Регрессия, регуляризация, отбор признаков

Дейвид Капаца, Анастасия Мандрикова, Елена Гоголева

29 октября 2021 г.

Содержание

1 Регрессия в ML и вероятностная модель

Сначала о том, какую задачу мы хотим решить и какие минимальные условия следует наложить на данные для того, чтобы она была решаемой.

Пусть имеется некоторый набор данных (обучающая выборка):

1. $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ — матрица данных, матрица плана (data matrix, design matrix), состоит из столбцов X_i (признаки) и строк $\mathbf{x}_i$ (индивиды, объекты);
2. $\boldsymbol{y} \in \mathbb{R}^n$ — ответ, вектор наблюдений (response vector).

Задача регрессии:

Уметь предсказывать y_i (ответы) по *новым* $\mathbf{x}_i$ (объектам, индивидам), установив некоторую зависимость на обучающей выборке.

1.1 ML подход: гипотеза непрерывности

Какими должны быть данные для того, чтобы данная задача была корректной? Нужно, чтобы все рассматриваемые объекты были в некотором смысле однородны и происходили из некоторой генеральной совокупности (если иначе, то как предсказать ответ, когда новый объект $\mathbf{x}_i$ совершенно не похож на объекты обучающей выборки).

В машинном обучении для обоснования использования методов регрессии используется так называемая **гипотеза непрерывности**:

«близким» объектам $\mathbf{x}_i$ соответствуют «близкие» ответы y_i

Такая гипотеза, несмотря на свою простоту и наглядность, допускает множество интерпретаций и имеет свои недостатки. Пожалуй, главная проблема заключается в полном отсутствии случайности.[1] Можно формализовать задачу регрессии на вероятностном языке. Приведём один из вариантов.

1.2 Вероятностная постановка задачи регрессии

Пусть $\boldsymbol{\xi} \in \mathbb{R}^p$ — случайный вектор, $\eta, \varepsilon \in \mathbb{R}$ — случайные величины.

Предполагаем, что η и $\boldsymbol{\xi}$ функционально зависимы:

$$\eta = \varphi(\boldsymbol{\xi}) + \varepsilon. \tag{1}$$

Обычно $\mathrm{E}\varepsilon = 0$, $\mathrm{D}\varepsilon = \sigma^2$, $\boldsymbol{\xi} \perp \varepsilon$; часто имеют место предположения о распределении ε.

Задачей в данном случае является нахождение функции $\boldsymbol{\varphi}$. Переходя к выборке, наблюдаем $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{L}(\boldsymbol{\xi})$ и $y_1, \ldots, y_n \sim \mathcal{L}(\eta)$. На основании этой выборки делаем предположение о $\boldsymbol{\varphi}$, получаем её приближение $\hat{\boldsymbol{\varphi}}$. Предсказанием ответа $\tilde{y}$ для нового объекта $\tilde{\mathbf{x}}$ на построенной модели будет $\hat{\boldsymbol{\varphi}}(\tilde{\mathbf{x}})$.

Таким образом, с помощью вероятностного подхода удаётся формализовать задачу и получить удобную для построения и анализа модели вероятностную интерпретацию. Далее мы рассмотрим стандартные этапы обучения модели с учителем, которые применяются и в регрессионной постановке.

[1] Если объект и ответ регистрируются со случайной ошибкой, то мы уже не можем сказать, что одинаковые объекты приводят к одинаковым ответам. Но тогда получаем противоречие с гипотезой непрерывности.

1.3 Этапы обучения модели

Так как тема регрессии в целом предполагается неплохо знакомой, а доклад является вводным в тематику обучения с учителем, напомним классическую схему обучения модели. Каждый из этапов приведён с соответствующим примером (случай классической линейной регрессии).

1. **Выбор модели регрессии** (класс рассматриваемых $\varphi(\cdot)$)

   Линейная модель: $\varphi(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_i \mathbf{x}_i[j], \quad i \in 1:n$

   Чаще всего на практике в качестве кандидатов на $\boldsymbol{\varphi}$ рассматривается именно класс линейных функций. Такой выбор обусловлен явным видом решения во многих случаях, его простотой и повышенной интерпретируемостью.

2. **Выбор функции потерь** (loss function)

   Квадратичная функция потерь: $\sum_{i=1}^{n}(y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}))^2$

   Значение функции потерь отражает качество рассматриваемой модели на тренировочной выборке, измеряя отличие между предсказаниями $\tilde{y}_i$ и наблюдениями y_i. Вариант по умолчанию — сумма квадратов остатков — обусловлен своей простотой и дифференцируемостью, а также тем, что он естественно возникает при предположении о нормальном распределении остатков $\boldsymbol{\varepsilon}$.

