Санкт-Петербургский государственный университет

Прикладная математика и информатика

Кафедра статистического моделирования

Регрессия и регуляризация

Романова Елизавета
Горбачук Анна
Сидоренко Денис

Санкт-Петербург
2019

Содержание

1 Обучение с учителем

Пусть наблюдается некоторый количественный отклик Y и p предикторов (признаков) $X_1, \ldots, X_p$. Будем предполагать, что между Y и $X = (X_1, \ldots, X_p)$ существует определенная связь, которую можно представить в виде

$$Y = f^*(X) + \varepsilon,$$

где f^* — фиксированная, но неизвестная функция от предикторов, ε — ошибка, которая не зависит от X и имеет нулевое среднее значение.

Можно рассматривать две различные задачи с разными целями: предсказание (prediction) и статистический вывод (inference).

- **Prediction.**

  Так как ошибки имеют нулевое среднее, можем предсказывать Y в соответствии с формулой

  $$\hat{Y} = \hat{f}(X),$$

  где $\hat{f}$ — оценка f^*, $\hat{Y}$ — предсказанное значение Y.

  Когда целью является предсказание, нам не важна точная форма функции $\hat{f}$, если она обеспечивает точные предсказания Y.

- **Inference.**

  Интересуемся тем, как изменения $X_1, \ldots, X_p$ влияют на Y. То есть, цель — оценить f^* (а не предсказать Y). В таком случае нужно знать точную форму $\hat{f}$.

В настоящем курсе нас больше интересуют предсказания.

Пусть имеется выборка из n отдельных наблюдений: $x_1, \ldots, x_n$ для которых известны соответствующие значения отклика. Обозначим x_{ij} — значение j-го признака i-го наблюдения, $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, y_i — отклик у i-го наблюдения.

Хотим найти такую функцию $\hat{f}$, что $y \approx \hat{f}(x)$ для любого наблюдения (x, y). Совокупность X_n пар $(x_1, y_1), \ldots, (x_n, y_n)$, которая участвует в оценке функции f^*, называется обучающей выборкой. Выборка $X'_k = (x'_i, y'_i)_{i=1}^k$, не участвующая в оценке функции f^*, называется тестовой (или контрольной).

2 Регрессия

2.1 Регрессия и МНК

Итак, пусть дана обучающая выборка $X_n = (x_i, y_i)_{i=1}^n$, где $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ и предполагается, что между ответами и объектами есть связь:

$$y_i = f^*(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

где ε_i — независимые одинаково распределенные случайные величины с $\mathrm{E}\,\varepsilon_i = 0$, $\mathrm{E}\,\varepsilon_i^2 = \sigma^2$.

Пусть задана модель регрессии — параметрическое семейство функций $f(x, \theta)$, где $\theta \in \Theta$ — вектор параметров модели, $\Theta \subset \mathbb{R}^p$ — пространство параметров, $f : \mathbb{R}^p \times \Theta \to \mathbb{R}$

— фиксированная функция. Выберем в качестве функционала качества аппроксимации целевой зависимости на выборке X_n среднеквадратическую ошибку:

$$\mathrm{MSE}_{\mathrm{train}} = Q(\theta, X_n) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i,\theta) - y_i)^2. \tag{1}$$

Обучение по методу наименьших квадратов (МНК) состоит в нахождении такого вектора параметров θ^*, при котором достигается минимум среднего квадрата ошибки на заданной обучающей выборке X_n:

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^p} Q(\theta, X_n).$$

MSE в (1) вычисляется на основе обучающей выборки, то есть наблюдений, которые были использованы для подгонки модели, так что это ошибка на обучающей выборке. В реальности нас интересует ошибка MSE на контрольной выборке, то есть то, насколько метод дает точное предсказание для наблюдений, которые не участвовали в оценке f^*. Нет гарантии, что метод с минимальной среднеквадратической ошибкой на обучающих данных также будет иметь минимальную MSE на контрольных данных.

Когда качество работы алгоритма на новых объектах, не вошедших в состав обучения, оказывается существенно хуже, чем на обучающей выборке ($\mathrm{MSE}_{\mathrm{test}} \gg \mathrm{MSE}_{\mathrm{train}}$), говорят об эффекте переобучения (overtraining) или переподгонки (overfitting).

