значение плотности из-за того, что окно широкое.

5.3.2. Многомерный случай, возникающие проблемы

Непараметрическое оценивание плотности хорошо обобщается на многомерный случай, при этом необходимо разность между точками x и x_i заменить метрикой и ввести нормировочную константу $V(h)$, чтобы плотность была нормирована:

$$p_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

$$V(h) = \int K\left(\frac{\rho(x, x_i)}{h}\right) dx$$

У многомерного непараметрического подхода к восстановлению плотности есть одна большая проблема. Для примера можно рассмотреть выборку, изображённую на рисунке 5.10. Если спроецировать эту выборку на одну ось, то для одномерного случая имеется достаточно объектов, чтобы восстановить плотность вероятности. Однако в двумерном пространстве на один элемент площади приходится гораздо меньше объектов,

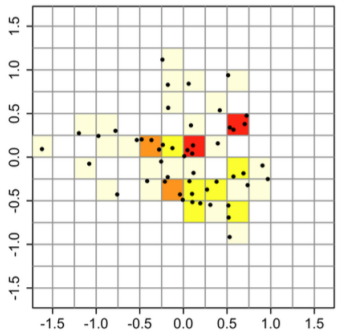

Рис. 5.10: Пример многомерное непараметрического восстановления плотности

чем на соответствующий ему отрезок в одномерном пространстве. Поэтому оценка плотности становится более шумной и менее репрезентативной. Число объектов, необходимых для восстановления плотности, растёт экспоненциально с увеличением размерности пространства, и на практике этот подход применим в пространствах не очень высокой размерности.

5.3.3. Обнаружение аномалий с использованием непараметрического подхода

Обнаружение аномалий с помощью этого подхода происходит так же, как и с использованием параметрического восстановления плотности: для каждого нового объекта вычисляется плотность вероятности $p(x)$, если $p(x) < t$, где t — заданный порог, то объект считается аномалией.

5.4. Одноклассовый SVM

5.4.1. Связь обнаружения аномалий с задачей классификации

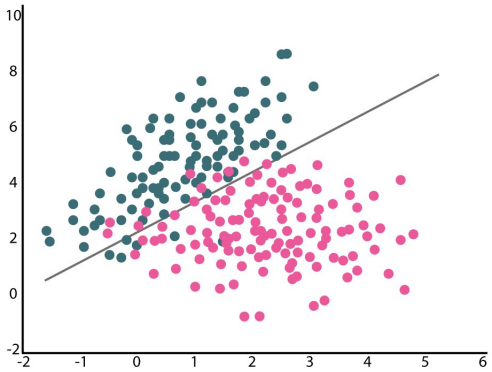

Рис. 5.11: Задача классификации

Задачу обнаружения аномалий можно связать с задачей классификацией. В задаче бинарной классификации (рисунок 5.11) 2 класса необходимо разделить прямой (в пространствах высокой размерности — гиперплоскостью).

В задаче обнаружения аномалий тоже имеется выборка, но требуется построить некоторую кривую, которая отделит выборку от всего остального (рисунок 5.12). Все, что находится вне этой кривой, будет объявляться аномалиями, потому что это нечто, что не попадает в множество типичных объектов из выборки.

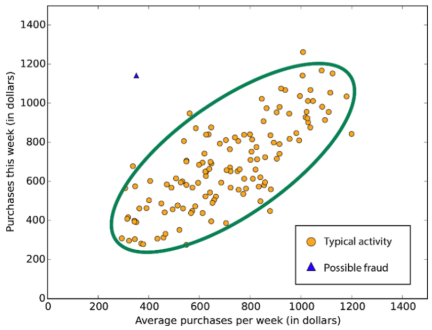

Рис. 5.12: Задача обнаружения аномалий

Как можно подружить эти две задачи? В задаче обнаружения аномалий тоже необходимо построить некоторую разделяющую поверхность, но при этом в наличии нет примеров аномалий, есть только примеры типичных объектов. Можно считать, что в задаче присутствуют 2 класса. Первый класс — это нормальные объекты, к нему относится вся обучающая выборка. Второй класс — это аномалии, и считается, что аномальным является начало координат. Теперь можно решать эту задачу, используя методы классификации классификации.