3. **Выбор метода обучения** (training)

   МНК: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}))^2$

   Обучение модели представляет из себя задачу нахождения наилучшей модели среди рассматриваемых (например, заданных семейством параметров $\boldsymbol{\beta}$). Обычно такой выбор происходит за счёт минимизации функции потерь. В указанном случае линейной регрессии решение находится явно, однако так получается не всегда. В зависимости от вида функции потерь приходится использовать различные методы оптимизации (градиентные, если есть производная; стохастические, если функция сложная; условные, если на параметры накладываются некоторые ограничения, и т.д.). Также обращают внимание на свойства полученной оценки: в указанном примере получаем наилучшую несмещённую оценку вектора коэффициентов (BLUE).

4. **Выбор метода проверки** (test)

   MSE: $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}}(y_i^{\text{test}} - \varphi(\mathbf{x}_i^{\text{test}}, \hat{\boldsymbol{\beta}}))^2$

   Для оценки качества построенной модели и сравнения модели с другими построенными используется тестовая выборка, в нашем случае это mean-squared error. Сравнение величин ошибок на тестовой и на тренировочной выборке может оказаться полезным для выявления проблемы переобучения: низкая ошибка при тренировке и высокая при тестировании могут свидетельствовать о наличии этой проблемы.

Далее мы подробнее остановимся на некоторых этапах.

2 Задача регрессии как задача оптимизации

Задача параметрической[2] регрессии может быть сформулирована как задача минимизации некоторого функционала от выборки. Рассмотрим такую формулировку.

Пусть заданы:

[2] непараметрическую мы не рассматриваем ввиду того, что с ней сложнее предсказывать

- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ — матрица данных (design matrix);
- $\boldsymbol{y} \in \mathbb{R}^n$ — вектор ответов;
- $\boldsymbol{\beta} \in \mathbb{R}^d$ — вектор параметров;
- $\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) := (\varphi(\mathbf{x}_1, \boldsymbol{\beta}), \ldots, \varphi(\mathbf{x}_n, \boldsymbol{\beta}))^{\mathrm{T}}$ — функция от выборки и параметров;
- $\mathcal{L}(\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}), \boldsymbol{y})$ — *некоторая* функция потерь.[3]

Тогда решением задачи регрессии будет вектор $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}), \boldsymbol{y}).$$

Такая запись сразу позволяет увидеть, что мы имеем дело с некоторой оптимизационной задачей. Минимизация функции $\mathcal{L}(\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}), \boldsymbol{y})$ на некотором пространстве параметров приводит к оптимальному решению.[4]

В редких случаях удаётся найти явное решение,[5] однако в общем случае решение находится приближённо с использованием методов приближённых вычислений, поэтому помимо статистической ошибки полученного результата следует учитывать и вычислительную погрешность, которая зависит от выбранного метода оптимизации, точности входных данных, а также момента остановки алгоритма.

Также нельзя забывать и про то, что при использовании некоторых методов численной оптимизации есть шанс попасть не в глобальный, а в локальный минимум или седловую точку, что приведёт к неудовлетворительной оценке $\hat{\boldsymbol{\beta}}$. Применяя ту и иную функцию оптимизации, следует проверить наличие ограничений или условий для сходимости алгоритма.

3 Выбор функции потерь

Выбор функции потерь $\mathcal{L}$ несёт критическую роль в решении задачи регрессии, поэтому он должен быть обоснован каждый раз, когда решается та или иная задача. Есть множество вариантов функций потерь для разных задач, в качестве примеров перечислим три из них, которые хорошо изучены и имеют свои применения в разных условиях:

- $\|\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) - \boldsymbol{y}\|_2^2$ — квадратичная ошибка (l_2-норма);
- $\|\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) - \boldsymbol{y}\|_1$ — модуль ошибки (l_1-норма);
- $\sum_{i \in 1:n} H_\delta(\varphi(\mathbf{x}_i, \boldsymbol{\beta}), y_i)$ — функция потерь Хубера,
  где H_δ — функция Хубера, квадратичная на $[-\delta, \delta]$ и линейная на $|x| > \delta$.

[3] здесь $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

[4] конечно, можно вводить ограничения на пространство (например, исходя из смысла задачи, можно рассматривать только положительные $\hat{\boldsymbol{\beta}}$), но в данном случае для простоты рассматриваем безусловную оптимизацию

[5] например, в случае линейной регрессии при выборе суммы квадратов остатков в качестве функции потерь

Как выбирать? Выбор должен быть основан на нескольких критериях, вот их приблизительная подборка (приоритет пунктов определяется в зависимости от конкретной задачи):

- Явный вид решения;
- Простота функции $\mathcal{L}$ для оптимизации;
- Точность данных/наличие выбросов;
- Конкретные предположения о распределении остатков $\boldsymbol{\varepsilon}$;
- Инвариантность решения относительно масштаба/сдвига для признаков.

В дальнейшем осветим эти пункты подробнее на примере линейной регрессии.