2.2 Bias–variance tradeoff

Пусть $(x', y') \in X'_k$ — объект данных из тестовой выборки, $y' = f^*(x') + \varepsilon$, $\mathrm{E}\,\varepsilon = 0$, $\mathrm{E}\,\varepsilon^2 = \sigma^2$.

Для математического ожидания квадрата ошибки предсказания на (x', y') справедливо

$$\mathrm{E}(\hat{f}(x') - y')^2 = \mathrm{D}\,\hat{f}(x') + (\mathrm{Bias}\hat{f}(x'))^2 + \sigma^2, \tag{2}$$

$\mathrm{D}\,\hat{f}$ — дисперсия оценки $\hat{f}$, $\mathrm{Bias}\hat{f}(x')$ — смещение оценки, σ^2 — неустранимая ошибка.

В контексте inference нас может больше интересовать аналогичная формула

$$\mathrm{E}(\hat{f}(x') - f^*(x'))^2 = \mathrm{D}\,\hat{f}(x') + (\mathrm{Bias}\hat{f}(x'))^2. \tag{3}$$

Таким образом, MSE на контрольной выборке зависит от дисперсии оценки и квадрата ее смещения. Дисперсия оценки определяет то количество, на которое изменится $\hat{f}$, если бы мы получали эту оценку с использованием другого набора данных. Мы хотим, чтобы оценка $\hat{f}$ не менялась сильно на разных обучающих выборках. Смещение $\hat{f}$ характеризует ошибку, возникающую при аппроксимации реальной сложной функции f^* более простой моделью. То есть для минимизации ожидаемой ошибки на контрольных данных нужен такой метод обучения, который обеспечивает и низкую дисперсию, и низкое смещение.

2.3 Множественная линейная регрессия

Предполагаем, что зависимость между ответами и признаками линейная, а также, что ответы и признаки центрированы (для краткости записи: чтобы не приходилось дописывать свободный член или добавлять в матрицу $\mathbb{X}$ столбец из единиц).

Модель множественной линейной регрессии ($\beta_1, \ldots, \beta_p$ — параметры модели):

$$y_i = f(x_i; \beta_1, \ldots, \beta_p) + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

Введем матричные обозначения:

$$\mathbb{X} = [X_1 : \ldots : X_p] = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad B = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Модель множественной линейной регрессии в матричной записи:

$$Y = \mathbb{X}B + \mathcal{E}.$$

Минимизируем среднеквадратическую ошибку

$$Q(B, X) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \frac{1}{n} ||Y - \mathbb{X}B||^2 = \frac{1}{n} (Y - \mathbb{X}B)^{\mathrm{T}} (Y - \mathbb{X}B) \to \min_B.$$

Решение МНК:

$$\hat{B} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}Y = \mathbb{X}^{-}Y, \qquad \hat{Y} = \mathbb{X}\hat{B}. \tag{4}$$

Вычислительная проблема: при плохой обусловленности матрицы $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ вычисление обратной к ней матрицы крайне нежелательно. Поэтому на практике лучше избегать прямого использования формул (4).

Варианты обхода:

- Переход к нормальной системе (p неизвестных, p уравнений):

  $$\mathbb{X}^{\mathrm{T}}Y = \mathbb{X}^{\mathrm{T}}\mathbb{X}B.$$

  Существует большое количество численных методов решения нормальной системы. Наибольшей популярностью пользуются методы, основанные на ортогональных разложениях матрицы $\mathbb{X}$ (QR-разложение, например). Эти методы эффективны, обладают хорошей численной устойчивостью и позволяют строить различные модификации и обобщения.

- Использование сингулярного разложения.

  Пусть $\mathbb{X} = \mathbb{V}\mathbb{\Lambda}\mathbb{U}^{\mathrm{T}}$ — сингулярное разложение $\mathbb{X}$. Тогда псевдообратную к $\mathbb{X}$ матрицу легко записать в виде

  $$\mathbb{X}^{-} = \mathbb{U}\mathbb{\Lambda}^{-1}\mathbb{V}^{\mathrm{T}} = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j V_j^{\mathrm{T}}.$$

  Вектор МНК-решения:

  $$\hat{B} = \mathbb{X}^{-}Y = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j (V_j^{\mathrm{T}} Y). \tag{5}$$

Оценка Y:

$$\hat{Y} = \mathbb{X}\hat{B} = \sum_{j=1}^{p} V_j(V_j^{\mathrm{T}}Y). \tag{6}$$

Норма вектора коэффициентов:

$$||\hat{B}||^2 = \sum_{j=1}^{p} \frac{1}{\lambda_j}(V_j^{\mathrm{T}}Y)^2. \tag{7}$$

Таким образом, имея сингулярное разложение, не приходится вычислять обратную матрицу. Эффективные численные алгоритмы, вычисляющие SVD, реализованы во многих стандартных математических пакетах.