5.4.2. Применение метода опорных векторов для обнаружения аномалий

Для решения будет использоваться линейный способ, при этом необходимо выбирать разделяющую гиперплоскость так, чтобы она давала максимальный размер зазора, тогда вероятность переобучения будет минимальный. Это лучше всего делать с помощью метода опорных векторов, или SVM. Вот, как выглядит задача отделения выборки от начала координат:

$$\begin{cases} \frac{1}{2}\| \mathrm{w} \|^2 + \frac{1}{v\ell}\sum\limits_{i=1}^{\ell} \xi_i - \rho \to \min\limits_{\mathrm{w},\xi,\rho} \\ \langle \mathrm{w}, x_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \end{cases}$$

Эта задача похожа на обычную задачу метода опорных векторов, но всё-таки отличается. Важный параметр задачи, ν, задаёт верхнюю оценку для доли аномальных объектов в выборке. Если выбрать ν так, что аномальными могут быть объявлены 3 объекта, то разделение может произойти как на рисунке 5.13.

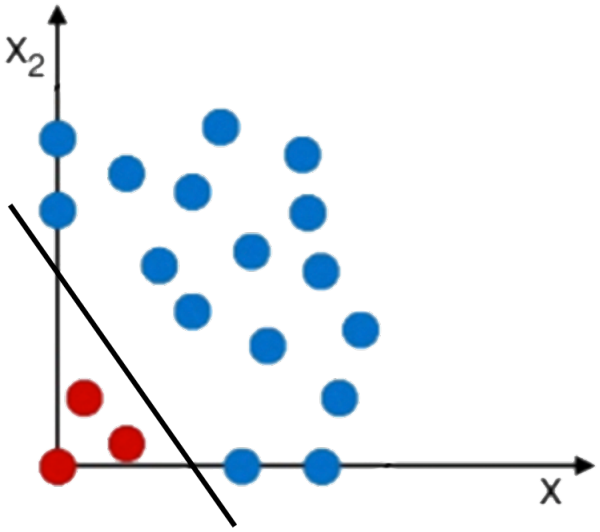

Рис. 5.13: Задача обнаружения аномалий

5.4.3. Классификация с использованием нелинейного ядра

Предположение о том, что 0 — это аномальный объект, очень странное. Например, если выборка центрирована вокруг начала координат, то этот подход в принципе не может дать правильный результат. На самом деле, одноклассовый SVM с линейным ядром никогда не используется. Скалярное произведение в сформулированной задаче можно заменить на ядро K. Популярный выбор — это RBF-ядро, которое вычисляется по формуле

$$K(x, z) = \exp\left(\frac{\|x - z\|^2}{\sigma^2}\right).$$

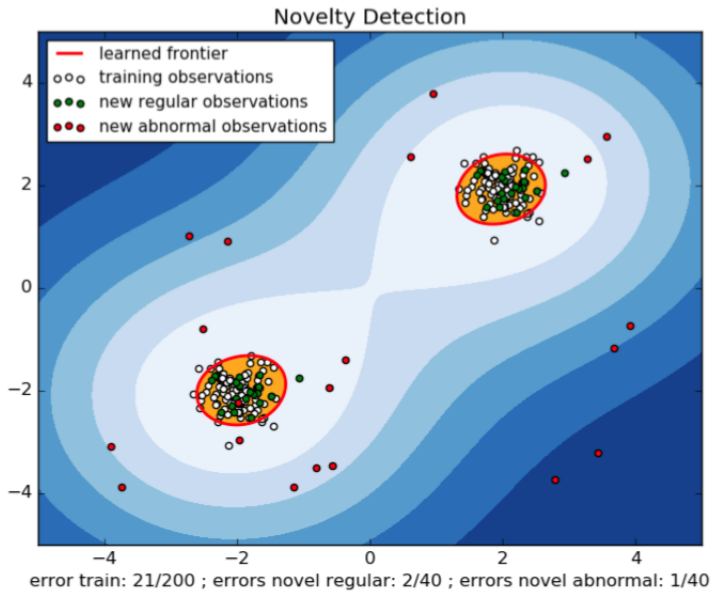

Рис. 5.14: Поиск аномалий с использованием ядрового одноклассового SVM с RBF-ядром

После замены ядра разделяющая плоскость будет строиться в пространстве более высокой размерности. Пример применения ядрового одноклассового SVM с RBF-ядром в выборке показан на рисунке 5.14.