4 Линейная регрессия

Частным случаем задачи регрессии является линейная регрессия. Мы делаем предположение о том, что модель данных имеет следующий вид:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

где

- $\boldsymbol{y} \in \mathbb{R}^n$ — вектор ответов, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ — вектор ошибок, $\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0}$;
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ — матрица данных (design matrix):
  - детерминированная (для простоты рассматриваем этот случай, то есть предполагаем, что случайность в модели происходит только от вектора шума);
  - случайная (результаты похожи, но уже с условными математическими ожиданиями из-за случайности матрицы, здесь не будем рассматривать);
- $\boldsymbol{\beta} \in \mathbb{R}^p$ — вектор параметров;
- $n \geqslant p$.

Заметим, что такое предположение обосновано не только простотой результирующей модели. Если столбцы матрицы $\boldsymbol{X}$ (то есть признаки) и вектор $\boldsymbol{y}$ распределены нормально, то известно, что $\boldsymbol{y}$ является *линейной* комбинацией столбцов матрицы $\boldsymbol{X}$.

На случайную ошибку обычно накладываются следующие требования:

$$\mathrm{E}\varepsilon_i = 0,\ \mathrm{E}\varepsilon_i^2 = \sigma^2 < +\infty,\ \mathrm{E}\varepsilon_i\varepsilon_j = 0. \tag{3}$$

Решение задачи линейной регрессии — вектор $\hat{\boldsymbol{\beta}}$.

Если не оговорено иное, под задачей линейной регрессии подразумевается задача минимизации квадратичной функции потерь:

Классическая задача:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

Полученную оценку $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ называют оценкой по методу наименьших квадратов (МНК-оценкой). Она имеет явный вид (если матрица $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ невырожденная):[6]

$$\hat{\boldsymbol{\beta}}_{\text{МНК}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{4}$$

Сразу отметим, что наличие явного вида решения крайне удобно в вычислительном плане. Оценка вычисляется достаточно быстро посредством применения сингулярного разложения матрицы данных $\boldsymbol{X}$, в чём мы убедимся далее.

4.1 Особенности МНК-оценки

В данном разделе отметим основные особенности оценки (4).

Начнём с математического ожидания и дисперсии оценки (напомним, что действуем в предположении, что матрица $\boldsymbol{X}$ — детерминированная). Математическое ожидание вычислим прямо здесь:

$$\begin{aligned}\mathrm{E}\hat{\boldsymbol{\beta}}_{\text{МНК}} = \mathrm{E}[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}] = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}\boldsymbol{y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \\ = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\underbrace{\mathrm{E}\boldsymbol{\varepsilon}}_{=0}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}) = \\ = \underbrace{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})}_{=\mathbf{I}}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned} \tag{5}$$

Таким образом, мы показали, что оценка $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ является *несмещённой*.

С дисперсией (ковариационной матрицей, если точнее) чуть посложнее, поэтому просто выпишем результат (при условии выполнения требований (3)):

$$\mathrm{D}\hat{\boldsymbol{\beta}}_{\text{МНК}} = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}. \tag{6}$$

Теорема Гаусса–Маркова утверждает, что $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ имеет наименьшую дисперсию среди всех несмещённых оценок (best linear unbiased estimate — BLUE). Таким образом, у найденной оценки отсутствует смещение, и в то же время она имеет наименьшую дисперсию среди всех возможных оценок.

Хорошая оценка $\hat{\boldsymbol{\beta}}$ должна иметь низкую среднеквадратическую[7] ошибку $\mathrm{E}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^2$, а она раскладывается в следующую сумму:

$$\mathrm{E}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^2 = \underbrace{\mathrm{D}\hat{\boldsymbol{\beta}}}_{\text{дисперсия}} + \underbrace{(\mathrm{E}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^2}_{\text{смещение}}. \tag{7}$$

Отсюда становится видно, что обе характеристики — дисперсия и смещение — одинаково важны для получения наилучшей оценки и что несмещённая оценка всё же может иметь большу́ю среднеквадратичную ошибку из-за достаточно большой дисперсии. О том, как можно уменьшить среднеквадратическую ошибку за счёт допущений на смещение оценки, мы поговорим в разделе 5.

Также добавим, что полученная оценка $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ (при выполении условий (3) и нормальности ошибок) является оценкой по методу максимума правдоподобия (ОМП), что позволяет говорить об асимптотической нормальности и асимптотической эффективности $\hat{\boldsymbol{\beta}}_{\text{МНК}}$.

[6] берутся частные производные по компонентам вектора $\boldsymbol{\beta}$ функции потерь и приравниваются к нулю. В результате получаем уравнение, которое и даёт указанную оценку.

[7] это один из наболее распространённых критериев того, что выбранная оценка является наилучшей. К примеру, выборочное среднее является решением задачи минимизации среднеквадратической ошибки $\min_a \frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2$

4.2 Распределение ошибки в модели

Если ошибки удовлетворяют данным в (3) условиям, а также распределены по нормальному закону, МНК-оценка является BLUE и ОМП. В таком случае использование хорошо известных на практике критериев о значимости регрессии является корректным.