2.4 Мультиколлинеарность

Проблема мультиколлинеарности является общей для многих методов корреляционного анализа. МНК не исключение.

Если матрица данных содержит несколько сильно коррелированных признаков, то есть матрица начинает приближаться к вырожденной, то минимальное собственное число становится близким к 0. Что будет происходить в таком случае с МНК-оценкой?

При очень малых собственных числах λ_j соответствующие знаменатели в формулах (5) и (7) близки к нулю. Поэтому в суммах появляются очень большие и неустойчивые слагаемые.

Теряется интерпретируемость оценок коэффициентов (это важно, если нас интересует inference), так как коэффициенты могут неоправданно принимать очень большие значения.

При этом мы не заметим проблем, работая только с обучающей выборкой, так как на ней Y по-прежнему будет хорошо приближаться (в формуле (6) не участвуют собственные числа). Но на тестовой выборке ответы $Y' = \mathbb{X}'\hat{B}$ неустойчивы.

Таким образом, в линейных моделях мультиколлинеарность приводит к переобучению.

Способы решения проблемы:

- Регуляризация: проблема зарождается в мультиколлинеарности, а проявляется в том, что норма вектора коэффициентов увеличивается. Регуляризация контролирует увеличение нормы вектора.
- Преобразование признаков.
- Отбор признаков.

3 Регуляризация

В соответствии с формулами (2), (3), MSE на контрольной выборке зависит от дисперсии оценки $\hat{f}$ и ее смещения.

Когда связь между откликом и предикторами (почти) линейна, оценки по методу наименьших квадратов обладают (почти) нулевым смещением, но при этом могут иметь высокую дисперсию.

Ковариационная матрица МНК-оценки $\hat{B}$:

$$\mathrm{Cov}(\hat{B}) = \sigma^2(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}. \tag{8}$$

Чем больше дисперсия оценки $\hat{B}$, тем больше дисперсия $\hat{f}$. Когда матрица $\mathbb{X}$ близка к вырожденной (это может произойти из-за наличия мультиколлинеарности или когда число предикторов p почти равно числу наблюдений n), дисперсия $\hat{B}$ становится большой и $\mathrm{MSE}_{\mathrm{test}}$ увеличивается. При $p > n$ или при полностью коллинеарных признаках оценки по методу наименьших квадратов не имеют уникального решения.

Введение небольшого смещения в оценке может привести к значительному уменьшению дисперсии и тем самым уменьшению $\mathrm{MSE}_{\mathrm{test}}$.

3.1 Регуляризация Тихонова

Метод наименьших квадратов решает нормальную систему $\mathbb{X}^{\mathrm{T}}\mathbb{X}B = \mathbb{X}^{\mathrm{T}}Y$. Идея метода регуляризации Тихонова состоит в том, чтобы прибавить к (возможно) плохо обусловленной матрице $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ другую матрицу $\mathbb{T}^{\mathrm{T}}\mathbb{T}$ так, чтобы их сумма была хорошо обусловлена. Переходим к следующей нормальной системе:

$$(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})B = \mathbb{X}^{\mathrm{T}}Y.$$

Решение последней системы соответствует минимизации функции

$$||Y - \mathbb{X}B||_2^2 + ||\mathbb{T}B||_2^2.$$

Решение:

$$\hat{B}_T = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}Y.$$

Оценка $\hat{B}_T$, конечно, уже будет иметь смещение:

$$\mathrm{E}\,\hat{B}_T = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}B.$$

Легко показать, что ковариационная матрица полученной оценки

$$\mathrm{Cov}(\hat{B}_T) = \sigma^2(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}((\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1})^{\mathrm{T}},$$

поэтому при наличии полностью коллинеарных признаков дисперсия соответствующей оценки $\hat{f}_{\mathrm{T}}$ (и, как следствие, $\mathrm{MSE}_{\mathrm{test}}$) будет конечна, в отличие от дисперсии (8) обычной МНК-оценки (4).

