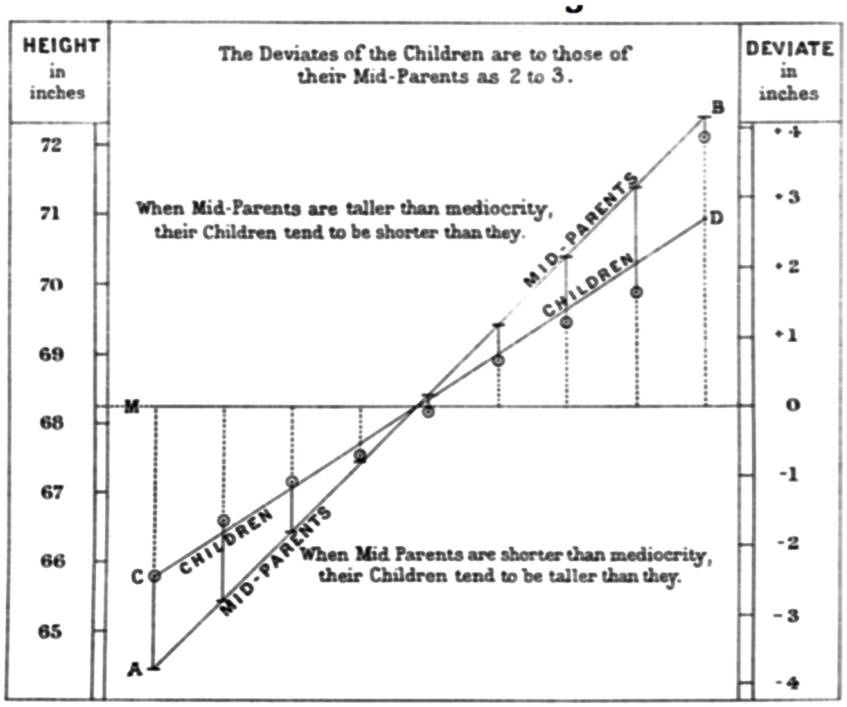
Урок 5

Линейные модели: статистический взгляд

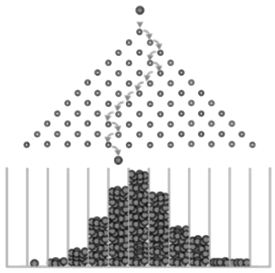
5.1. Задача регрессии

5.1.1. История термина

Впервые термин регрессия появился в конце XIX века в работе Френсиса Гальтона:



В этой работе «Регрессия к середине в наследственности роста» Френсис Гальтон исследовал зависимость между средним ростом детей и средним ростом их родителей и обнаружил, что отклонение роста детей от среднего составляет примерно 2/3 отклонения роста родителей от среднего. Казалось бы, со временем люди должны рождаться все ближе и ближе к среднему росту. На самом деле, естественно, этого не происходит.



Лучше понять эффект регрессии к среднему позволяет другое творение Френсиса Гальтона, которое называется машина, или доска, Гальтона.

Это механическая машина, в которой сверху в центральной части находятся шарики. Когда открывается заслонка, шарики начинают постепенно сыпаться вниз, ударяясь о штырьки, которые расположены на одинаковом расстоянии друг от друга. При каждом соударении шарика со штырьком вероятности того, что он упадет налево и направо от штырька, равны. Постепенно шарики начинают собираться в секциях внизу в гауссиану, или плотность нормального распределения.

Чтобы понять эффект регрессии к среднему, давайте мысленно подставим к машине Гальтона снизу еще одну такую же машину. Если теперь убрать перегородку, которая удерживает шарики в верхней половине, они начнут постепенно осыпаться вниз и сформируют внизу еще одну такую же гауссиану. Если зафиксировать какой-то конкретный шарик в нижней половине ближе к краю, то откажется, что с достаточно большой вероятностью этот шарик пришел не из ячейки, которая находится в верхней половине прямо над ячейкой, в которой он оказался внизу, а от ячейки ближе к середине. Это происходит просто потому, что в середине шариков больше.

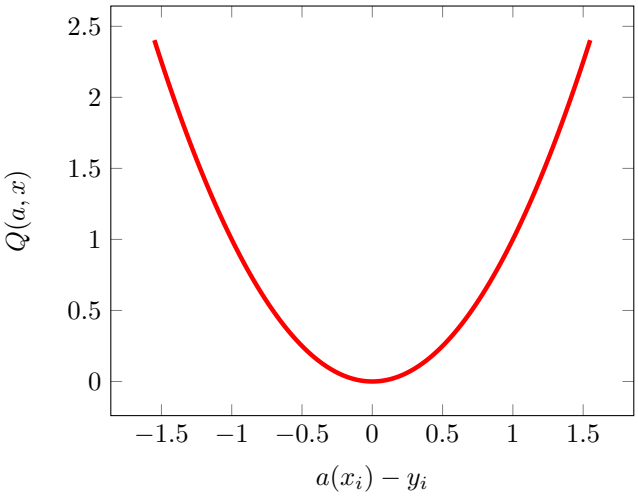
Эффект регрессии к среднему проявляется во многих практических задачах. Например, если дать студентам очень сложный тест, большую роль в том, насколько хорошо они его пройдут, будут играть не только их знания по предмету, но и везение, то есть случайный фактор. Поэтому, если изолировать 10% студентов, которые прошли тест лучше всех (набрали больше всего баллов) и дать им еще один вариант теста, то средний балл в этой группе скорее всего упадет. Просто потому что люди, которым повезло в первый раз, скорее всего уже не будут так удачливы во второй — в этом и состоит эффект регрессии к середине.