Однако, такие требования выполняются не всегда. Данные могут иметь выделяющиеся наблюдения (outliers), а ошибки могут быть распределены не нормально; также, при отклонении от линейной модели возможно возникновение проблемы гетероскедастичности.

В таком случае становится необходимостью использование других функций потерь и других методов оптимизации. Говорить о явном решении уже не приходится, однако на практике применение таких подходов приводит к более устойчивым и лучшим в плане среднеквадратической ошибки оценкам.

4.3 Вычисление МНК-оценки: сингулярное разложение

В этом разделе покажем, как используется сингулярное разложение матрицы для упрощения задачи вычисления $\hat{\boldsymbol{\beta}}_{\text{МНК}}$, а затем обсудим вопросы вычислительной устойчивости полученной оценки.

Напомним некоторые необходимые факты про сингулярное разложение. *Сингулярным разложением* матрицы $\boldsymbol{X}$ называется разложение $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}$, где

- $\boldsymbol{V}$ и $\boldsymbol{U}$ — ортогональные,[8] $\boldsymbol{D}$ — диагональная;
- $\boldsymbol{V} = (V_1, V_2, \ldots, V_n) \in \mathbb{R}^{n\times n}$, V_i — собственные векторы $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$;
- $\boldsymbol{U} = (U_1, U_2, \ldots, U_n) \in \mathbb{R}^{p\times n}$, U_i — собственные векторы $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$;
- $\boldsymbol{D} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, $\lambda_j \geqslant 0$ — собственные значения $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

Для простоты предположим, что имеем дело с матрицей полного ранга, $p = n$ (результаты распространяются на случай $n > p$).

Подставим в формулу для $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ вместо матрицы $\boldsymbol{X}$ её сингулярное разложение и получим

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{МНК}} = \\ = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \\ = (\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{D}^{-2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \\ = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y}, \quad (8)\end{aligned}$$

где $\boldsymbol{D}^{-1} = \operatorname{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})$. Если предположить, что вычисление сингулярного разложения на компьютере происходит быстро и с малой погрешностью (в целом так и есть), то такой подход к вычислению $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ оказывается наиболее предпочтительным.

4.4 Мультиколлинеарность

Подход с сингулярным разложением также позволяет пролить свет на одну из главных проблем в задаче регрессии — проблему коллинеарности признаков (столбцов матрицы $\boldsymbol{X}$).

Обратим внимание на разложение (8). Формула для $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ содержит матрицу $\boldsymbol{D}^{-1} = \operatorname{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})$, состоящую из корней собственных чисел $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Ясно, что когда

[8] основное свойство ортогональных матриц: $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$ (используется при выводе формулы для $\hat{\boldsymbol{\beta}}_{\text{МНК}}$)

собственные числа оказываются приближённо равными нулю, возникает проблема вычислительной неустойчивости, когда погрешность вычислений может возрасти существенно. В таком случае говорят, что матрица $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ близка к вырожденной или плохо обусловлена.

Когда собственные значения $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ малы?
Ответ на этот вопрос можно получить из следующего факта:

Если существует $\boldsymbol{v} \in \mathbb{R}^p$ такой, что $\boldsymbol{X}\boldsymbol{v} \approx \boldsymbol{0}$, то некоторые $\lambda_i \approx 0$.

Таким образом, когда некоторые из признаков близки к коллинеарным, некоторые собственные числа становятся близкими к нулю. Следующее определение формализует данную проблему.

Определение 1 $\mu(\boldsymbol{S}) = \|\boldsymbol{S}\|\|\boldsymbol{S}^{-1}\| = \lambda_{max}/\lambda_{min}$ *называется* ***числом обусловленности*** *матрицы* $\boldsymbol{S}$

Ещё один известный факт[9] состоит в том, что при вычислении $\boldsymbol{S}^{-1}\boldsymbol{v} = \boldsymbol{z}$ (что, вообще говоря, и делается при подсчёте оценки $\hat{\boldsymbol{\beta}}_{\text{МНК}}$) происходит увеличение погрешности в $\mu(\boldsymbol{S})$ раз:

$$\frac{\|\delta\boldsymbol{z}\|}{\|\boldsymbol{z}\|} \leqslant \mu(\boldsymbol{S})\frac{\|\delta\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$$

Помимо вычислительной, у проблемы мультиколлинеарности есть и статистическая сторона. Напомним, что дисперсия оценки $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ вычисляется по формуле (6):

$$\mathrm{D}\hat{\boldsymbol{\beta}}_{\text{МНК}} = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}.$$

Норма полученной матрицы увеличивается с увеличением коллинеарности ($\|(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\|_2 = \lambda_{min}^{-1}$), что свидетельствует об увеличении дисперсии оценки $\hat{\boldsymbol{\beta}}_{\text{МНК}}$. Это, в свою очередь, автоматически приводит к увеличению MSE, что негативно сказывается на точности и предсказательной силе.