3.2 Гребневая регрессия (Ridge regression)

Гребневая регрессия — это наиболее распространенный частный случай метода регуляризации Тихонова с $\mathbb{T} = \sqrt{\tau}\mathbb{I}$.

3.2.1 Задача гребневой регрессии

Вводится штраф за увеличение нормы вектора B и минимизируется следующая функция:

$$Q_\tau(B) = ||\mathbb{X}B - Y||^2 + \tau||B||^2 \to \min_B,$$

где τ — неотрицательный параметр регуляризации.

В развернутом виде задача оптимизации записывается так:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \tau\sum_{j=1}^{p}\beta_j^2 \to \min_B.$$

Решение задачи гребневой регрессии:

$$\hat{B}_\tau = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \tau\mathbb{I}_p)^{-1}\mathbb{X}^{\mathrm{T}}Y.$$

3.2.2 Параметр регуляризации

Чем больше коэффициент регуляризации τ, тем устойчивее решение, но больше смещение. Когда $\tau = 0$, гребневая регрессия совпадает с обычной регрессией, но при $\tau \to \infty$ коэффициенты регрессии стремятся к нулю. Для каждого значения τ гребневая регрессия порождает свой оптимальный набор оценок коэффициентов $\hat{\beta}_1, \ldots, \hat{\beta}$. Важно подобрать хорошее значение параметра τ, чтобы достичь компромисса между смещением и неустойчивостью.

Подход на основе сингулярного разложения $\mathbb{X} = \mathbb{V}\mathbb{\Lambda}\mathbb{U}^{\mathrm{T}}$ позволяет подбирать параметр τ, вычислив SVD только один раз.

Решение гребневой регрессии через SVD:

$$\hat{B}_\tau = \mathbb{U}(\mathbb{\Lambda}^2 + \tau\mathbb{I}_p)^{-1}\mathbb{\Lambda}\mathbb{V}^{\mathrm{T}}Y = \sum_{j=1}^{p}\frac{\sqrt{\lambda_j}}{\lambda_j + \tau}U_j(V_j^{\mathrm{T}}Y).$$

Оценка функции f^* для выборки X_n через SVD:

$$\mathbb{X}\hat{B}_\tau = \mathbb{V}\mathbb{\Lambda}\mathbb{U}^{\mathrm{T}}\hat{B}_\tau = \mathbb{V}\mathrm{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right)\mathbb{V}^{\mathrm{T}}Y = \sum_{j=1}^{p}\frac{\lambda_j}{\lambda_j + \tau}V_j(V_j^{\mathrm{T}}Y).$$

Таким образом, необходимо один раз произвести сингулярное разложение матрицы $\mathbb{X}$, а затем несложным образом вычислять вектор оценок параметров для интересующих значений параметра τ.

Добавление в знаменатель положительного числа τ приводит к тому, что проблема неустойчивости уходит.

3.2.3 Вероятностная интерпретация гребневой регрессии

Пусть в линейной модели выполнены следующие предположения:

- ошибки независимы и $\varepsilon_i \in N(0, \sigma^2)\ \forall i = 1, \ldots, n$.

- вектор параметров B имеет априорное распределение $\pi(B) \stackrel{d}{=} N(\mathbf{0}, \frac{\sigma^2}{\tau}\mathbb{I})$.

Пусть $L(X, Y, B)$ — функция правдоподобия выборки. По теореме Байеса при фиксированном X апостериорное распределение $q(B|X, Y)$ пропорционально $L(X, Y, B)\pi(B)$.

Покажем, что в условии выполненных предположений оценка апостериорного максимума B совпадает с решением гребневой регрессии.