Френсис Гальтон был основоположником дактилоскопии, исследовал явление синестезии, внес существенный вклад в метеорологию, впервые описав циклоны и антициклоны, а также, например, изобрел ультразвуковой свисток для собак. Но именно регрессия и по сей день остается одним из наиболее важнейших инструментов, к которому он приложил руку.

5.1.2. Регрессия

Чаще всего под регрессией понимают минимизацию среднеквадратичной ошибки: квадратов отклонений откликов y от их предсказанных значений a(x).

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$



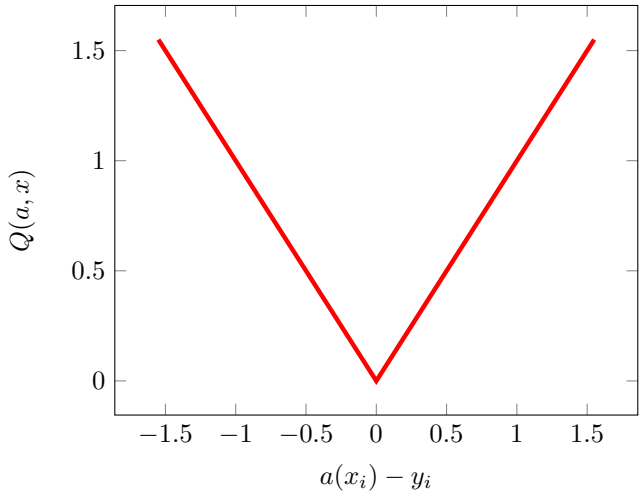
Поскольку минимизируется сумма квадратов отклонений, этот метод называется методом наименьших квадратов (сокращенно МНК). Для линейной регрессии задача имеет аналитическое решение.

$$w_*(x) = \operatorname{argmin}_w Q(w, X) = (X^T X)^{-1} X^T y.$$

Именно этим частично объясняется популярность среднеквадратичной ошибки.

В XIX веке, когда эта задача впервые возникла, никакого способа ее решения, кроме аналитического, быть не могло. Сейчас можно численно минимизировать не только среднеквадратичную ошибку, но и, например, среднюю абсолютную, то есть сумму модулей отклонений нашей модели от отклика:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|, \qquad a_*(x) = \operatorname{argmin}_a Q(a, X).$$



Такая задача является частным случаем класса задач квантильной регрессии.

5.2. Метод максимизации правдоподобия

Пусть x — некоторая случайная величина с функцией распределения $F(x, \theta)$, которая зависит от неизвестного параметра θ, а $X^n = (X_1, ..., X_n)$ — выборка размера n, сгенерированная из распределения $F(x, \theta)$. Необходимо оценить по данной выборке неизвестный параметр.

5.2.1. Метод максимизации правдоподобия: пример

Чтобы понять метод максимального правдоподобия, можно рассмотреть еще один исторический пример. Эти данные собраны в конце XIX века. В Генеральный штаб прусской армии ежегодно в течение 20 лет от десяти кавалерийских корпусов поступали данные о количестве смертей кавалеристов в результате гибели под ними коня.

Кол-во погибших	0	1	2	3	4	5	Всего
Кол-во донесений	109	65	22	3	1	0	200

Поскольку данная случайная величина — счетчик, ее необходимо моделировать распределением Пуассона, функция вероятности которого имеет вид:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Поскольку выборка состоит из независимых, одинаково распределенных случайных величин, вероятность получения строго определенной выборки равна произведению вероятностей получения каждого из элементов этой выборки:

$$P(X^n, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda).$$

Функция L зависит от неизвестного параметра λ и называется правдоподобием выборки. В качестве оценки для λ можно взять такое значение, которое максимизирует функцию правдоподобия:

$$\hat{\lambda}_{\text{ОМП}} = \text{argmax}_{\lambda}\, L(X^N, \lambda)$$

Эта оценка называется оценкой максимального правдоподобия.

Несложно показать, что в рассматриваемой задаче оценка максимального правдоподобия для параметра λ совпадает с выборочным средним:

$$\hat{\lambda}_{\text{ОМП}} = \bar{X}_n = 0.61$$

Чтобы это сделать, можно взять производную от логарифма функции правдоподобия и приравнять ее к нулю. Здесь используется тот факт, что логарифмирование L не меняет точку максимума функции.

5.2.2. Метод максимизации правдоподобия: общий вид

В общем виде метод максимума правдоподобия записывается следующим образом. Пусть некоторая случайная величина x имеет распределение $F(x, \theta)$, X^n — выборка размера n:

$$X \sim F(x, \theta), \quad X^n = (X_1, ..., X_n),$$

Тогда функция правдоподобия имеет вид:

$$L(X^n, \lambda) = \prod_{i=1}^{n} P(X = X_i, \theta).$$

Поскольку при логарифмировании не меняются положения максимумов функции, удобно работать не с самим правдоподобием, а с логарифмом правдоподобия:

$$\ln L(X^n, \lambda) = \sum_{i=1}^{n} \ln P(X = X_i, \theta).$$

Оценкой максимального правдоподобия называется величина:

$$\hat{\lambda}_{\text{ОМП}} = \text{argmax}_{\lambda} \ln L(X^N, \lambda)$$

В случае непрерывной случайной величины метод максимального правдоподобия записывается аналогично:

$$X \sim F(x, \theta), \qquad L(X^n, \lambda) = \prod_{i=1}^{n} f(X = X_i, \theta), \qquad \hat{\lambda}_{\text{ОМП}} = \text{argmax}_{\lambda}\, L(X^N, \lambda).$$