Таким образом, проблема мультиколлинеарности приводит как к проблеме вычислительной неустойчивости, так и к увеличению дисперсии у полученной оценки. Существует несколько методов борьбы с этой проблемой:

- Уменьшение числа признаков (отбор признаков) (см. 6);
- Регуляризация (см. 5);
- Преобразование признаков (Анализ главных компонент и т.п., в данном докладе не рассматривается).

В следующих разделах мы подробнее рассмотрим другие модификации и методы регрессии, и увидим, что они в той или иной мере помогают решить эту и остальные приведённые выше проблемы МНК-оценок.

5 Регуляризация в линейной регрессии

Как уже говорилось ранее, одним из главных критериев качества оценки является MSE, разложение которого имеет вид (7)

$$\mathrm{E}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^2 = \underbrace{\mathrm{D}\hat{\boldsymbol{\beta}}}_{\text{дисперсия}} + \underbrace{(\mathrm{E}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2}_{\text{смещение}}.$$

[9] любой учебник по вычислительным методам и теории приближённых вычислений в линейной алгебре

И так как мы рассмотрели МНК-оценку, которая имеет наименьшее смещение, но *не гарантирует* минимизацию всего MSE, можно попробовать допустить смещение и надеяться на то, что за счёт уменьшения дисперсии MSE удастся уменьшить.

Модель остаётся той же, то есть остаёмся в контексте линейной регрессии и предполагаем выполнение условий 3:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

5.1 Гребневая регрессия

Возьмём оценку по МНК и сделаем её смещённой, добавив к матрице $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ слагаемое $\lambda\mathbf{I}$, где $\lambda > 0$ — параметр, регулирующий смещение:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \quad \lambda > 0. \tag{9}$$

Если расписать данную оценку с помощью SVD, получим

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \sum_{i=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \lambda} U_j (V_j^{\mathrm{T}}\boldsymbol{y}).$$

За счёт положительного параметра $\lambda > 0$ получается отделить знаменатель от нуля. То есть устойчивость вычислений повышается.

Что происходит с дисперсией оценки? Исходя из вида коэффициента

$$\frac{\sqrt{\lambda_j}}{\lambda_j + \lambda},$$

можно заметить, что параметром $\lambda > 0$ штрафуются все компоненты, но в особенности те, которые малы, что уменьшает дисперсию оценки.

На рис. 1 хорошо видно, как с изменением λ достигается компромисс между дисперсией и смещением. Однако такой график для реальных данных мы получить не можем, так как изначально неизвестно, чему равно смещение и дисперсия оценки. Чаще всего **выбор параметра** λ осуществляется на основании кросс-валидации, то есть минимизации ошибки на валидационной выборке (есть и теоретические оценки, но они зависят от многих параметров, поэтому обычно не используются).

Задача, явное решение которой записано в (9), называется задачей *гребневой регрессии* (ridge regression). Она может быть сформулирована двумя эквивалентными способами

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}(\mu) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_2^2, \quad \mu > 0, \tag{10}$$

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \text{ т.ч. } \|\boldsymbol{\beta}\|_2^2 \leqslant \lambda, \quad \lambda > 0. \tag{11}$$

Заметим, что λ и μ при равенстве не дают одинаковых задач: $\mu = \infty$ и $\lambda = 0$ приводят к нулевой оценке $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \mathbf{0}$, а $\mu = 0$ и $\lambda = \infty$ приводят к классической оценке по МНК.

По ссылке `https://www.desmos.com/calculator/3fp4awzeyp` можно посмотреть на геометрическое построение задачи (11) в случае, когда $n = p = 2$, то есть оси — значения двумерного вектора $\boldsymbol{\beta}$. На рисунке 3 — пример такого построения.

5.2 Lasso

По аналогии с задачей гребневой регрессии, которая может быть сформулирована как задача (11), можно рассмотреть такую задачу:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \text{ т.ч. } \|\boldsymbol{\beta}\|_1 \leqslant \lambda, \quad \lambda > 0. \tag{12}$$

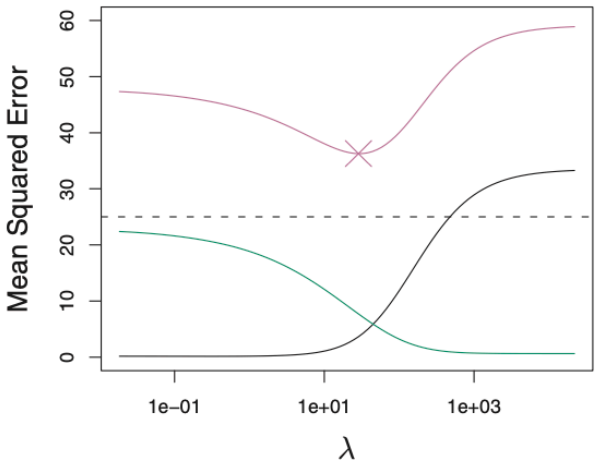

Рис. 1: Пример для сгенерированных данных. По оси x отложены значения параметра регуляризация λ. Зелёная кривая — дисперсия оценки, чёрная — её смещение, среднеквадратическая ошибка с обозначенным минимумом — красная.