Оценка максимума апостериорной вероятности:

$$\arg\max_B q(B|X,Y) = \arg\max_B L(X,Y,B)\pi(B) =$$

$$= \arg\max_{\beta_1,\ldots,\beta_p} \exp\left(-\sum_{j=1}^{n}\frac{\varepsilon_j^2}{2\sigma^2}\right)\exp\left(-\sum_{i=1}^{p}\frac{\tau\beta_i^2}{2\sigma^2}\right) =$$

$$= \arg\max_{\beta_1,\ldots,\beta_p} \exp\left(-\sum_{j=1}^{n}\frac{(y_j - \sum_{i=1}^{p}\beta_i x_{ij})^2}{\sigma^2}\right)\exp\exp\left(-\sum_{i=1}^{p}\frac{\tau\beta_i^2}{2\sigma^2}\right) =$$

$$= \arg\max_B \exp\left(-\frac{||Y - \mathbb{X}B||^2}{2\sigma^2} - \frac{\tau||B||^2}{2\sigma^2}\right) =$$

$$= \arg\min_B(||Y - \mathbb{X}B||^2 + \tau||B||^2).$$

Пришли к решению задачи гребневой регрессией с параметром регуляризации τ.

3.2.4 Проблемы и замечания

- Стандартные МНК-оценки инварианты относительно умножения признака на константу, то есть значение $X_j\hat{\beta}_j$ не зависит от масштаба j-го признака. Оценки МНК гребневой регрессии не обладают свойством инвариантности и могут существенно меняться. Поэтому гребневую регрессию нужно использовать после стандартизации признаков.

- В конечную модель входят все начальные признаки, если признаков много, то усложняется интерпретация.

3.3 Лассо (Lasso)

С задачей отбора признаков справляется Лассо регрессия, в которой в качестве штрафа на норму коэффициентов используется l_1-норма вектора коэффициентов.

3.3.1 Задача Lasso-регрессии

Метод LASSO решает следующую задачу минимизации:

$$||\mathbb{X}B - Y||_2^2 + \tau||B||_1^2 \to \min_B,$$

где τ — неотрицательный параметр регуляризации.

Задача оптимизации в развернутом виде:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \tau\sum_{j=1}^{p}|\beta_j| \to \min_{\beta_1,\ldots,\beta_p}.$$

Сложность задачи состоит в ее негладкости, из-за которой мы не можем сразу применить теорему 1 (теорему Куна-Таккера).

Задачу lasso-оптимизации можно переписать в форме с ограничениями:

$$\begin{cases}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \to \min_{\beta_1,\ldots,\beta_p},\\ \sum_{j=1}^{p}|\beta_j| \leq æ,\end{cases}$$

где $æ = 1/\tau$.

Приведем задачу к каноничному виду. Представим каждый параметр β_j в виде разности положительной и отрицательной частей: $\beta_j = \beta_j^+ - \beta_j^-$. Тогда $|\beta_j| = \beta_j^+ + \beta_j^-$. После замены переменных переходим к задаче ($2p$ переменных, $2p+1$ ограничений):

$$\begin{cases}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}(\beta_j^+ - \beta_j^-)x_{ij}\right)^2 \to \min_{\beta_1^+,\ldots,\beta_p^+,\beta_1^-,\ldots,\beta_p^-},\\ \sum_{j=1}^{p}\beta_j^+ + \beta_j^- \leq æ, \quad \beta_j^+ \geq 0, \quad \beta_j^- \geq 0.\end{cases}$$

Получили выпуклую задачу квадратичного программирования с линейными ограничениями-неравенствами, к которой применима теорема Куна-Таккера.

Чем меньше параметр æ, тем больше ограничений обращаются в равенства: $\beta_j^+ = \beta_j^- = 0$, что соответствует обнулению коэффициента β_j и исключению j-го признака.

3.3.2 Вероятностная интерпретация Лассо

Пусть в линейной модели выполнены следующие предположения:

- ошибки независимы и $\varepsilon_i \in N(0,\sigma^2)\ \forall i = 1,\ldots,n$.
- вектор параметров B имеет априорное распределение $\pi(B) = \prod_{j=1}^{p} g(\beta_j)$, где g — плотность распределения Лапласа $\text{Laplace}(0,\tau)$.

Пусть $L(X,Y,B)$ — функция правдоподобия выборки. По теореме Байеса при фиксированном X апостериорное распределение $q(B|X,Y)$ пропорционально $L(X,Y,B)\pi(B)$.

В условии выполненных предположений оценка апостериорного максимума B совпадает с решением лассо-регрессии.