5.2.3. Свойства метода максимального правдоподобия

Метод максимального правдоподобия обладает рядом очень полезных свойств:

- Состоятельность, то есть получаемые оценки при увеличении объема выборки начинают стремиться к истинным значениям:
$$\hat{\lambda}_{\text{ОМП}} \to \theta \qquad \text{при} \qquad n \to \infty.$$
- Асимптотическая нормальность, то есть с ростом объема выборки, оценки максимального правдоподобия все лучше описываются нормальным распределением с средним, равным истинному значению θ, и дисперсией, равной величине, обратной к информации Фишера:
$$\hat{\lambda}_{\text{ОМП}} \sim N\left(\theta, I^{-1}(\theta)\right) \qquad \text{при} \qquad n \to \infty.$$

5.3. Регрессия как максимизация правдоподобия

5.3.1. Модель шума: нормальное распределение

При решении задачи регрессии значение отклика приближается в виде

$$y = a(x) + \varepsilon,$$

где $a(x)$ — регрессионная функция, а компонента ε описывает случайный шум.

Если этот случайный шум имеет нормальное распределение с нулевым средним и дисперсией σ^2, оказывается, что задача минимизации среднеквадратичной ошибки

$$a_* = \operatorname{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

дает оценку максимального правдоподобия для регрессионной функции $a(x)$.

Этот факт позволяет использовать в задаче с регрессии свойства метода максимального правдоподобия. Например, используя асимптотическую нормальность, можно определять значимость признаков x^j в модели и делать их отбор. Также можно строить доверительные интервалы для значения отклика на новых объектах, которых нет в обучающей выборке.

5.3.2. Модель шума: распределение Лапласа

Распределение шума не обязательно должно быть нормальным и может быть каким-то другим. Например, можно попытаться описать его распределением Лапласа с нулевым средним. Формула для функции плотности вероятности такого распределения:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}.$$

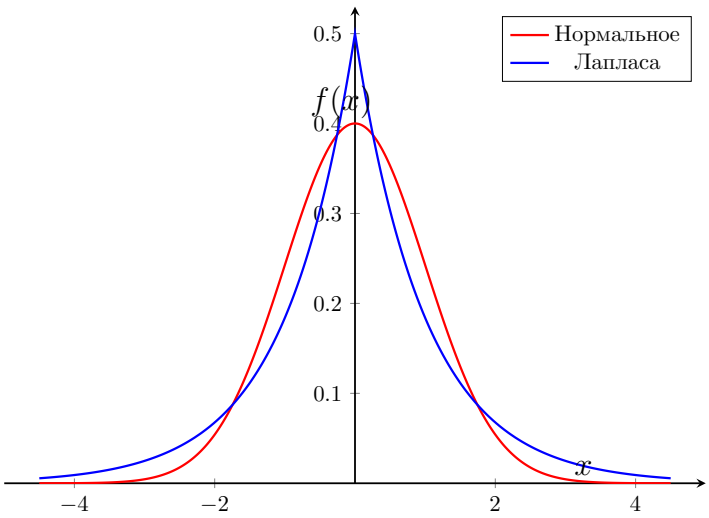


Рис. 5.1: График функции Лапласа с нулевым средним и нормального распределения.

По сравнению с нормальным распределением, распределение Лапласа имеет более тяжелые хвосты, то есть для него более вероятны большие значения ε. Другими словами, если моделировать шум распределением Лапласа, то наблюдения могут сильнее отклоняться от выбранной модели. За счет этого получается решение, которое более устойчивое к выбросам. Оказывается, что если шум действительно описывается распределением Лапласа, то к оценке максимального правдоподобия приводит минимизация средних абсолютных отклонений:

$$a_* = \operatorname{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|.$$

5.4. Регрессия как оценка среднего

5.4.1. Среднеквадратичная ошибка

Пусть для начала a — константа (что соответствует ситуации, когда отсутствуют признаки), а y является случайной функцией с плотностью распределения $f(t)$. В таком случае среднеквадратичная ошибка имеет вид:

$$Q(a) = \int_t (a - t)^2 f(t) dt.$$

Нетрудно показать, что:

$$a_* = \operatorname{argmin}_a Q(a) = \mathbb{E}y,$$

то есть наилучшая константа, которая аппроксимирует значение y в смысле среднеквадратичной ошибки — это математическое ожидание.

Если $a(x)$ — произвольная функция признаков x, функционал среднеквадратичной ошибки имеет вид:

$$Q(a, X) = \int_t (a(x) - t)^2 f(t) dt,$$

а его минимум будет доставлять условное математическое ожидание:

$$a_*(x) = \operatorname{argmin}_a Q\big(a(x)\big) = \mathbb{E}(y|x).$$

Таким образом, в случае с конечной выборкой:

$$Q(a(x), X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(a(x_i) - y_i\right)^2,$$

оценка, получаемая при минимизации среднеквадратичной ошибки:

$$a_*(x) = \operatorname{argmin}_a Q(a, X)$$

является лучше аппроксимацией условного математического ожидания $\mathbb{E}(y|x)$. В случае линейной регрессии, то есть когда отклик моделируется линейной комбинацией $\langle w, x_i \rangle$:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2, \qquad w_* = \operatorname{argmin}_w Q(w, X),$$

выражение $\langle w_*, x_i \rangle$ является наилучшей линейной аппроксимацией условного математического ожидания $\mathbb{E}(y|x)$.

Полученный результат согласуется с интуитивными представлениями. Действительно, пусть $y_i = 2$, график зависимости ошибки на этом объекте в зависимости от предсказания алгоритма $a(x)$ выглядит следующим образом.

