ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Логистическая регрессия

Высшая Школа Цифровой Культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Логистическая регрессия

Этот раздел посвящен еще одному разделу обучения с учителем – задаче классификации. Сегодня, однако, мы не будем рассматривать эту задачу в самой общей постановке, а остановимся на ситуации, когда классов всего два. Несмотря на такое существенное, как может показаться, ограничение, двухклассовая классификация – это очень жизненная задача. Ведь и правда, если на вопрос можно ответить либо да, либо нет, то это ровно-таки она. Достоин ли человек нобелевской премии, кредитоспособен ли заемщик, пойдет ли сегодня дождь? – таких вопросов множество! Так что ограничение на два класса, конечно, существенное, но область применимости все равно широчайшая.

Подход, который мы будем изучать, называется логистической регрессией. Внимательный слушатель, наверное, сразу встанет в ступор: при чем здесь регрессия? Ведь регрессия – это одна задача, а классификация – совсем другая. Регрессия дает число, а классификация – класс. Дело тут вот в чем. Логистическая регрессия выдает вероятность отнесения того или иного объекта к одному из двух классов. Вероятность – это число, поэтому и регрессия, но тут же присутствуют классы – вот и классификация. Запутались? Тогда контрольный. Как думаете, почему нельзя одному классу присвоить число 0, другому число 1, и по аналогии с только что описанным подходом просто воспользоваться уже разобранной нами линейной регрессией?

В общем, вопросов пока что больше, чем ответов. Попробуем разобраться.

1.1 Метод максимального правдоподобия (ММП)

Начнем с метода максимального правдоподобия, который лежит в основе логистической регрессии.

Наводящие соображения

Предположим, что проводится серия из n независимых одинаковых испытаний, вероятность успеха в каждом из которых равна $p \in (0, 1)$. Например, стрельба по мишени или серия послематчевых пенальти. Напомним, что в такой схеме вероятность получить ровно $k \in \{0, 1, .., n\}$ успехов вычисляется по формуле Бернулли:

$$P_n(k) = C_n^k \cdot p^k \cdot (1 - p)^{n-k},$$

где C_n^k – число сочетаний, равное:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Пусть серия состоит из пяти испытаний с вероятностью успеха p равное 0.7 в каждом из них. Известно, что произошло событие, где при первом подходе произошло два успеха, а при втором – четыре. Математики в этом случае говорят, что имеется выборка X_1, X_2, где $X_1 = 2$, $X_2 = 4$.

Какова же вероятность рассматриваемого события? В силу независимости, она равна произведению вероятностей соответствующих событий, каждая их которых вычисляется по формуле Бернулли:

$$\mathsf{P}\left(X_1 = 2, X_2 = 4\right) = P_5(2) \cdot P_5(4) = C_5^2 \cdot 0.7^2 \cdot (1-0.7)^3 \cdot C_5^4 \cdot 0.7^4 \cdot (1-0.7)^1.$$

Проведя вычисления, получаем, что вероятность рассматриваемого события невелика и примерно равна 0.048:

$$\mathsf{P}\left(X_1 = 2, X_2 = 4\right) \approx 0.048.$$

Очевидно, что на результат влияет вероятность успеха в каждом испытании схемы Бернулли, а значит интересно задаться вопросом: при каком параметре p вероятность рассматриваемого эксперимента будет максимальной? Запишем так называемую функцию правдоподобия, заменив известную вероятность 0.7 на неизвестную p:

$$f\left(X_1 = 2, X_2 = 4, p\right) = C_5^2 \cdot p^2 \cdot (1-p)^3 \cdot C_5^4 \cdot p^4 \cdot (1-p)^1.$$

Теперь нам хочется найти такое значение p, при котором эта функция на множестве $[0, 1]$ примет свое наибольшее значение. Легко видеть, что это не 0 и не 1, так как в этих точках функция равна нулю, так что впредь будем считать p от нуля до единицы не включительно.

В итоге мы приходим к классической задаче математического анализа – к задаче нахождения точки, в которой заданная функция принимает наибольшее значение на заданном множестве. Для решения этой задачи можно воспользоваться следующим достаточно вольным алгоритмом.

1. Найти производную первого порядка.

2. Найти точки из области определения функции, в которых производная либо обращается в ноль, либо не существует. Все эти точки называются точками, подозрительными на экстремум.

3. Для проверки того, что точка, подозрительная на экстремум, является точкой локального максимума, воспользоваться каким-нибудь достаточным условием. Например, точка, подозрительная на экстремум будет точкой локального максимума, если при переходе через нее производная меняет знак с плюса на минус.

4. Сравнить значения функции в найденных точках локального максимума со значениями на границах множества (если таковые есть), выбрать наибольшее, а затем определить точку, в которой это наибольшее значение достигается.

Полученная функция $f\left(X_1 = 2, X_2 = 4, p\right)$ представляет собой произведение, что усложняет поиск производной, так как приходится многократно применять правило дифференцирования произведения функций, что, в свою очередь, довольно объемно. Для упрощения вычислений ее можно прологарифмировать и использовать так называемую логарифмическую функцию правдоподобия:

$$L\left(X_1 = 2, X_2 = 4, p\right) = \ln f\left(X_1 = 2, X_2 = 4, p\right).$$

Так как логарифм – монотонная функция, то точки экстремума функции $f\left(X_1 = 2, X_2 = 4, p\right)$ перейдут в точки экстремума функции $L\left(X_1 = 2, X_2 = 4, p\right)$, и наоборот. В итоге, после логарифмирования, произведение распадется в сумму логарифмов:

$$L\left(X_1 = 2, X_2 = 4, p\right) = \ln\left(C_5^2 p^2 (1-p)^3 \cdot C_5^4 p^4 (1-p)^1\right) =$$

$$= \ln C_5^2 + \ln p^2 + \ln(1-p)^3 + \ln C_5^4 + \ln p^4 + \ln(1-p)^1.$$

После чего, используя свойства логарифмов, последнее выражение приводится к упрощенному виду:

$$\ln C_5^2 + \ln C_5^4 + 6\ln p + 4\ln(1-p).$$

Такие преобразования, конечно, приводят к изменению области определения функции, но раз $p \in (0,1)$, преобразование законно. В результате получаем упрощенную логарифмическую функцию правдоподобия:

$$L\left(X_1 = 2, X_2 = 4, p\right) = \ln(C_5^2 \cdot C_5^4) + 6\ln p + 4\ln(1-p).$$

Вычислим производную этой функции по параметру p. Напомним, что $(C)' = 0$, где C – произвольное число, а $(\ln|x|)' = \frac{1}{x}$, в результате чего, получим:

$$\left(L\left(X_1 = 2, X_2 = 4, p\right)\right)'_p = \frac{6}{p} - \frac{4}{1-p}.$$

Тогда, для нахождения точки, подозрительной на экстремум, решим уравнение, приравняв к нулю найденную производную:

$$\frac{6}{p} - \frac{4}{1-p} = 0.$$

Откуда p составит 0.6:

$$6(1-p) = 4p,$$

$$p = 0.6.$$

Кроме того, есть две точки, 0 и 1, когда производная не существует, но они не принадлежат области определения функции и не рассматриваются.

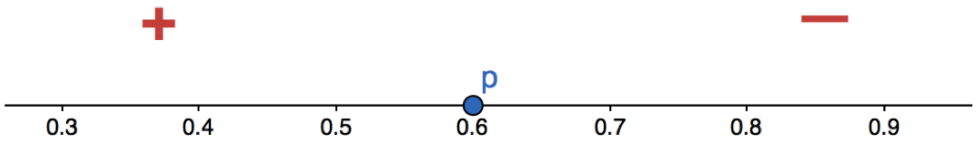

Рис. 1: Интервалы возрастания и убывания функции L.

Проверяем знаки производной функции слева и справа от точки 0.6. Как видим, производная меняет свой знак с плюса на минус, а следовательно p равное 0.6 – точка максимума. Итак, вероятность события, что в первый раз произошло два успеха, а во второй – четыре максимальна при $p = 0.6$.