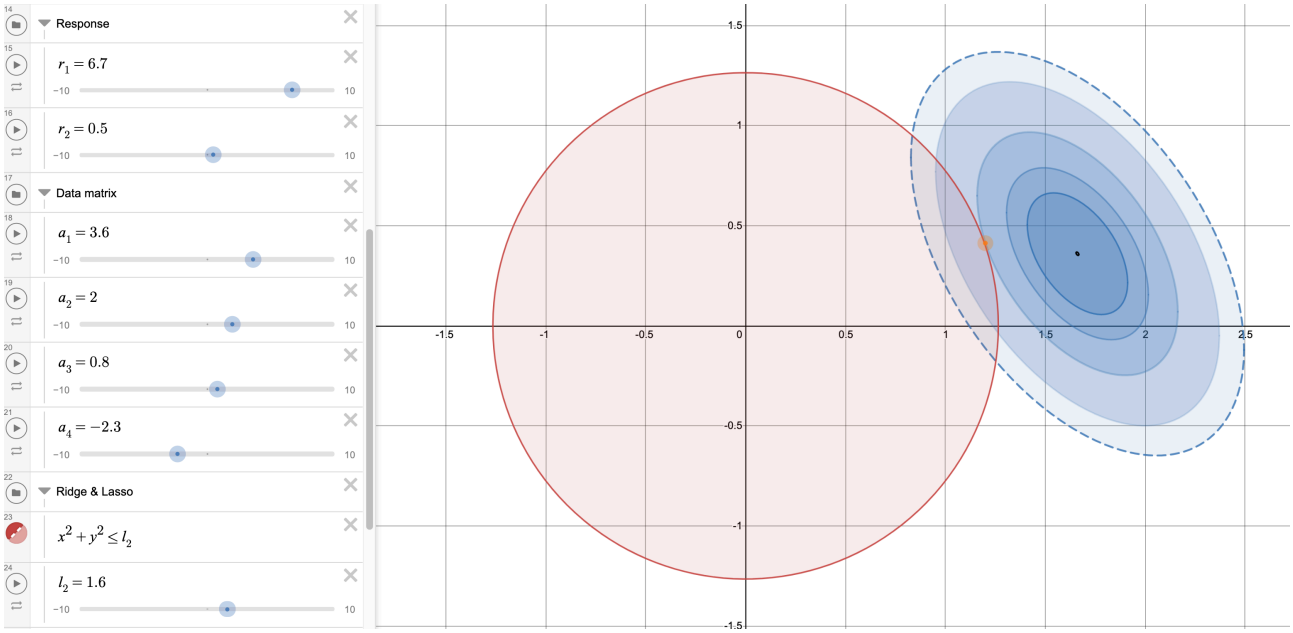

Рис. 2: Синим цветом показаны линии уровня целевой функции $z((\beta_1, \beta_2)^{\mathrm{T}}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, черная точка — её безусловный минимум, который достигается на МНК решении. Однако это не будет решением по методу гребневой регрессии: теперь искать решение мы можем только внутри красного круга, радиус которого определяется параметром λ. Исходя из рисунка, минимум целевой функции при заданном ограничении достигается в оранжевой точке.