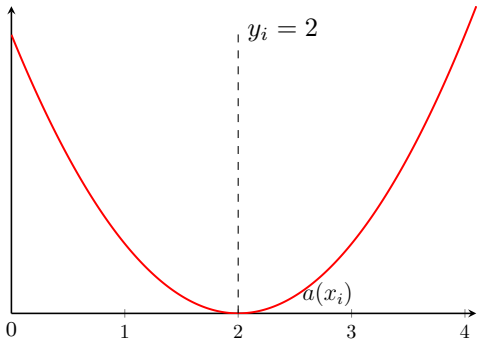


Рис. 5.2: Зависимость ошибки от предсказания алгоритма в случае среднеквадратичной ошибки.

По графику видно, что одинаково штрафуются отклонения предсказания как в большую, так и в меньшую сторону от истинного значения y_i. Поэтому не удивительно, что функция, которая доставляет минимум функции, представляет собой какое-то среднее.

5.4.2. Дивергенции Брегмана

Однако оказывается, что условное математическое ожидание доставляет минимум не только среднеквадратичной ошибки, но и более широкого класса функций потерь, которые называются дивергенциями Брегмана.

Дивергенции Брегмана порождаются любой непрерывной дифференцируемой выпуклой функцией φ:

$$Q(a, X) = \varphi(y) - \varphi(a(X)) - \varphi'(a(X))(y - a(X)).$$

Среднеквадратичная ошибка является частным случаем дивергенции Брегмана. Минимизируя любую дивергенцию Брегмана, мы получаем оценку для условного математического ожидания:

$$a_* = \operatorname{argmin}_a Q(a, X) \text{ — лучшая аппроксимация } \mathbb{E}(y|x).$$

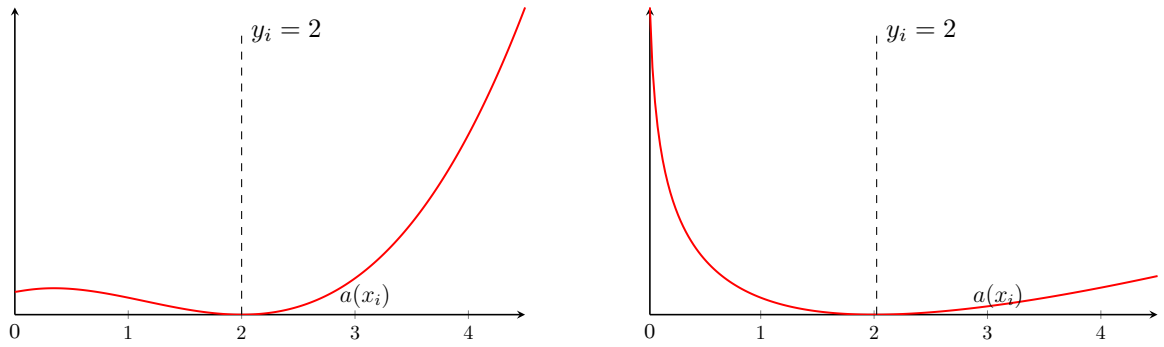


Рис. 5.3: Несколько функции потерь из класса дивергенций Брегмана.

Этот результат уже является несколько странным, поскольку в семействе дивергенций Брегмана можно найти, в том числе, несимметричные относительно y функции. Такие функции больше штрафуют за отклонение модели в большую или меньшую сторону. Это результат может быть несколько контринтуитивным и получен не так давно.

5.4.3. Средняя абсолютная ошибка и несимметричная абсолютная ошибка

Средняя абсолютная ошибка:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

не входит в семейство дивергенций Брегмана. При ее минимизации получается оценка не условного математического ожидания, а оценка условной медианы:

$$a_* = \operatorname{argmin}_a Q(a, X) \text{ — лучшая аппроксимация } \operatorname{med}(y|x).$$

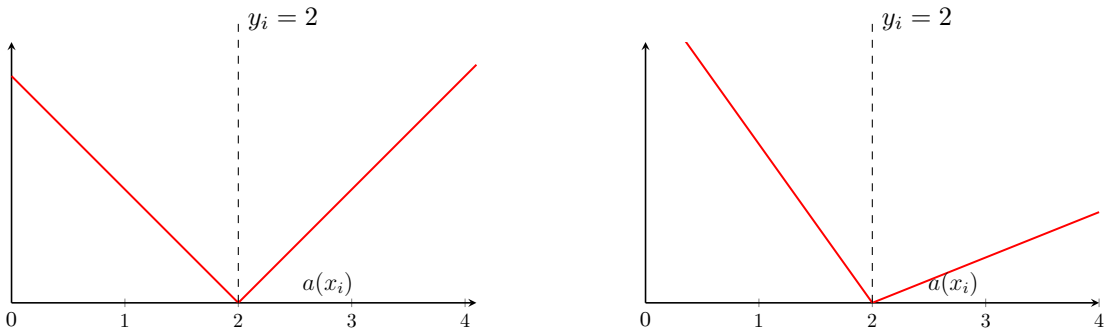


Рис. 5.4: Графики симметричной средней абсолютной и несимметричной абсолютной функций ошибок.