Давайте найдем наибольшую вероятность рассматриваемого события и подставим найденное значение p в ранее рассмотренный пример, где расчет был выполнен для 0.7:

$$\mathsf{P}\left(X_1 = 2, X_2 = 4\right) = P_5(2) \cdot P_5(4) = C_5^2 \cdot 0.6^2 \cdot (1-0.6)^3 \cdot C_5^4 \cdot 0.6^4 \cdot (1-0.6)^1.$$

Вычисления приводят нас к результату 0.06, что больше прошлого.

$$\mathsf{P}\left(X_1 = 2, X_2 = 4\right) \approx 0.06.$$

Таким образом, мы нашли такое значение p, которое позволяет максимизировать вероятность для конкретных условий, два успеха в первом испытании и четыре во втором.

Сам метод максимального правдоподобия

Перейдем от частного примера, к общим положениям. Предположим, что имеется выборка X объема n, элементы которой $X_1, X_2, \ldots, X_n$ независимы, одинаково распределены и имеют некоторое распределение P_θ, известным образом зависящее от параметра θ. Этот параметр может принимать

значения из какого-то множества Θ, то есть $\theta \in \Theta$. Например, в только что рассмотренном примере семейство распределений – биномиальное распределение $\mathfrak{F}_\theta = \mathsf{Bin}(n, p)$, зависящее от параметра $\theta = p$, причем множество значений параметра Θ – это отрезок $[0, 1]$.

Метод максимального правдоподобия – один из статистических методов, который состоит в построении оценки этого параметра. Грубо говоря, в качестве максимально правдоподобного значения θ берут такое, что при n испытаниях максимизируется вероятность получения данной выборки $X_1, X_2, \ldots, X_n$. В дискретном случае функция правдоподобия $f(X, \theta)$ есть вероятность выборке в данной серии экспериментов равняться $x_1, x_2, \ldots, x_n$:

$$f(X, \theta) = \mathsf{P}_\theta \left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right).$$

Учитывая независимость элементов выборки $X_1, X_2, \ldots, X_n$ мы можем перейти к произведению вероятностей:

$$f(X, \theta) = \mathsf{P}_\theta \left(X_1 = x_1\right) \cdot \mathsf{P}_\theta \left(X_2 = x_2\right) \cdot \ldots \cdot \mathsf{P}_\theta \left(X_n = x_n\right).$$

Оценкой максимального правдоподобия θ^* неизвестного параметра θ называется такое значение, при котором $f(X, \theta)$ достигает максимума, как функция от θ при фиксированных $X_1, X_2, \ldots, X_n$. Математическим языком, задача может быть переписана с использованием оператора $\arg\max$ от рассматриваемой функции f:

$$\theta^* = \arg\max_{\theta \in \Theta} f(X, \theta).$$

Само название оператора говорит, что мы ищем аргумент, максимизирующий функцию, а не значение максимума функции.

Для вычислительных удобств рассматривают не функцию правдоподобия $f(X, \theta)$, а ее логарифм – логарифмическую функцию правдоподобия, что позволяет перейти от произведения вероятностей, к сумме логарифмов от вероятностей:

$$L(X, \theta) = \ln f(X, \theta) = \ln \mathsf{P}_\theta \left(X_1 = x_1\right) + \ldots + \ln \mathsf{P}_\theta \left(X_n = x_n\right).$$

Такой переход возможен, так как логарифм – монотонная функция, а значит точки экстремума функции $f(X, \theta)$ будут и точками экстремума функции $\ln f(X, \theta)$, и наоборот. В итоге, оценкой максимального правдоподобия можно назвать аргумент, максимизирующий логарифмическую функцию правдоподобия:

$$\theta^* = \arg\max_{\theta \in \Theta} L(X, \theta).$$

1.2 Описание логистической регрессии

Рассмотрим логистическую регрессию, которая, в отличие от линейной регрессии, не производит предсказание значения отклика по значению предикторов. Вместо этого она позволяет получить вероятность отнесения объекта к определенному классу, и, как мы уже отмечали в самом начале, решает задачу классификации. Понять это отличие поможет следующий пример.

Пусть предиктор X показывает площадь дома (в квадратных метрах), а отклик Y – стоимость этого дома. Тогда мы можем использовать линейную регрессию для прогнозирования цены продажи в зависимости от размера дома. Зададим теперь несколько иной вопрос. Исходя из заданной площади, будет ли дом продаваться более, чем за 5 млн. рублей? В таком случае возможно два ответа, либо «Да», либо «Нет». И именно для такой задачи подходит логистическая регрессия. Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами.

В общем, мы имеем дело с ситуацией, когда объект может относиться к одному из классов «плюс» или «минус» при разных значениях предикторов $X_1, \ldots, X_p$. Наша цель – по новому, ранее неизвестному набору предикторов определить вероятность отнесения объекта к классу «плюс» или «минус». По сути можно было бы использовать линейную регрессию, но непонятен «вес» классов, так как отклик может быть любым и встает вопрос интерпретации результатов. Логистическая регрессия как раз предлагает подход перевода метки класса в некоторое число.

Итак, нам достаточно научиться определять вероятность $\mathsf{P}_{(+)}$ попадания интересующего нас значения в определенный класс. Тогда попадание в противоположный класс будет иметь вероятность $\mathsf{P}_{(-)} = 1 - \mathsf{P}_{(+)}$. Таким образом, нам нужно разобраться, как значения предикторов можно перевести в вероятности. Давайте это и сделаем.

Рассмотрим величину $\mathsf{P}_+ \in (0, 1)$ – вероятность какого-либо события, к примеру, что сборная России победит в матче против Хорватии. Тогда P_- – вероятность противоположного события, то есть вероятность проигрыша в матче (ничья приравнивается к проигрышу). Перейдем от вероятностной величины к шансу (odds), для этого достаточно разделить вероятность P_+ на P_-:

$$\text{odds}_+ = \frac{\mathsf{P}_+}{\mathsf{P}_-},$$

где плюс обозначает шансы на отнесение к классу $(+)$.

Например, если вероятность выиграть матч $\mathsf{P}_+ = 0.8$, то шанс составит четыре к одному.

$$\text{odds}_+ = \frac{0.8}{1 - 0.8} = 4.$$

Таким образом, в отличие от вероятности, шансы будут принимать любые неотрицательные значения:

$$\text{odds}_+ \in (0, +\infty).$$

Если далее перейти к логарифму шансов, то значения будут принадлежать уже любым вещественным числам:

$$\ln(\text{odds}_+) = \ln\left(\frac{\mathsf{P}_+}{1-\mathsf{P}_+}\right) \in (-\infty, +\infty).$$

При этом логарифм шансов, мы и будем приближать с помощью многомерной линейной регрессии. Для этого приравняем полученное выражение к линейному уравнению многомерной регрессии:

$$\ln\left(\frac{\mathsf{P}_+}{1-\mathsf{P}_+}\right) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p.$$

Для дальнейшего удобства обозначим правую часть выражения за Ψ.

$$\Psi = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p$$

В полученной формуле вероятность P_+ является искомым значением, и его нетрудно выразить явно:

$$e^{\Psi} = \frac{\mathsf{P}_+}{1-\mathsf{P}_+},$$
$$(1-\mathsf{P}_+) \cdot e^{\Psi} = \mathsf{P}_+,$$
$$e^{\Psi} = \mathsf{P}_+ \cdot \left(1 + e^{\Psi}\right).$$

На последнем шаге получаем явную формулу для P_+, однако ее использовать не рекомендуется, так как она может привести к значительной погрешности модели:

$$\mathsf{P}_+ = \frac{e^{\Psi}}{1+e^{\Psi}}.$$

Поэтому продолжим алгебраические преобразования и получим формулу, которую будем использовать далее:

$$\mathsf{P}_+ = \frac{1}{e^{-\Psi} \cdot (1+e^{\Psi})},$$

$$\mathsf{P}_+ = \frac{1}{1+e^{-\Psi}}.$$

Еще раз отметим, что $\mathsf{P}_+ \in (0,1)$, так как $e^x \in (0,+\infty)$, а, следовательно, знаменатель будет строго больше единицы, таким образом правая часть принимает значения из промежутка $(0,1)$. При этом, если записать Ψ, как

$$\Psi = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p,$$

где $X_0 = 1$, то выражение может быть записано как скалярное произведение векторов θ и X:

$$(\theta, X) = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p,$$

1.3 Логистическая функция

Полученная в прошлом фрагменте формула для вероятности P_+ на самом деле может быть соотнесена с функцией, называемой сигмоидой.

$$\mathsf{P}_+ = \frac{1}{1+e^{-\Psi}} \longleftrightarrow \sigma(x) = \frac{1}{1+e^{-x}}$$

Сигмоиду также называют логистической функцией, а ее график вы можете наблюдать на рисунке 2. Простейший анализ (как, в свою очередь, и картинка) показывает, что функция $\sigma(x)$ возрастает, непрерывна и ограничена двумя горизонтальными асимптотами: нулем и единицей.