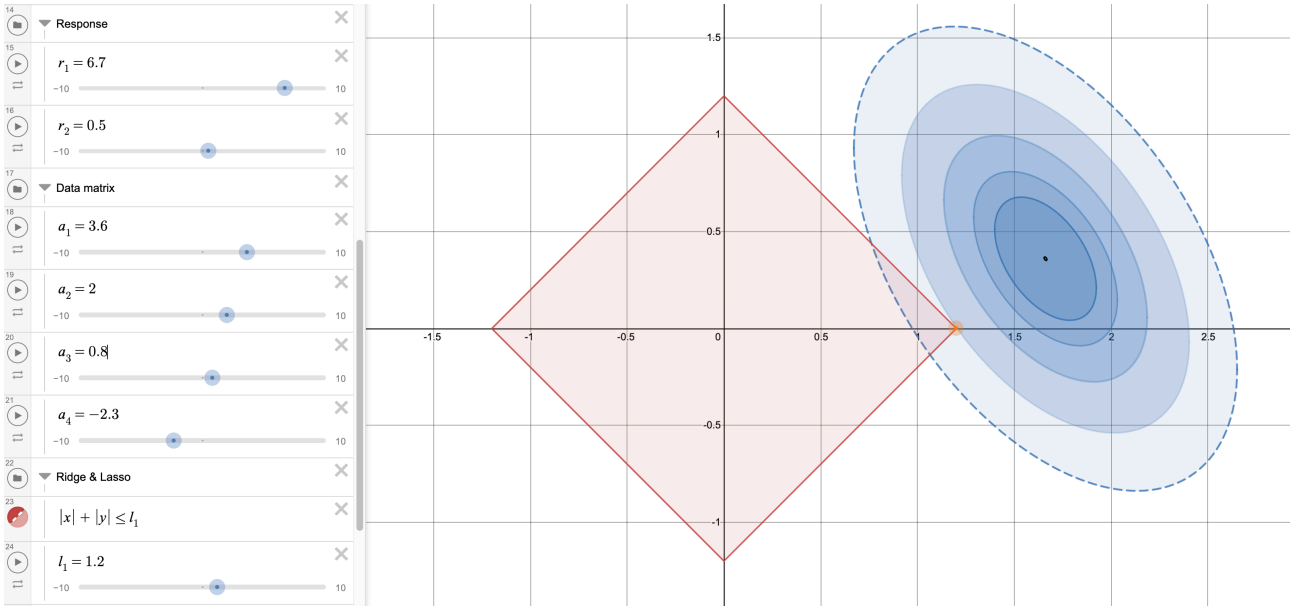

Рис. 3: Синим цветом показаны линии уровня целевой функции $z((\beta_1, \beta_2)^{\mathrm{T}}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, черная точка — её безусловный минимум, который достигается на МНК решении. Решение по методу lasso показано оранжевой точкой: видно, что благодаря «острому» множеству ограничений, приходим к решению, у которого одна из координат — нулевая. Решение по методу lasso отмечено оранжевой точкой.

Как можно видеть, изменилась только норма у ограничения $\|\boldsymbol{\beta}\|$; теперь рассматривается ограничение не на сумму квадратов, а на сумму модулей. Это влияет на вид множества ограничений: если раньше оно представляло из себя шар (см. рис 3), то теперь это ромб (в двумерном случае), размер которого также увеличивается с увеличением параметра λ.

Благодаря геометрическим особенностям нового множества ограничений $\|\boldsymbol{\beta}\|_1 \leqslant \lambda$ (см. двумерный пример на рис. 3), в результате решения задачи получаем вектор $\hat{\boldsymbol{\beta}}_{\text{lasso}}$, у которого некоторые компоненты равны нулю. С увеличением λ некоторые координаты перестают быть нулевыми, а при $\lambda = +\infty$ получаем классическое МНК-решение.

Особенности $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ Чаще всего отмечают два главных достоинства метода lasso:

1. Уменьшение MSE;
2. Увеличение интерпретируемости модели.

Первое обусловлено тем, что опять же рассматриваем смещённую оценку, а наилучшее значение MSE достигается на компромиссном значении λ.

Второе: имеется ввиду простота итоговой модели. Объясняется обнулением некоторых координат $\hat{\boldsymbol{\beta}}_{\text{lasso}}$: получается, что количество признаков, влияющих на ответ в полученной модели контролируется с помощью параметра λ. При небольших значениях λ и, соответственно, малом количестве значимых признаков (тех, коэффициент при которых не равен нулю), модель становится легко интерпретируемой. При малых λ будем иметь большее смещение, так что обычно параметр всё равно выбирается с помощью кросс-валидации, а не выбором «удобного» для интерпретации значения λ.

У данного метода так же есть и лимитации. Несмотря на то, что благодаря введённому ограничению (и, как следствие, уменьшению числа значимых признаков) появляется возможность рассмотрения случая $p > n$, адекватной оценки может не получится. Большинство теоретических результатов относительно сходимости решений lasso к истинному

значению $\boldsymbol{\beta}$ основаны на том, что **сам вектор $\boldsymbol{\beta}$ является разреженным**, то есть имеет только небольшую долю ненулевых координат (например, задача с $n = 100, p = 40000$, а число значимых признаков составляет порядка 10 признаков; ясно, что ни одна из данных ранее задач не привела бы к адекватному решению, но есть шанс, что lasso с этой задачей справится).

Особенности нахождения оценки В отличие от МНК- и Ridge-оценок, явного решения у задачи lasso нет. В то же время, благодаря возможности переформулировки данной задачи в задачу квадратичного программирования,[10] у неё есть очень быстрая реализация в пакете `glmnet` — решение для отдельного λ вычисляется со скоростью, сопоставимой со скоростью вычисления МНК-оценки; также очень быстро вычисляется множество решений на сетке значений λ (если полученное решение соответствует началу или концу построенной сетки, следует сдвинуть границы — может оптимальное решение окажется там).

6 Отбор признаков

Как отмечалось в предыдущем разделе, в результате применения метода lasso получается вектор коэффициентов с большим количеством нулей, что приводит к итоговой модели с малым числом признаков. По сути, осуществляется процедура *отбора признаков*. В этом разделе мы обсудим критерии выбора модели, а также ещё несколько подходов к решению задачи отбора признаков для модели.

6.1 Критерии выбора модели

Часто происходит так, что в рассмотрение берётся некоторый конечный набор моделей, а от статистика требуется выбрать наилучшую в некотором смысле модель.[11] Выбор критериев зависит от предположений и конечной цели.