Несимметричная абсолютная функция ошибки («несимметричность» определяется параметром τ):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \big((\tau - 1)\big[y_i < a(x_i)\big] + \tau\big[y_i \geq a(x_i)\big]\big)\,(y_i - a(x_i))$$

имеет в некотором смысле наклоненный график по сравнению с графиком симметричной абсолютной ошибки. При минимизации такого функционала получается лучшая оценка для соответствующего условного квантиля:

$$a_* = \operatorname{argmin}_a Q(a, X) \text{ — лучшая аппроксимация } y|x \text{ порядка } \tau.$$

5.5. Регуляризация

5.5.1. Переобучение регрессионных моделей

Если используется слишком сложная модель, а данных недостаточно, чтобы точно определить ее параметры, эта модель легко может получиться переобученной, то есть хорошо описывать обучающую выборку и плохо — тестовую. Бороться с этим можно различными способами:

- **Взять больше данных.** Такой вариант обычно недоступен, поскольку дополнительные данные стоят дополнительных денег, а также иногда недоступны совсем. Например, в задачах веб-поиска, несмотря на наличие терабайтов данных, эффективный объем выборки, описывающей персонализированные данные, существенно ограничен: в этом случае можно использовать только историю посещений данного пользователя.
- **Выбрать более простую модель** или упростить модель, например исключив из рассмотрения некоторые признаки. Процесс отбора признаков представляет собой нетривиальную задачу. В частности, не понятно, какой из двух похожих признаков следует оставлять, если признаки сильно зашумлены.
- **Использовать регуляризацию.** Ранее было показано, что у переобученной линейной модели значения весов в модели становятся огромными и разными по знаку. Если ограничить значения весов модели, то с переобучением можно до какой-то степени побороться.

5.5.2. L_1-регуляризация и L_2-регуляризация

Есть несколько способов провести регуляризацию:

- L_2-регуляризатор (ridge-регрессия или гребневая регрессия):

$$w_* = \operatorname{argmin}_w \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right).$$

- L_1-регуляризатор (lasso-регрессия или лассо-регрессия):

$$w_* = \operatorname{argmin}_w \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} |w_j| \right).$$

Понять различия между L_1 и L_2 регулязаторами можно на модельном примере. Пусть матрица «объекты–признаки» X является единичной матрицей размера $\ell \times \ell$:

$$X = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}.$$

Тогда при решении задачи линейной регрессии использование метода наименьших квадратов без регуляризации:

$$w_* = \operatorname{argmin}_w \sum_{i=1}^{\ell} (w_i - y_i)^2,$$

дает следующий вектор весов:

$$w_{*j} = y_j$$

При использовании гребневой регуляризации (L_2–регуляризация) компоненты вектора весов имеют вид:

$$w_{*j} = \frac{y_j}{1+\lambda},$$

а при использовании L_1–регуляризатора (lasso):

$$w_{*j} = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \leq \lambda/2. \end{cases}$$

При использовании только МНК без регуляризации $w_{*j} = y_j$. Соответствующая линия изображена пунктиром на обоих графиках. При использовании L_2 регуляризации зависимость w_{*j} от y_j все еще линейная, компоненты вектора весов ближе расположены к нулю.

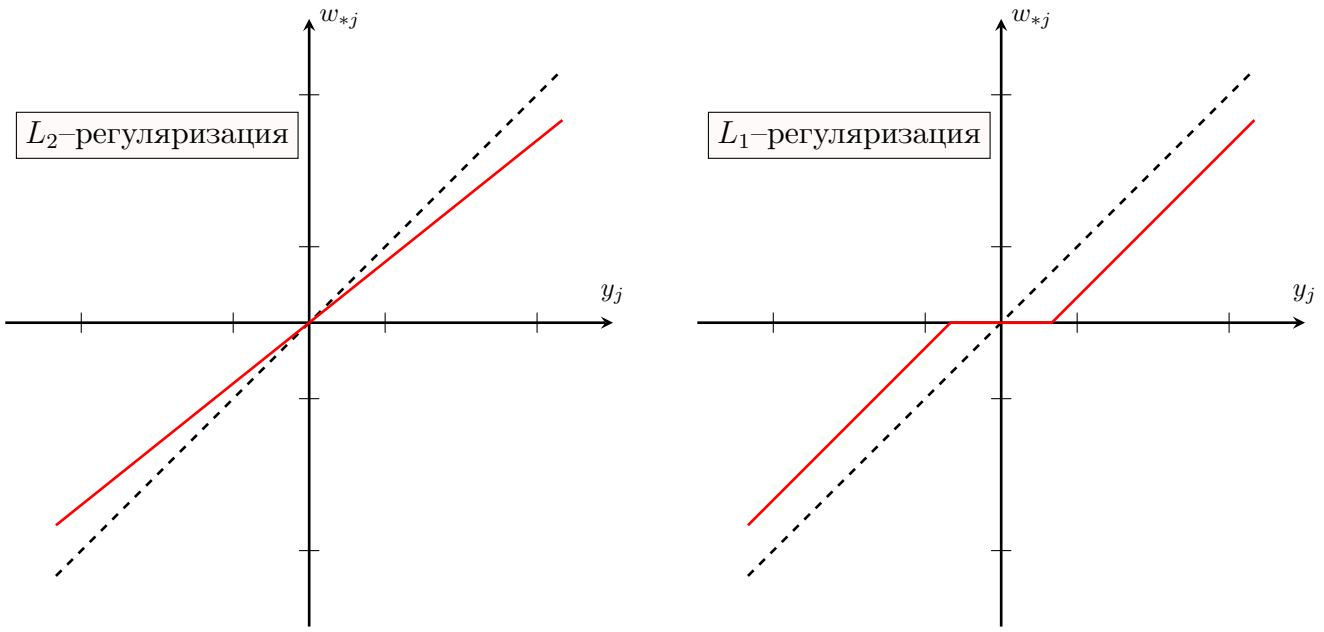


Рис. 5.5: Зависимость w_{*j} от значения отклика y_j при использовании различных регуляризаторов.