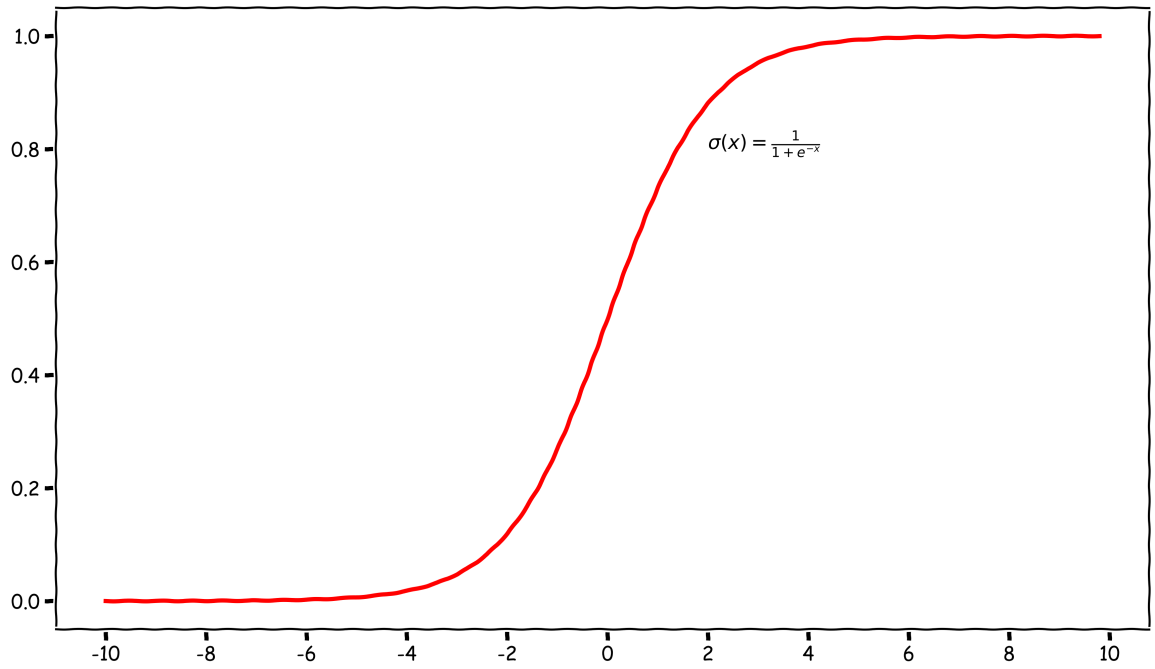

Рис. 2: Логистическая кривая (сигмоида).

Покажем, что функция $\sigma(x)$ возрастает. Действительно, e^{-x} – убывает, $1+e^{-x}$ – убывает, $\frac{1}{x}$ – убывает, а следовательно композиция $\frac{1}{1+e^{-x}}$ – возрастает.

Уравнения асимптот получаются из следующих пределов:

$$\lim_{x\to+\infty}\sigma(x)=1,\quad \lim_{x\to-\infty}\sigma(x)=0,$$

а непрерывность следует из непрерывности композиции элементарных функций.

Это значит, что $\sigma(x)$ является функцией распределения некоторой случайной величины ξ

$$\sigma(x)=\mathsf{P}(\xi<x),$$

а следовательно вероятность P_+ – это вероятность того, что наша величина «ниже» гиперплоскости, так как в качестве аргумента x у нас выступает выражение Ψ:

$$\mathsf{P}_+=\sigma[(\theta,X)]=\sigma\left(\Psi\right)=\mathsf{P}\left(\xi<\Psi\right).$$

Резюмируя, алгоритм прогноза методом логистической регрессии будет состоять из следующих двух шагов (при условии, что параметры $\theta_0,\theta_1,\ldots,\theta_p$ уже получены, то есть модель обучена):

1. Вычислить значение Ψ:

$$\Psi=\theta_0+\theta_1X_1+\theta_2X_2+\ldots+\theta_pX_p.$$

2. Вычислить вероятность P_+:

$$\mathsf{P}_+=\frac{1}{1+e^{-\Psi}}.$$

Рассмотрим пример и приведем крайне упрощенную модель, которая использует три фактора. В качестве данных будут выступать данные статистики футбольного матча (удары в створ ворот, процент владения мяча и удары в сторону ворот), а предсказывать модель будет победу команды. В качестве предикторов выступают три переменные: X_1 – количество ударов в створ ворот; X_2 – процент владения мячом; X_3 – количество ударов в сторону ворот:

$$X_1=\{7,0,3,4,\ldots\},$$

$$X_2=\{40,60,43,57\ldots\},$$

$$X_3=\{13,6,8,14,\ldots\}.$$

Отклик Y представлен двумя значениями, 1 или 0 соответствующих победе или проигрышу команды и классам «плюс» или «минус»:

$$Y=\{1,0,0,1,\ldots\}.$$

На основе данных получена модель с указанным набором значений θ:

$$\theta_0 = -0.046, \theta_1 = 0.541, \theta_2 = -0.014, \theta_3 = -0.132.$$

Далее мы рассмотрим, как вычислить значения этих переменных.

Применим такую модель для данных о команде, которая ударила в створ 1 раз, владела мячом 40 процентов игрового времени и нанесла 3 удара в сторону ворот. Используя формулу, получим вероятность победы команды в матче, которая составит 38% в процентном соотношении:

$$\mathsf{P}_+ = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \cdot 1 + \theta_2 \cdot 40 + \theta_3 \cdot 3)}} \approx 0.38.$$

Остается вопрос обучения модели и нахождения коэффициентов $\theta_0, \theta_1, \ldots, \theta_p$.

1.4 Обучение модели

Давайте зададимся вопросом обучения модели, или вопросом того, как ее «собрать». Вспомним, что логистическая регрессия дает вероятность отнесения объекта к «плюсу» по следующей формуле:

$$\mathsf{P}_+ = \frac{1}{1 + e^{-\Psi}} = \sigma[\Psi],$$

тогда вероятность отнесение к «минусу» это $1 - \mathsf{P}_+$:

$$\mathsf{P}_- = 1 - \mathsf{P}_+ = 1 - \sigma[\Psi].$$

Перепишем последнее соотношение с помощью алгебраических преобразований:

$$1 - \mathsf{P}_+ = 1 - \frac{1}{1 + e^{-\Psi}} = \frac{e^{-\Psi}}{1 + e^{-\Psi}} = \frac{1}{1 + e^{\Psi}}.$$

Таким образом P_- вычисляется через логистическую функцию от аргумента с противоположным знаком:

$$\mathsf{P}_- = \frac{1}{1 + e^{\Psi}} = \sigma[-\Psi].$$

Оба этих выражения можно объединит, если ввести y_i метку ответа, отвечающую за отнесение результата к положительному или отрицательному исходу. Так как имеется всего два класса: «плюс» и «минус», то логично им сопоставить отклики $+1$ и -1. В итоге, отклик $Y = \{+1, -1\}$ принимает значения $y_i = \pm 1$, а выражения для вероятности отнесения объекта к соответствующему классу можно объединить, например, следующим образом:

$$\mathsf{P}_{y_i} = \sigma\left[y_i \cdot \Psi\right].$$

Такой аргумент функции сигма обозначается $M_i(\Psi)$ и называется отступом (margin) классификации. Отступ позволяет судить об уверенности модели в классификации новых объектов. На практике оказывается, что чем меньше значение отступа M_i, тем ближе объект подходит к границе классов и тем выше становится вероятность ошибки.