3.4 Сравнение гребневой регрессии и Лассо

Сначала заметим, что задачу гребневой регрессии можно представить в виде задачи минимизации с ограничениями

$$\begin{cases}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \to \min_{\beta_1,\ldots,\beta_p},\\ \sum_{j=1}^{p}\beta_j^2 \leq æ.\end{cases}$$

Ранее мы также получали соответствующую форму записи для лассо-регрессии:

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \to \min_{\beta_1,\ldots,\beta_p}, \\ \sum_{j=1}^{p} |\beta_j| \leq æ. \end{cases}$$

Рассмотрим простой случай, когда $p = 2$. Тогда выражение $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2$ — это эллипс с центром в точке $\hat{B}$. Предположим, что центр эллипса не удовлетворяет ограничениям $\sum_{j=1}^{p} \beta_j^2 \leq æ$ и $\sum_{j=1}^{p} |\beta_j| \leq æ$, то есть лежит вне круга в случае гребневой регрессии и вне ромба в случае Лассо. Тогда решения задач минимизации будут лежать на границе возможных значений. На рис. 1 видно, что для Лассо существует гораздо больше различных эллипсов, которые пересекались бы с ромбом (ограничениями) таким образом, чтобы один из коэффициентов был равен нулю.

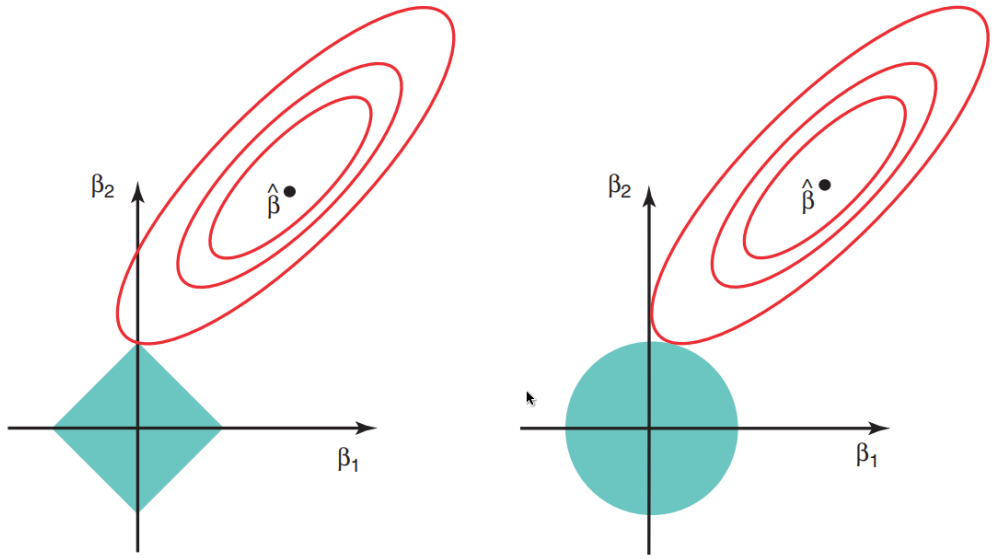

Рис. 1: Границы ошибки $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2$ и ограничений $\sum_{j=1}^{p} |\beta_j| \leq ae$ для Лассо (слева) и $\sum_{j=1}^{p} \beta_j^2 \leq æ$ для гребневой регрессии (справа).

Замечания:

- Обычно Лассо подходит лучше в случае наличия в данных большого количества лишних (незначимых) признаков.
- Для реальных данных обычно заранее не известно количество признаков, значимо влияющих на зависимую переменную.
- С помощью кросс-валидации можно определить какой подход лучше для конкретных данных.

3.5 Elastic net regularization

Решается задача оптимизации

$$||Y - \aleph B||_2^2 + \tau_1 ||B||_1^2 + \tau_2 ||B||_2^2 \to \min_B .$$

- Elastic net — это комбинация методов Lasso и Ridge:
  - Когда $\tau_1 = 0$: Ridge регрессия;
  - Когда $\tau_2 = 0$: Lasso регрессия;
- Elastic net обычно дает лучшие результаты, чем Lasso, при наличии коррелированных признаков;
- При наличии группы релевантных и избыточных признаков Lasso обычно имеет тенденцию отказываться от всех, кроме одного признака из этой группы, в то время как Elastic net будет выбирать всю группу признаков.
- Elastic net можно свести к SVM, для которого разработано много быстрых решений.