В случае L_1 регуляризации график выглядит несколько иначе: существует область (размера λ) значений y_j, для которых $w_j = 0$. То есть lasso, или L_1–регуляризация, позволяет отбирать признаки, а именно: веса признаков, обладающих низкой предсказательной способностью, оказываются равными нулю.

5.5.3. Смещение и дисперсия

Можно показать, что математическое ожидание квадрата ошибки регрессии представляет собой сумму трех компонент:

$$\mathbb{E}\left(a_*(x) - y\right)^2 = \underbrace{\left(\mathbb{E}a_*(x) - a(x)\right)^2}_{\text{Квадрат смещения}} + \underbrace{\mathrm{D}a_*(x)}_{\text{Дисперсия оценки}} + \underbrace{\sigma^2}_{\text{Шум}}.$$

От выбора модели зависит квадрат смещения и дисперсия оценки, но не шум, который является свойством данных, а не модели.

Метод наименьших квадратов дает оценки, которые имеют нулевое смещение. Регуляризация позволяет получить смещенные оценки с меньшим $\mathbb{E}\left(a_*(x) - y\right)^2$ за счет того, что у этой оценки будет меньше дисперсия.

Следующая аналогия позволяет лучше понять баланс между смещением и дисперсией. При стрельбе по мишени среднее число набранных очков зависит от положения средней точки попадания и разбросом относительно этого среднего.

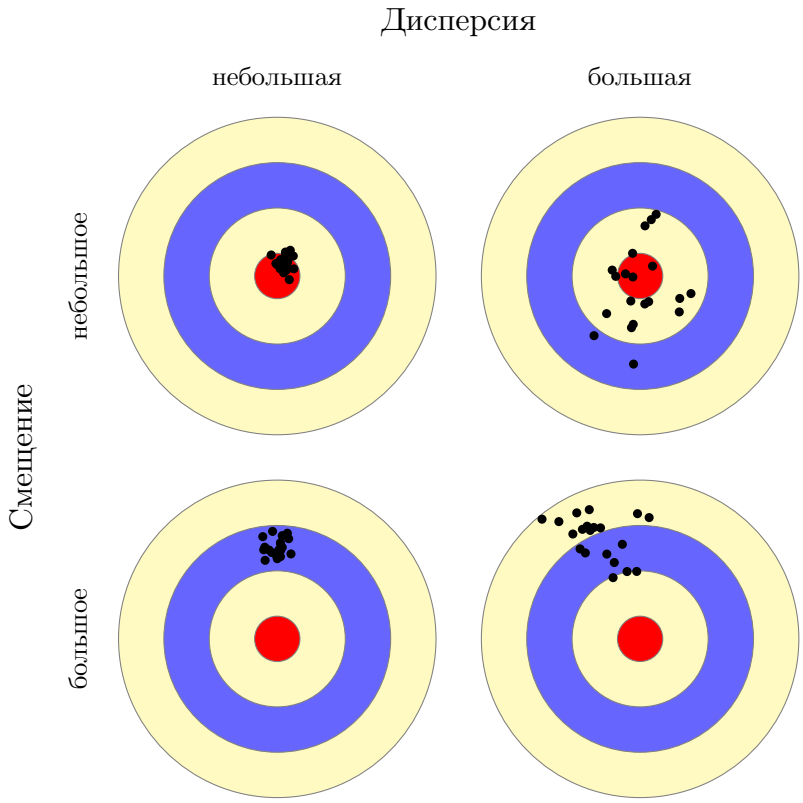


Рис. 5.6: Дисперсия и смещения при различных характерах стрельбы.

Лучший результат будет, если стрелять без смещения и без разброса. Переобучению линейных моделей соответствует стрельба без смещения, но с огромным разбросом. И часто оказывается, что можно набрать больше очков, стреляя не совсем в цель, то есть со смещением, но зато более точно. Именно это и позволяет добиться регуляризация.

5.5.4. Решение задач гребневой регрессии и лассо

В байесовской статистике гребневая регрессия соответствует заданию нормального априорного распределения на коэффициенты линейной модели, а метод лассо — заданию Лапласовского априорного распределения. Подробнее о байесовской статистике написано в соответствующем уроке.

Задача гребневой регрессии имеет аналитическое решение:

$$w_* = \left(X^T X + \lambda I\right)^{-1} X^T y.$$

Для решения задачи лассо аналитического решения не существует, однако есть очень эффективный численный способ получения решения.

5.6. Логистическая регрессия

5.6.1. Логистическая регрессия

Пусть $\mathbb{X}$ — пространство объектов, $\mathbb{Y}$ — пространство ответов, $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка, $x = (x^1, ..., x^d)$ — признаковое описание.

Логистическая регрессия – это метод обучения с учителем в задаче бинарной классификации $\mathbb{Y} = \{0, 1\}$.

5.6.2. Метод линейного дискриминанта Фишера

Метод линейного дискриминанта Фишера, один из самых старых методов классификации, заключается в минимизации среднеквадратичной ошибки:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - y_i \right)^2 .$$

В результате получается вектор весов:

$$w_* = \operatorname{argmin}_a Q(w, X),$$

причем, если для некоторого объекта $\langle w, x_i \rangle > 0.5$, объект относится к первому классу $y = 1$, в ином случае — к нулевому $y = 0$. На самом деле, хочется предсказывать не просто метки классов, а вероятности того, что объекты относятся к какому-то из классов:

$$P(y = 1|x) \equiv \pi(x).$$

Хотя $\pi(x)$ совпадает с условным математическим ожиданием $\mathbb{E}(y|x)$:

$$\pi(x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) = \mathbb{E}(y|x),$$

использовать для оценки вероятности обычную линейную регрессию

$$\pi(x) \approx \langle w, x \rangle$$

не получится: получаемая линейная комбинация факторов не обязательно лежит на отрезке от 0 до 1.