Вот мы и подошли к завершающему этапу, давайте объединим все полученные шаги в итоговый результат. Поставим задачу поиска параметров $\theta_0, \ldots, \theta_p$, имея p предикторов и n наборов данных $x_{i1}, \ldots, x_{ip}$, где $i = \{1, \ldots, n\}$ с двумя классами $Y = \{+1, -1\}$. Тогда отступ для каждого набора данных i будем обозначать $M_i(\theta, X_i)$. Для нахождения параметров, согласно рассмотренному ранее методу максимального правдоподобия, нужно максимизировать логарифмическую функцию правдоподобия:

$$L(X, \theta) = \sum_{i=1}^{n} \ln\left(\sigma\left[M_i\left(\theta, X_i\right)\right]\right) \to \max.$$

С другой стороны, из свойств логарифмов моментально получается следующая сумма, которую нужно минимизировать, а называется она logloss или логистическая функция потерь.

$$\log\text{loss}(X, \theta) = \sum_{i=1}^{n} \ln\left(1 + e^{-M_i(\theta, X_i)}\right) \to \min.$$

Последнее выражения получается благодаря свойствам логарифма

$$\ln\left(\sigma\left[M_i\left(\theta, X_i\right)\right]\right) = \ln\left(\frac{1}{1 + e^{-M_i(\theta, X_i)}}\right) = -\ln\left(1 + e^{-M_i(\theta, X_i)}\right),$$

таким образом функция правдоподобия примет вид

$$L(X, \theta) = -\sum_{i=1}^{n} \ln\left(1 + e^{-M_i(\theta, X_i)}\right) \to \max.$$

Максимизировать заданную функцию, то же самое, что минимизировать функцию, убрав знаки минус.

Но как найти производную и точки подозрительные на экстремум? Теперь у нас не одна переменная, а $p + 1$ предикторов. Найдем частные производные по $\theta_0, \ldots, \theta_p$ и приравняем каждую к нулю, записав все в систему

уравнений:

$$\begin{cases} \left(\sum\limits_{i=1}^{n} \ln\left(1+e^{-M_i(\theta,X_i)}\right)\right)'_{\theta_0} = 0, \\ \left(\sum\limits_{i=1}^{n} \ln\left(1+e^{-M_i(\theta,X_i)}\right)\right)'_{\theta_1} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ \left(\sum\limits_{i=1}^{n} \ln\left(1+e^{-M_i(\theta,X_i)}\right)\right)'_{\theta_p} = 0. \end{cases}$$

Такую систему уравнений, как правило, решают численно. Среди полученных точек ищут ту или те, которые дают минимум, используя достаточное условие экстремума. Однако в рамках машинного обучения, для решения задачи максимизации или минимизации, можно воспользоваться инструментами моделирования или математическими пакетами. В дополнительных материалах вы сможете ознакомиться с примером и описанием выполнения расчетов.

Рассмотрим последовательность действий, приводящих к численному решению. Для начала рассмотрим урезанные данные по футбольной статистике (с полной таблицей можно ознакомиться в дополнительных материалах к лекции).

Победа или проигрыш	Количество ударов в створ	Процент владения мячом	Удары по воротам
1	7	40	13
0	0	60	6
0	3	43	8
1	4	57	14
...	...	...	...

Составим логистическую функцию потерь, которую впоследствии мы будем минимизировать, для наглядности расчетов возьмем выборку объема 4 из представленной таблицы. Функция потерь будет зависеть от трех предикторов, так как первый столбец является откликом.

$$\log\text{loss}\,(X_1, X_2, X_3, \theta_0, \theta_1, \theta_2, \theta_3) = \sum_{i=1}^{4} \ln\left(1+e^{-M_i(X_1,X_2,X_3,\theta_0,\theta_1,\theta_2,\theta_3)}\right) \to \min\,.$$

Для удобства сперва запишем значения отступов M_i для $i = \{1, 2, 3, 4\}$. Как помните, отступ зависит от значения отклика, в данном примере он принимает значения нуля и единицы, что соответствует проигрышу и победе в матче.

Таким образом индикатор + будет соответствовать единице, а − нулю.

$$
\begin{aligned}
M_1 &= +\left(\theta_0 + 7 \cdot \theta_1 + 40 \cdot \theta_2 + 13 \cdot \theta_3\right), \\
M_2 &= -\left(\theta_0 + 0 \cdot \theta_1 + 60 \cdot \theta_2 + 6 \cdot \theta_3\right), \\
M_3 &= -\left(\theta_0 + 3 \cdot \theta_1 + 43 \cdot \theta_2 + 8 \cdot \theta_3\right), \\
M_4 &= +\left(\theta_0 + 4 \cdot \theta_1 + 57 \cdot \theta_2 + 14 \cdot \theta_3\right).
\end{aligned}
$$

Полученные значения отступов используем в общей формуле и получим сумму четырех логарифмов:

$$
\log \operatorname{loss}\left(X_1, X_2, X_3, \theta_0, \theta_1, \theta_2, \theta_3\right) = \sum_{i=1}^{4} \ln\left(1 + e^{-M_i}\right).
$$

Полученная функция является функцией четырех переменных, следовательно, для поиска точек, подозрительных на экстремум аналитически, необходимо решить систему из четырех уравнений. Однако, как мы уже отметили ранее, значения находятся численно. Пример может быть найден в дополнительных материалах.

Обратимся еще к одному примеру и рассмотрим логистическую регрессию в контексте классификации. Мы уже говорили, что в результате вычислений получается уравнение гиперплоскости $(\theta, X) = 0$, разделяющей объекты на классы. В случае двух измерений — это прямая линия. В случае трех — плоскость. Чтобы классификатор работал, точки, конечно, должны интуитивно разделяться на эти классы, как на рисунке 3.

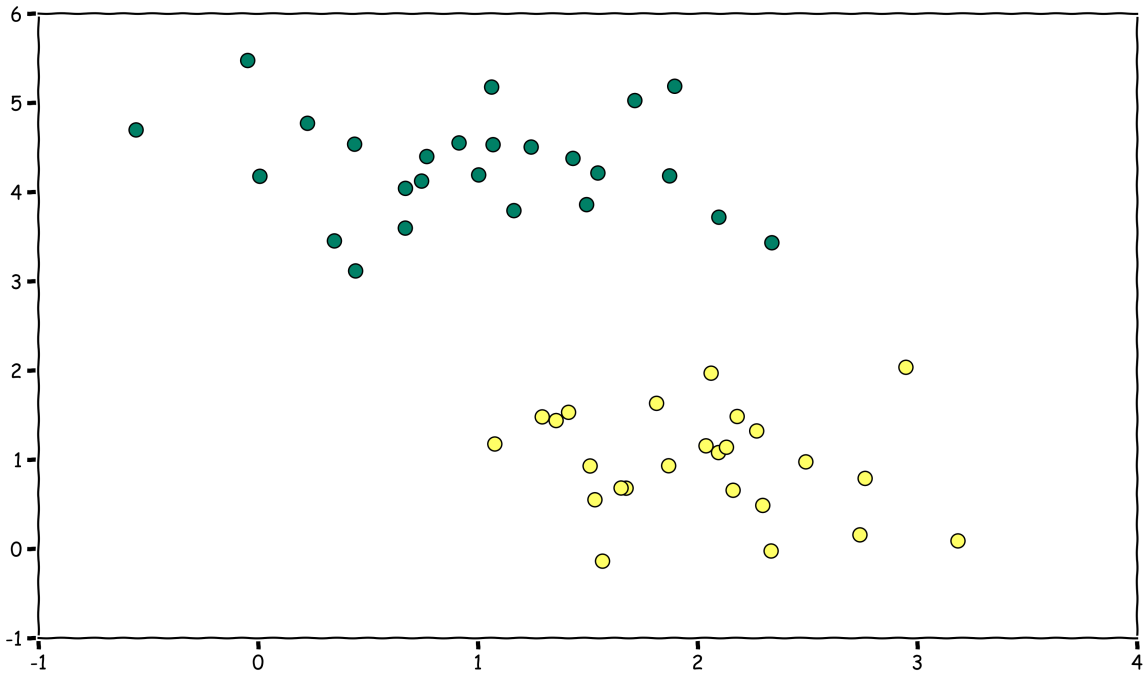

Рис. 3: Объекты интуитивно разделимы.

Моделирование и расчет неизвестных коэффициентов $\theta_0, \theta_1, \ldots, \theta_p$ приводит нас к нахождению явного уравнения гиперплоскости:

$$1.07 + 1.65 \cdot X_1 - 1.55 \cdot X_2 = 0,$$

В данном случае это прямая, которая разделяет плоскость на две области (рисунок 4).