Итак, предположим, что есть некоторое семейство построенных моделей $\{\mathcal{M}_i\}_{i \in I}$. Хотим выбрать лучшую модель $\mathcal{M}^*$ для предсказания. Перечислим далее некоторые подходы к решению этой задачи.

- Кросс-валидация:
  - **leave-one-out CV**

    Строим модель для всех элементов выборки, кроме одного. Считаем на нём ошибку предсказания. Проделываем то же самое для каждого элемента выборки. Берём среднее по всем ошибкам.
  - **k-fold CV**

    Пример для $k = 5$: случайным образом разбиваем выборку на 5 частей. Строим модель по четырём частям и считаем ошибку на пятой нетронутой. Делаем так для 5 возможных комбинаций (1,2,3,4 и 5; 1,2,3,5 и 4; ...) и считаем среднюю ошибку по всем «фолдам».
- Информационные критерии и R^2

[10] и многим другим улучшениям, о которых здесь не будем распространяться, смотрите документацию и статьи к библиотеке `glmnet` на R или Python

[11] например, модель линейной регрессии со всеми признаками или же модель, где признаков меньше и они выбраны на основании мнения экспертов

– **AIC и BIC**

Два стандартных критерия, которые используются не только в регрессионных моделях.

Пусть $\mathcal{L}(\boldsymbol{X}; \mathcal{M}_i)$ — максимум (по параметрам распределения) функции правдоподобия для модели $\mathcal{M}_i$, p_i - число параметров в модели i. Тогда

$$\text{AIC}_i = 2p_i - 2\ln \mathcal{L}(\boldsymbol{X}; \mathcal{M}_i);$$
$$\text{BIC}_i = p_i \ln n - 2\ln \mathcal{L}(\boldsymbol{X}; \mathcal{M}_i).$$

Они представляют из себя функцию правдоподобия выборки с поправкой-штрафом, зависящей от числа параметров и размера выборки. Исходя из вида критериев, чем меньше значение, тем лучше; выбирается модель, у которой BIC или AIC наименьший.[12]

В BIC штраф за число параметров в модели больше. Так как в формуле участвует функция правдоподобия $\mathcal{L}(\boldsymbol{X}; \boldsymbol{\theta})$, мы должны принять некоторое предположение о распределении выборки.

– R^2

С увеличением числа признаков коэффициент детерминации R^2 только увеличивается, поэтому может привести к переобученной модели.

– **adj.**R^2

Здесь уже накладывается некоторый штраф за размерность пространства параметров.

6.2 Отбор признаков в регрессии

В разделе про методы регуляризации 5 уже осветили один из способов снижения размерности пространства признаков — lasso. В этом разделе кратко расскажем о классических методах: best subset selection, а также forward- и backward- subset selection.

Best subset selection Если имеется p признаков, наивный вариант — рассмотреть все возможные модели с $\tilde{p} = 1$ признаком, $\tilde{p} = 2$, и так далее до $\tilde{p} = p$, а затем выбрать наилучшую с помощью критериев из предыдущего раздела, 6.1. Количество таких моделей будет равно 2^p. Если для примера взять $p = 20$, получим, что $2^p = 1,048,576$. Это уже довольно большое число моделей. При $p > 40$ данный подход становится затруднительным даже для построения МНК-оценок.

Также заметим, что из-за рассмотрения большого числа моделей, применение метода best subset selection может привести к проблеме переобучения (происходит подгонка модели под тренировочную выборку).

Forward и backward subset selection Также существуют и «жадные» альтернативы методу best subset selection. Один из вариантов (*Forward subset selection*) состоит в выборе наилучшей модели с одним признаком, а затем последовательное добавление признаков, которые оказывают наилучшее влияние на критерий выбора. В итоге получаем $p(p+1)/2$ моделей.[13] Далее можем выбирать на основании желаемого числа признаков или опять же на основании тех же критериев.

[12] нельзя забывать, что полученные значения зависят от выборки, то есть представляют из себя случайные величины; если значения AIC и BIC близки, то однозначного вывода отсюда сделать не получится

[13] для $p = 20$ получаем 210 моделей — значительно меньше, чем у best subset

Аналогично можно начинать со всех признаков и последовательно удалять по одному, пока не придём к модели с одним признаком (*Backward subset selection*). Замечание: в случае, когда $p > n$ и считаются МНК-оценки (к примеру), метод Backward subset selection уже не сработает, так как нет возможности начать процедуру с полного пространства признаков.

7 Источники и рекомендуемая литература

- ESL (Elements of Statistical Learning) — Hastie, Tibshirani, Friedman;
- ISLR (An Introduction to Statistical Learning) — James, Witten, Hastie, Tibshirani;
- Лекции Н.Э. и А.И., СтатМод;
- Лекции Воронцова по ML;
- Лекции Larry Wasserman — Statistical Learning;
- All of Statistics — Larry Wasserman.