Пусть, например, решается следующая задача. Необходимо предсказать вероятность невозврата платежа по кредитной карте в зависимости от размера задолженности.

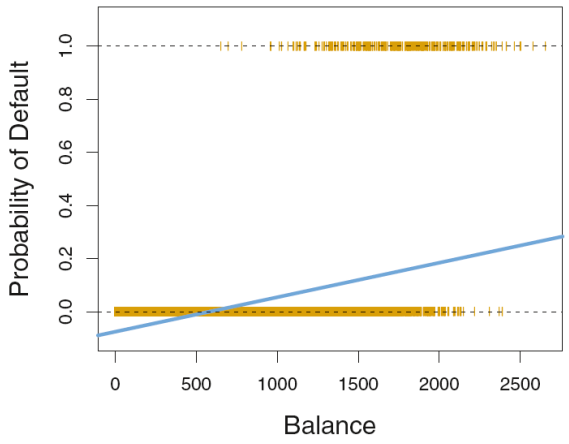


Рис. 5.7: Применение линейной регрессии в задаче оценки вероятности просрочки платежа.

По обучающей выборке была настроена модель линейной регрессии. Получается, что при задолженности 2000$ вероятность просрочить платеж по кредиту равна 0.2, при задолженности 500$ — нулю, а при меньших значениях и вовсе отрицательная. Также, если задолженность больше 10000$, вероятность просрочки будет больше 1. Не понятно, как интерпретировать этот результат.

5.6.3. Обобщенные линейные модели

Пусть функция $g : (0, 1) \mapsto \mathbb{R}$ переводит интервал $(0, 1)$ на множество всех действительных чисел, тогда можно решать задачу линейой регрессии:

$$g\left(\mathbb{E}(y|x)\right) \approx \langle w, x \rangle,$$

в которой строится оценка не для условного матожидания $\mathbb{E}(y|x)$, а для $g\left(\mathbb{E}(y|x)\right)$. Что то же самое:

$$\mathbb{E}(y|x) \approx g^{-1}\left(\langle w, x \rangle\right)$$

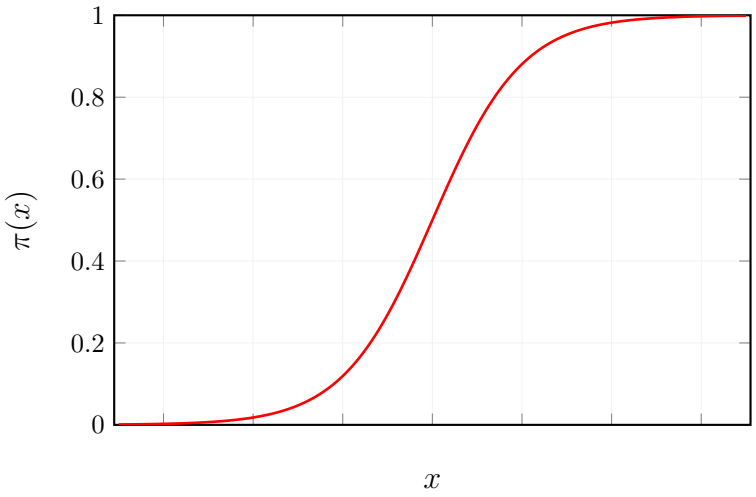
В статистике такое семейство моделей называется обобщенными линейными моделями.

В задаче бинарной классификации в качестве g^{-1} используется сигмоида:

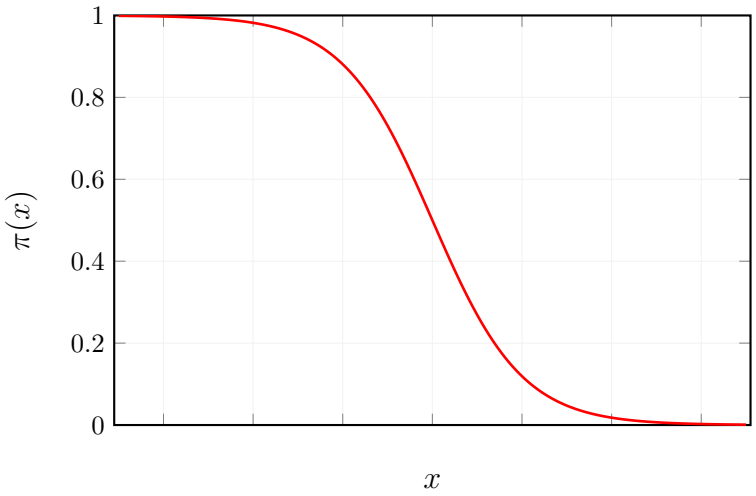
$$\pi(x) \approx \frac{e^{\langle w,x\rangle}}{1+e^{\langle w,x\rangle}}.$$

В одномерном случае значение параметр w_0 сигмоиды определяет положение её центра на числовой оси, а w_1 — форму это сигмоиды:

- Если $w_1 > 0$, сигмойда возрастающая:



- Если $w_1 < 0$, сигмоида убывающая:



Чем больше по модулю значение w_1, тем круче наклон сигмоиды в области ее середины.