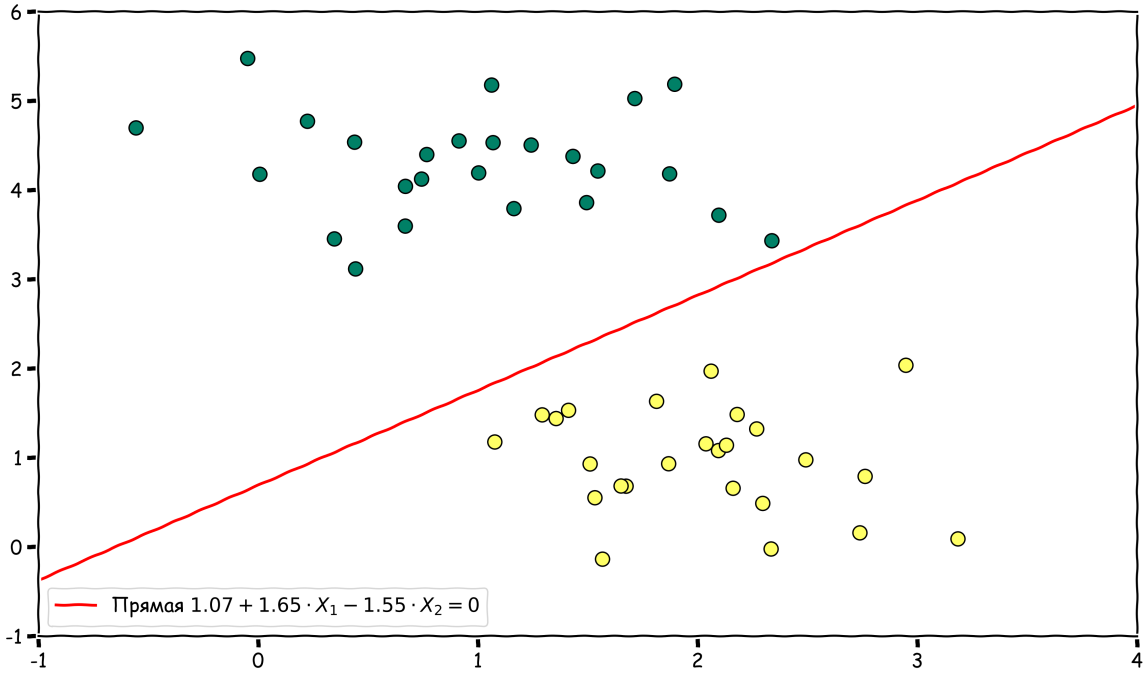

Рис. 4: Разделение объектов.

Теперь мы хотим оценить построенную нами модель. Рассмотрим набор тестовых значений (то есть реальных данных, не участвовавших в обучении) A, B, C, D с откликами из Y_{test}:

$$A = (0, 2), B = (1, 2), C = (3.5, 4), D = (3, 0),$$

$$Y_{test} = \{+1, -1, +1, -1\}.$$

Обратимся к рисунку 5 с отмеченным точками, их цвет соответствует исходному отклику (желтые – классу $+$ или $+1$, а зеленые – классу $-$ или -1). Как видим, объект A классифицирован неправильно, его знак $+$ отличен от класса где он располагается. Как это понять аналитически? Ведь часто данные просто невозможно представить графически, особенно, когда много предикторов.

Помните, мы ввели скалярное произведение (θ, X)? Если приравнять его к нулю, то мы получаем наше заветное уравнение гиперплоскости. То есть если выражение равно нулю, значит объект X лежит на гиперплоскости (в

нашем примере на прямой). Но что если мы не получим ноль? Как теперь понять, к какой части пространства, или к какому из двух классов относится наше наблюдение?

На самом деле все довольно просто. В выражении (θ, X), коэффициенты $\theta_1, \theta_2, \ldots, \theta_p$ – координаты нормального вектора гиперплоскости $(\theta, X) = 0$. Подзабыли геометрию? Не беда! Нормальный вектор – это вектор, перпендикулярный к гиперплоскости. Оказывается, что его направление указывает на класс «плюс», именно по этой причине не стоит использовать слова «выше» и «ниже» для обозначения классов «плюс» и «минус».

Помните, как мы определили вероятность отнесения объекта к классу +? Мы использовали сигмоиду, а именно выражение показанное на экране

$$\mathsf{P}_+ = \frac{1}{1 + e^{-(\theta, X)}}.$$

Если $(\theta, X) = 0$, то вероятность равна 0.5, а классификатор не знает, куда отнести объект, и именно поэтому он, объект, лежит на гиперплоскости. В обучающих данных мы четко знаем, какие объекты отнесены к классу +, а какие к −, поэтому логично, что в случае линейно разделимой выборки для объектов класса + должно выполняться $\mathsf{P}_+ > 0.5$ (иначе логичнее их отнести к классу −). Это означает, исходя из соотношения для P_+, что $(\theta, X) > 0$, если вектор нормали и объект находятся в одном полупространтсве.

Но как отсюда следует требуемое нами направление? Объясним идею, допустив, что $\theta_0 = 0$, в другом случае все похоже, но не хотим вас отягощать :) Так вот, так как (θ, X) в случае $\theta_0 = 0$ и есть скалярное произведение нормального вектора с координатами $\theta_1, \theta_2, \ldots, \theta_p$ на вектор X с координатами $X_1, X_2, \ldots, X_p$, векторы указывают в одно полупространство тогда и только тогда, когда скалярное произведение положительно. Почему? Да потому, что скалярное произведение можно найти как произведение длин векторов на косинус угла между ними, где первые две величины будут положительны, а значит знак зависит от косинуса, а точней угла. И если угол острый, получим знак плюс, а это происходит тогда и только тогда, когда векторы направлены в одно и то же полупространство. А если угол тупой, то знак минус и векторы будут направлены в разные полупространства.

В рассматриваемом примере, вектор имеет координаты $\theta = (1.65, -1.55)$, а следовательно, направлен в нижнюю область, там где находятся желтые точки, а они как раз из класса «плюс».

Вернемся к объекту A, и проверим, что он классифицирован неправильно, только теперь не на глаз, а аналитически, ведь на практике построить такие красивые картинки не выйдет, пространства будут многомерные.

Итак, подставим координаты точки A в уравнение прямой и получим

значение -2.03:

$$1.07 + 1.65 \cdot 0 - 1.55 \cdot 2 = -2.03.$$

О чем нам говорит отрицательное значение? О том, что классификатор относит эту точку в класс $-$.

Объекты B и C находятся близко к прямой, таким образом нельзя полностью доверять результату, однако знак скалярного произведения совпадает с классом и объекты классифицированы верно. Аналогично это можно проверить аналитически:

$$1.07 + 1.65 \cdot 1 - 1.55 \cdot 2 = -0.38,$$

$$1.07 + 1.65 \cdot 3.5 - 1.55 \cdot 4 = 0.645,$$

Объект D находится на значительном расстоянии от прямой, обычно такие объекты должны совпадать по знаку, но в данном примере знак противоположный, что, вероятно, говорит об аномалии или попросту об ошибке метки класса.

$$1.07 + 1.65 \cdot 3 - 1.55 \cdot 0 = 6.02.$$

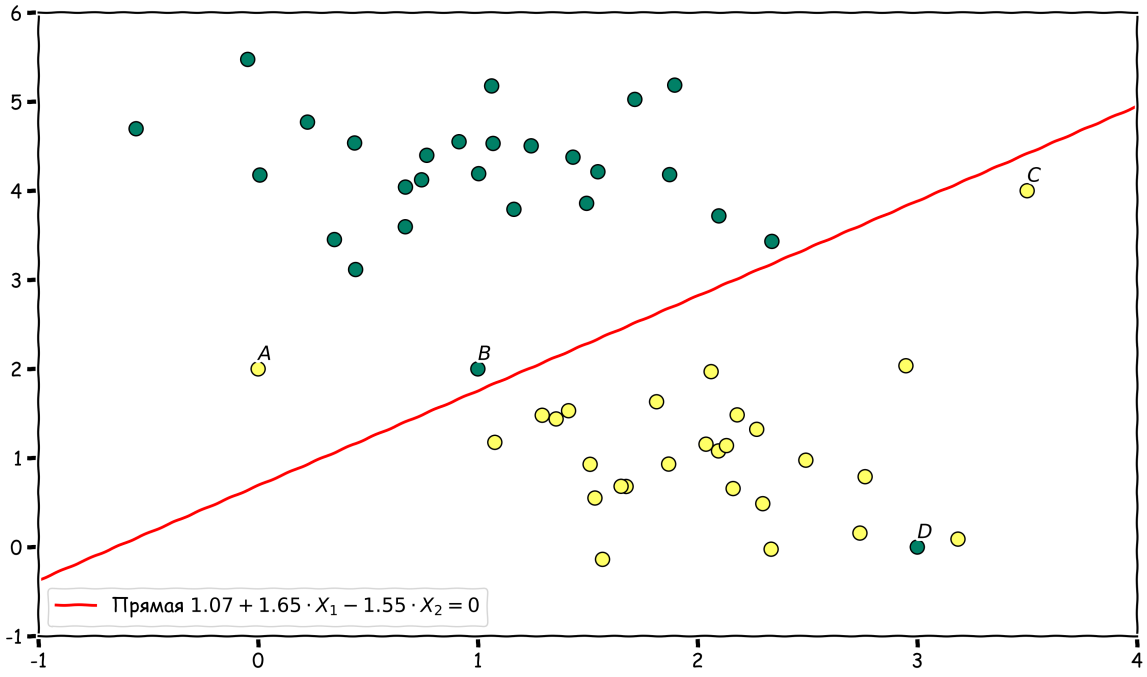

Рис. 5: Классификация новых объектов.