4 Нелинейная регрессия

Пусть задана нелинейная модель регрессии $f(x, \theta)$, $\theta \in \mathbb{R}^k$. Решаем задачу минимизации функционала среднеквадратичного отклонения:

$$Q(\theta, X) = \sum_{i=1}^{n} (f(x_i, \theta) - y_i)^2 \to \min_B .$$

К численному решению этой задачи можно применять метод стохастического градиента, но он имеет первый порядок сходимости и может сходиться не слишком быстро. Далее рассмотрим методы второго порядка.

4.1 Метод Ньютона-Рафсона

Выберем начальное приближение $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$ и организуем итерационный процесс

$$\theta^{t+1} := \theta^t - h_t (Q''(\theta^t))^{-1} Q'(\theta^t), \tag{9}$$

где $Q'(\theta^t)$ — градиент Q в точке θ^t, $Q''(\theta^t)$ — гессиан Q в точке θ^t (матрица порядка $k \times k$), h_t — величина шага (простейший вариант: $h_t = 1$).

Компоненты градиента:

$$\frac{\partial Q(\theta)}{\partial \theta_j} = 2 \sum_{i=1}^{n} (f(x_i, \theta) - y_i) \frac{\partial f(x_i, \theta)}{\partial \theta_j}. \tag{10}$$

Компоненты гессиана:

$$\frac{\partial^2 Q(\theta)}{\partial \theta_j \partial \theta_m} = 2 \sum_{i=1}^{n} \frac{\partial f(x_i, \theta)}{\partial \theta_j} \frac{\partial f(x_i, \theta)}{\partial \theta_m} - 2 \sum_{i=1}^{n} (f(x_i, \theta) - y_i) \frac{\partial^2 f(x_i, \theta)}{\partial \theta_j \partial \theta_m}. \tag{11}$$

Основная сложность — обращение гессиана (матрицы порядка $k \times k$) на каждой итерации (9).

4.2 Метод Ньютона-Гаусса

Рассмотрим модификацию метода Ньютона-Рафсона, основанную на линеаризации функции f.

В методе Ньютона-Рафсона мы считаем градиент и гессиан в конкретных точках θ^t. Можно считать, что в окрестности точки θ^t функция f — линейная.

Линеаризация $f(x_i, \theta)$ в окрестности θ^t:

$$f(x_i, \theta) = f(x_i, \theta^t) + \sum_{j=1}^{k} \frac{\partial f(x_i, \theta_j)}{\partial \theta_j}(\theta_j - \theta_j^t) + o(\theta_j - \theta_j^t).$$

Подставляем это представление f в формулы (10), (11). Первые производные не изменятся, а вторые производные будут равны нулю. Таким образом избавились от второго слагаемого в формуле для компонент гессиана.

Введем обозначения:

- $\mathbb{F}_t = (\frac{\partial f}{\partial \theta_j}(x_i, \theta^t))_{n \times k}$ — матрица первых производных,
- $f_t = (f(x_i, \theta^t))_{n \times 1}$ — вектор значений f.

Итерация метода Ньютона-Гаусса:

$$\theta^{t+1} := \theta^t - h_t \underbrace{(\mathbb{F}_t^{\mathrm{T}} \mathbb{F}_t)^{-1} \mathbb{F}_t^{\mathrm{T}} (f_t - Y)}_{\widetilde{B}}.$$

$\widetilde{B}$ — решение задачи множественной линейной регрессии

$$||\mathbb{F}_t B - (f_t - Y)||^2 \to \min_B.$$

Таким образом нелинейная регрессия сводится к серии линейных регрессий.

5 Приложение

5.1 Теорема Куна-Таккера

Пусть $x \in \mathbb{R}^n$. Рассмотрим задачу

$$\begin{gathered} f(x) \to \min, \\ g_i(x) \le 0, \quad i = 0, \ldots, m. \end{gathered}$$

Теорема 1. *Пусть $f(x)$ выпукла и дифференцируема на допустимом множестве. Все ограничения регулярные (аффинные функции). Тогда x_* — оптимальное решение тогда и только тогда, когда $\exists \lambda_i$ такие, что*

$$\begin{gathered} \frac{\partial f(x_*)}{\partial x_j} + \sum_{i=1}^{m} \lambda_i \frac{\partial g(x_*)}{\partial x_j} = 0 \quad j = 1, \ldots, n, \\ g_i(x_*) \le 0, \quad \lambda_i \ge 0, \quad \lambda_i g_i(x_*) = 0, \quad i = 1, \ldots, m. \end{gathered}$$