5.6.4. Предсказание вероятностей

Если использовать сигмоиду:

$$\pi(x) \approx \frac{e^{\langle w,x\rangle}}{1+e^{\langle w,x\rangle}}$$

в обобщенной линейной модели в задаче логистической регрессии, результат будет более адекватным:

- Вероятность $\pi(x) \in [0,1]$, как и требуется.

- На краях области значений x функция (вероятность) $\pi(x)$ слабо меняется при небольших изменениях x, когда как существенно изменяется, если x находится в середине диапазона своих значений.

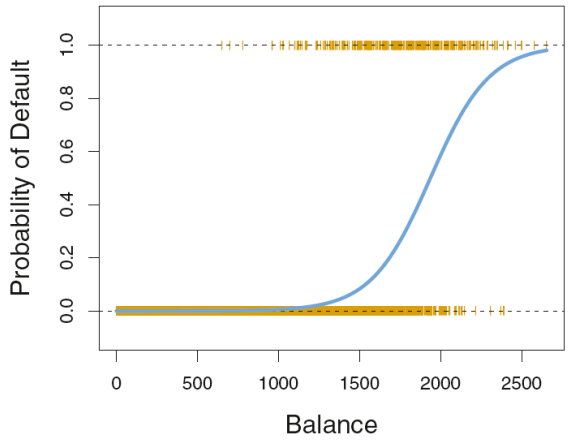


Рис. 5.8: Применение логистической регрессии в задаче оценки вероятности не вернуть задолженность.

Последнее свойство является весьма полезным. Например, в уже рассмотренной задаче при размере задолженности в районе 2000\$ оценка вероятности просрочки платежа сильно изменяется при увеличении или уменьшении задолженности на 100\$. С другой стороны при размере задолженности в 500\$ увеличение задолженности на 100\$ приводит только к незначительным изменениям требуемой оценки.

5.6.5. Оценка параметров

По функции $\pi(x)$ можно восстановить функцию g, которая фигурирует в определении обобщенной линейной модели:

$$\pi(x) \approx \frac{e^{\langle w,x\rangle}}{1+e^{\langle w,x\rangle}} \qquad \Longleftrightarrow \qquad \langle w,x\rangle \approx \underbrace{\ln \overbrace{\frac{\pi(x)}{1-\pi(x)}}^{\text{Риск}}}_{\text{Логит}}.$$

Отношение, стоящее под логарифмом, называется риском, а весь логарифм называется «логит». Именно поэтому метод называется логистической регрессией: логит приближается линейной комбинацией факторов.

Настройка модели происходит методом максимизации правдоподобия $L(X)$. Удобнее однако не максимизировать правдоподобие, а минимизировать минус логарифм от правдоподобия:

$$-\ln L(X) = -\sum_{i=1}^{\ell} \Big(y_i \ln \pi(x_i) + (1-y_i)\ln(1-\pi(x_i)) \Big)$$

Такой функционал также имеет названия log-loss, кросс-энтропия и другие.

Если изменить метку нулевого класса на -1, то получится логистическая функция потерь в таком виде, в котором она встречалась в курсе до этого:

$$Q(w,X) = \sum_{i=1}^{\ell} \ln\left(1+\exp\left(-y_i\langle w,x\rangle\right)\right)$$

5.6.6. Решение задачи максимизации правдоподобия

Задача максимизации правдоподобия в логистической регрессии очень хорошо решается численно, поскольку правдоподобие — выпуклая функция, а следовательно, она имеет единственный глобальный максимум. Кроме того, ее градиент и гессиан могут быть хорошо оценены.

Если объекты из разных классов линейно разделимы в пространстве признаков, возникает проблема переобучения: сигмоида вырождается в «ступеньку».

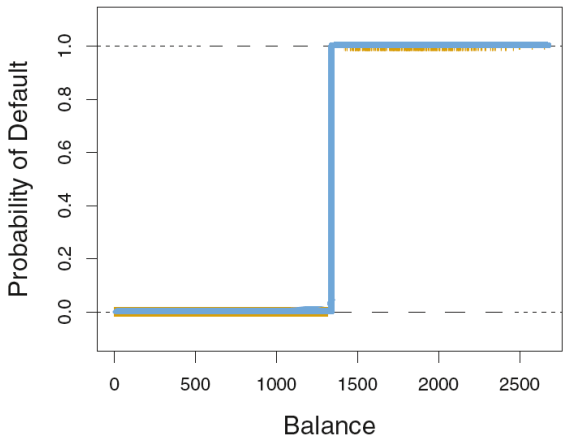


Рис. 5.9: Проблема переобучения в задаче логистической регрессии.

Например, такая ситуация возникает, если в уже упомянутой задаче оценки вероятности вернуть задолженность обучающая выборка такова, что все клиенты с задолженностью менее 1300$ вернули платеж вовремя, а все клиенты с задолженностью более 1300$ — нет.

В этом случае максимизация правдоподобия приводит к тому, что $\|w\| \to \infty$. В таких случаях необходимо использовать методы регуляризации, например L_1 или L_2 регуляризатор.

5.6.7. Предсказание отклика

Вероятности, которые дает логистическая регрессия, можно использовать для классификации, то есть для предсказания итоговых меток классов. Для этого выбирается порог p_0 и объект относится к классу 1 только в случае $\pi(x) > p_0$. В остальных случаях объект относится к классу 0.

Порог p_0 не следует выбирать всегда равным 0.5, как это может показаться из интуитивных соображений. Его необходимо подбирать для каждой задачи отдельно таким образом, чтобы обеспечить оптимальный баланс между точностью и полнотой классификатора.