1.5 Сравнение линейной и логистической регрессий

Давайте посмотрим на существенную разницу в подходах линейной и логистической регрессии. Для этого возьмем в качестве данных, например, уровень дохода и факт получения кредита. На рисунке 6 представлены наши обучающие данные. По горизонтальной оси отложен доход в тысячах рублей в месяц, а по вертикальной – факт получения кредита. Единица означает, что кредит получен, ноль, что не получен. Желтые точки отображают людей, которые получили кредит, а зеленые – нет, таким образом люди с одинаковым доходом не всегда получают одинаковый результат. В нашей модели два параметра θ_0 и θ_1.

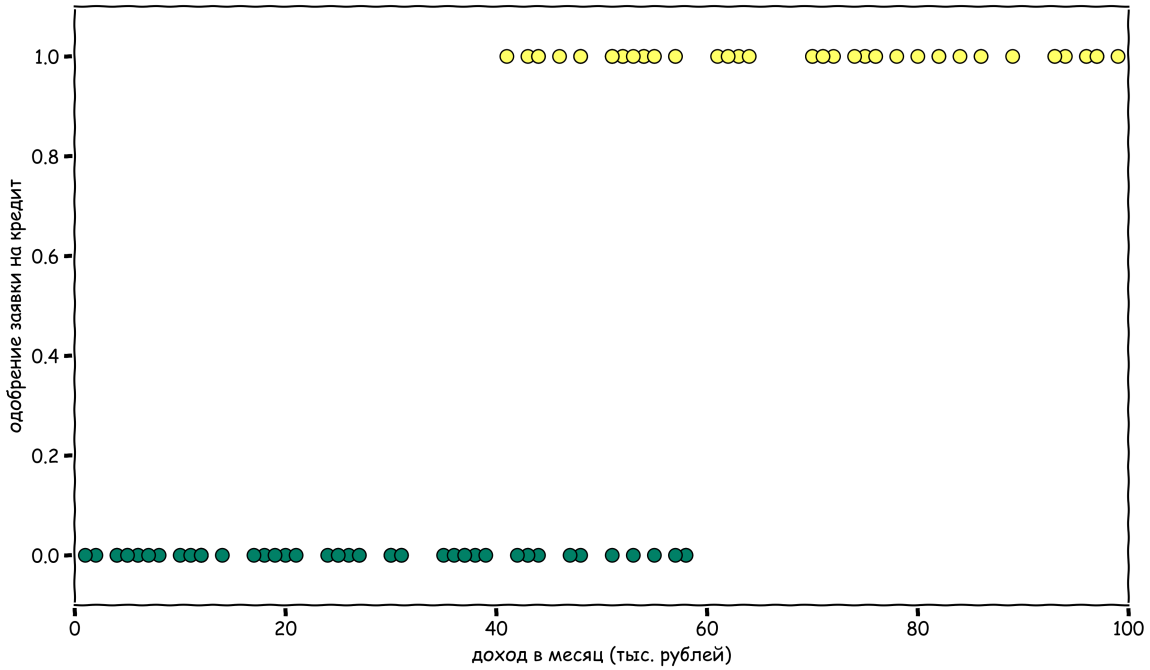

Рис. 6: Обучающие данные по выдаче кредитов.

В результате расчетов логистической модели получена следующая сигмоида, которая изображена красным цветом на рисунке 7. Синим же цветом, на рисунке 8 нарисована линейная регрессия.

Возьмем набор новых клиентов с доходом 35, 40, 60, 70, 80 тысяч рублей и вычислим вероятности получения кредита. Они составят $0.11, 0.22, 0.89, 0.98, 0.99$ для логистической регрессии. Линейная регрессия дает вероятности $0.32, 0.39, 0.67, 0.82, 0.96$.

Какие выводы можно сделать? Да, в указанном примере классификация прошла одинаково с точки зрения конечного результата, то есть обе модели одинаково указали на клиентов, которым кредит, скорее всего, не дадут (вероятность меньше 0.5), и которым кредит одобрят. Однако можно наблю-

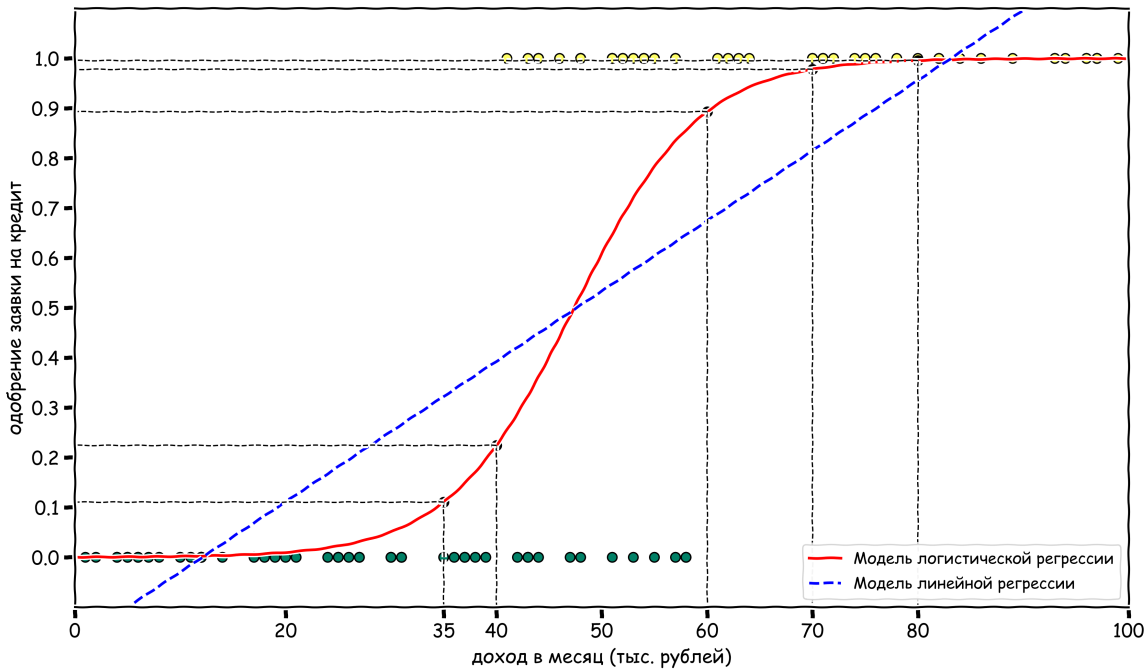

Рис. 7: Результаты логистической регрессии.

дать следующее: при средних значениях предикторов (в примере в районе 50 тыс.руб.) модели ведут себя по-разному. Так, в логистической регрессии видно сильное изменение значений вероятности за счет изменения выпуклости функции и ее стремительного роста при «средних» значениях дохода, тогда как в линейной значения вероятностей изменяются с одинаковой скоростью на всем диапазоне. С увеличением числа предикторов (например, с добавление возраста, пола, наличия недвижимости), эти нюансы моделей будут играть значительный эффект в классификации.

Кроме того видно, что линейная регрессия даже на тренировочных данных дает странные результаты. Так, есть люди, для которых она выдает «результат», меньший нуля, а есть – для которых выдает «результат» больший, чем единица. И как это интерпретировать? Этот пример еще раз иллюстрирует одну из проблем использования линейной регрессии в этой задаче: потеря нормировки.

1.6 ROC-анализ

ROC-кривая или кривая ошибок – кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении. Кривая ошибок показывает зависимость доли истинно положительных примеров от доли ложно положительных примеров. При этом предполагается, что у классификатора имеется некоторый параметр, изменение которого влияет на то или иное разбиение на эти два класса, положительный и отрицательный. Давайте разберемся с тем, как найти эти доли.

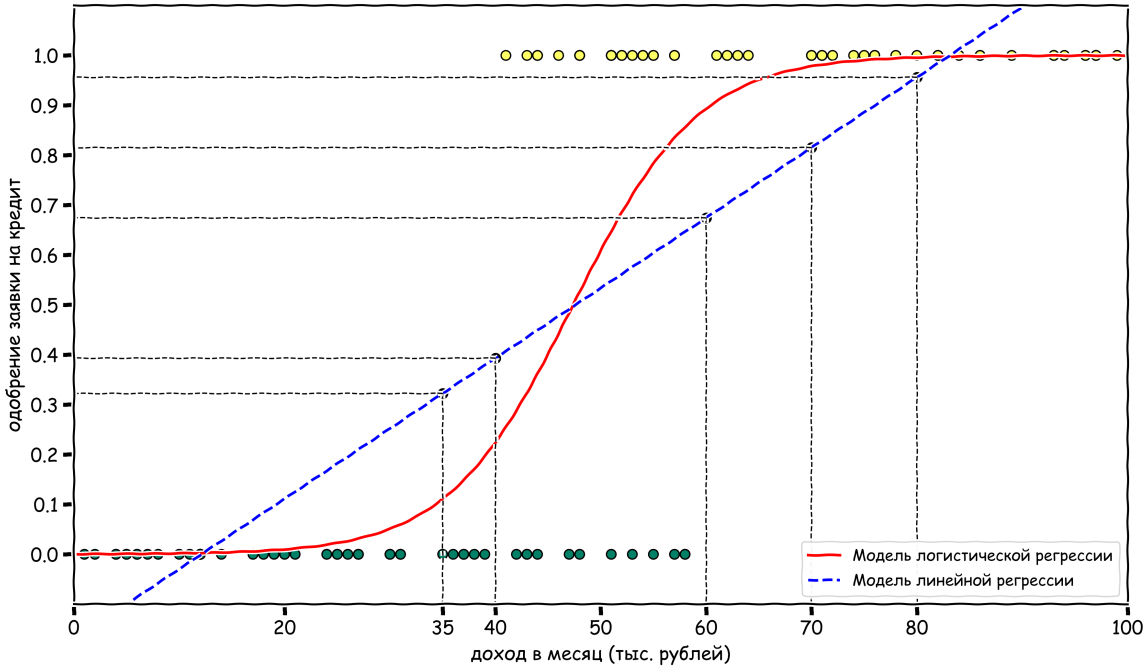

Рис. 8: Результаты линейной регрессии.

Для этого проанализируем полученные результаты классификации моделью и объективные, фактические данные и составим следующую матрицу ошибок (confusion matrix) на основе подсчитанных величин:

- TP (True Positives) – верно классифицированные положительные примеры;
- TN (True Negatives) – верно классифицированные отрицательные примеры;
- FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода);
- FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода).

Матрица ошибок		Верный класс	
		+	−
Прогноз	+	TP	FP
	−	FN	TN

При этом чаще оперируют не абсолютными показателями, а относительными – долями (rates) выраженными числами от нуля до единицы:

- доля истинно положительных примеров обозначается как TPR (True Positives Rate):
$$TPR = \frac{TP}{TP + FN};$$
- доля ложно положительных примеров обозначается как FPR (False Positives Rate):
$$FPR = \frac{FP}{TN + FP}.$$

Введем еще два понятия, которыми определяется объективная ценность любого бинарного классификатора:

- Чувствительность (Sensitivity) – отражает долю положительных результатов, которые правильно идентифицированы. Эта характеристика равна доле истинно положительных примеров:
$$S_e = TPR;$$
- Специфичность (Specificity) – отражает долю отрицательных результатов, которые правильно идентифицированы как таковые:
$$S_p = 1 - FPR = \frac{TN}{TN + FP}.$$

Данные понятия очень хорошо иллюстрирует такой пример из медицины. Чувствительный диагностический тест – это тот, который правильно идентифицирует пациентов с заболеванием, то есть, если тест на 100% чувствителен, то он определит всех пациентов, у которых есть заболевание. Однако он может записать к заболевшим и тех кто не болен. Высокочувствительный тест полезен для исключения заболевания, если человек имеет отрицательный результат.

Специфичный диагностический тест диагностирует только доподлинно больных, то есть, если тест имеет 100% специфичность, он будет идентифицировать 100% пациентов, которые не имеют заболевания, но может записать к здоровым и больных. Такой тест важен при лечении пациентов с определенным заболеванием.

Кроме указанных критериев оценки, использует еще два понятия, чаще встречающиеся в различных инструментах и библиотеках программирования – это точность (precision) и полнота (recall).

Precision – это доля объектов, действительно являющихся положительными к тем, что названы положительными в результате классификации. На

основании значений матрицы ошибок точность можно вычислить следующим образом:

$$Precision = \frac{TP}{TP + FP}.$$

Recall – характеризует долю объектов, классифицированных как положительные к тем, что действительно являются положительными и полностью соответствует значению TPR:

$$Recall = \frac{TP}{TP + FN}.$$

Вернемся к примеру по футбольной статистике, где мы уже обучили модель и нашли значения θ:

$$\theta_0 = -0.046, \theta_1 = 0.541, \theta_2 = -0.014, \theta_3 = -0.132.$$

Используем тестовый набор данных, представленный в таблице ниже и составим матрицу ошибок.

Победа или проигрыш	Количество ударов в створ	Процент владения мячом	Удары по воротам
1	5	60	10
1	2	35	3
0	3	45	6
0	1	53	10
1	7	70	11
1	3	65	12
1	1	30	2
0	2	40	9
1	10	71	15
1	6	54	12
0	7	65	15
0	0	30	3

Для начала найдем вероятность победы для каждого набора данных. Так для команды попавшей в створ ворот в 5 из 10 ударов и владевшей мячом 60 процентов игрового времени, вероятность составит 0.588:

$$\mathsf{P}_+ = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \cdot 5 + \theta_2 \cdot 60 + \theta_3 \cdot 10)}} \approx 0.588.$$

Но какой класс мы присвоим этому результату? Если порог отсечения 0.5 то победа, а если 0.6 то проигрыш. Найдем оставшиеся вероятности победы

в матче для всех тестовых данных, вычисления для которых аналогичны. Поместим округленные результаты в таблицу. На практике значения лучше не округлять:

Вероятность победы
0.588
0.520
0.517
0.186
0.743
0.285
0.446
0.337
0.886
0.671
0.666
0.307

В качестве порога выберем значение 0.5, и назначим классы.

Вероятность победы	Победа или проигрыш (предсказание)
0.588	1
0.520	1
0.517	1
0.186	0
0.743	1
0.285	0
0.446	0
0.337	0
0.886	1
0.671	1
0.666	1
0.307	0

Легко сопоставить предсказанные классы с исходными. Как видим, они не всегда одинаковы, а это значит, что модель ошибается.

Победа или проигрыш	Вероятность победы	Победа или проигрыш (предсказание)
1	0.588	1
1	0.520	1
0	0.517	1
0	0.186	0
1	0.743	1
1	0.285	0
1	0.446	0
0	0.337	0
1	0.886	1
1	0.671	1
0	0.666	1
0	0.307	0

Теперь мы можем составить матрицу ошибок. Подсчитаем число команд, которые победили в матче как по исходным данным, так и по прогнозу модели, таких команд пять. Аналогично подсчитаем число команд проигравших в матче как по исходным данным, так и по прогнозу, их три. Ошибок первого и второго рода по две. Тогда доля истинно положительных примеров составит

Матрица ошибок		Исходный класс	
		+	−
Прогноз	+	TP=5	FP=2
	−	FN=2	TN=3

около 0.71:

$$TPR = \frac{TP}{TP + FN} = \frac{5}{5 + 2} \approx 0.71,$$

а доля ложно положительных 0.4:

$$FPR = \frac{FP}{TN + FP} = \frac{2}{3 + 2} = 0.4.$$

Мы получили значения метрик для конкретного порога отсечения, но как же построить ROC-кривую зависимости TPR от FPR? Для этого надо повторить все вычисления для разных значений порога отсечения.

Существуют различные подходы. Можно менять значение порога от нуля до единицы с некоторым шагом, а можно в качестве значений взять полученные по модели вероятности, отсортированные по возрастанию. Занятие это монотонное и долгое, так что обычно выполняется автоматизированными инструментами. В результате, каждой паре значений FPR_i, TPR_i соответсвует точка на плоскости, которые соединяются прямыми. При этом при переходе

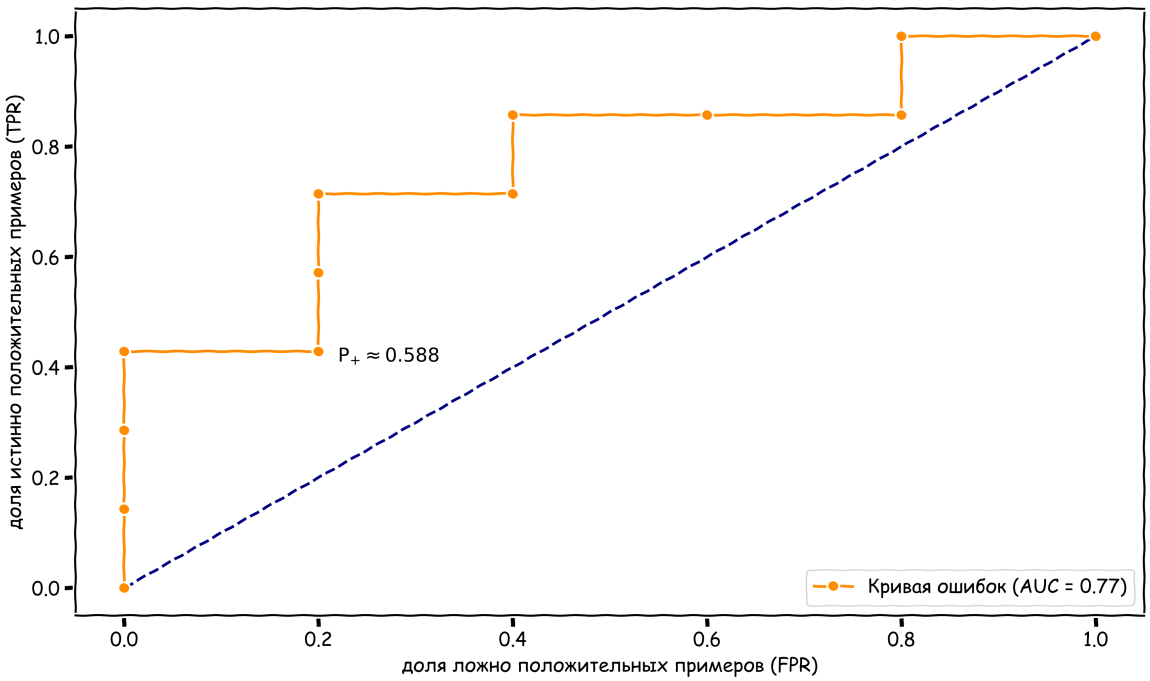

Рис. 9: Кривая ошибок.

через следующее значение вероятности меняется только одно значение классификации, в следствии чего скачок происходит либо вверх, либо вправо, а фигура будет ступенчатой (рисунок 9). Точки с координатами $(0, 0)$ и $(1, 1)$ всегда отмечаются и являются началом и концом кривой. Например, значению вероятности 0.588, соответствует доля истинно положительных примеров 0.2 и доля ложно положительных примеров 0.429.

В идеальной вселенной, кривая проходит через верхний левый угол, где процент истинно положительных случаев составляет 100%. Поэтому, чем больше выгнута ROC-кривая, тем более точным является прогнозирование результатов модели. Оценка модели может быть получена непосредственно вычислением площади под ROC-кривой. Показатель обозначается как AUC (Area Under Curve – площадь под кривой). Получившаяся фигура является ступенчатой, а следовательно, ее площадь может быть легко вычислена как сумма площадей прямоугольников на основе значений TPR_i, FPR_i.

Значения AUC на практике соотносят со следующим качеством модели: Отличное $\in (0.9, 1]$, Хорошее $\in (0.7, 0.9]$, Среднее $\in (0.60.7]$, Неудовлетворительное $\in (0.5, 0.6]$.

Как же настроить модель, чтобы она обеспечивала оптимальное соотношение чувствительности и специфичности? Как найти заветное значение порога отсечения?

Оптимальным значением порога, будет точка пересечения графика чувствительности и специфичности. График строится аналогично ROC-кривой, только мы строим на одной плоскости зависимость чувствительности от поро-

га, и специфичности от порога (рисунок 10). Для рассматриваемого примера это значение составит около 0.52. Еще раз отметим, что на практике значения вероятностей и прочих величин не стоит округлять.

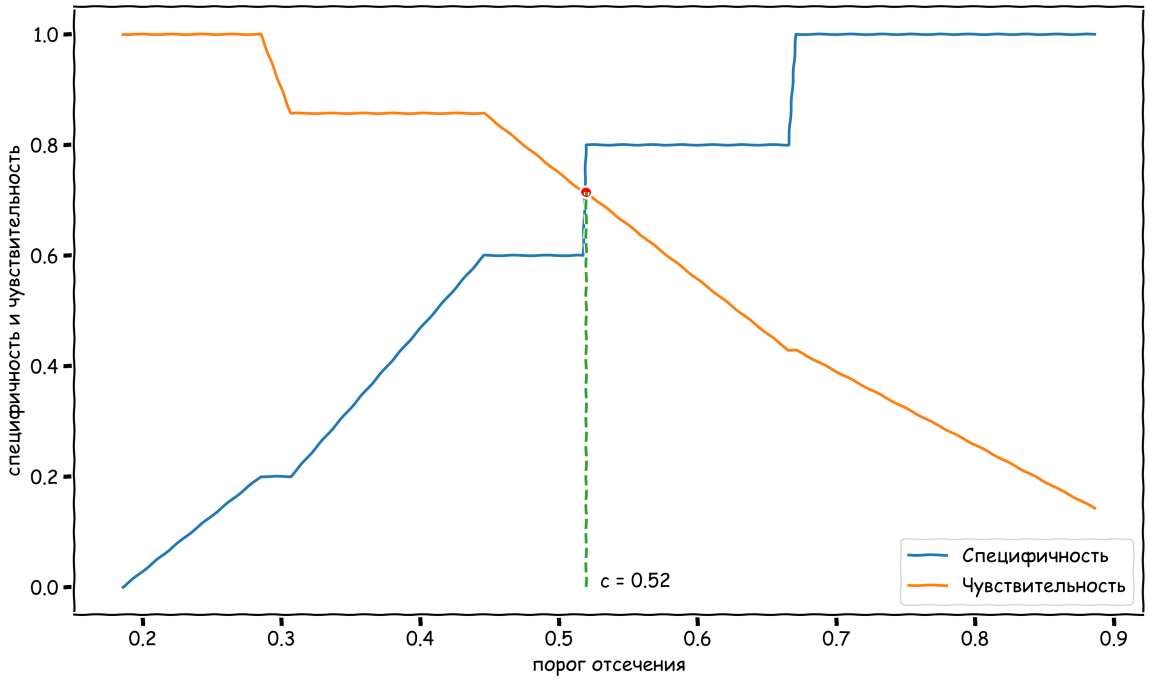

Рис. 10: Определение порога отсечения.

Порог мало отличается от рассмотренного ранее, да и объем данных невелик, однако матрица ошибок уже будет другой:

Матрица ошибок		Исходный класс	
		+	−
Прогноз	+	TP=4	FP=1
	−	FN=3	TN=4

При таком пороге отсечения доля истинно положительных примеров составит 0.571 и доля ложно положительных примеров 0.2. То есть модель стала лучше отсекать отрицательные данные за счет увеличения ее специфичности.

Таким образом, цель ROC-анализа заключается в том, чтобы подобрать такое значение точки отсечения, которое позволит модели с наибольшей точностью распознавать положительные или отрицательные исходы и выдавать наименьшее количество ложноположительных или ложноотрицательных ошибок.