Регрессия, регуляризация, отбор признаков

Михаил Козак, Павел Мехнин, Данил Шкурат

5 октября 2022 г.

1 Обучение с учителем

Обучение с учителем — это направление машинного обучения, объединяющее алгоритмы и методы построения моделей на основе множества примеров, содержащих пары «известный вход — известный выход». Иными словами, чтобы алгоритм относился к обучению с учителем, он должен работать с примерами, которые содержат не только вектор независимых переменных (атрибутов, признаков), но и значение, которое должна выдавать модель после обучения (такое значение называется целевым). Разность между целевым и фактическим выходами модели называется ошибкой обучения (невязкой, остатками), которая минимизируется в процессе обучения.

2 Регрессия

Рассмотрим задачу обучения с учителем, частным случаем которой является задача регрессии.

$\mathbf{X}$ — множество объектов, заданное своими признаками (точки p-мерного пространства)

$\mathbf{Y}$ — множество ответов (действительные числа)

Предполагаем наличие неизвестной зависимости между объектами $\mathbf{x} \in \mathbf{X}$ и ответами $y \in \mathbf{Y}$:

$$y = f^*(\mathbf{x}) + \boldsymbol{\varepsilon}$$

$(\mathbf{x}_i, y_i)_{i=1}^n$ — обучающая выборка (пары объект-известный ответ), случайным образом выбранная из генеральной совокупности.

Задача:

На основе обучающей выборки найти $\hat{f}$, такую что $y \approx \hat{f}(\mathbf{x})$ для любого наблюдения $(\mathbf{x}, y)$, то есть восстановить зависимость, способную для любого объекта выдать достаточно точный ответ.

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.

2.1 Задача регрессии как задача оптимизации

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

$$y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

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

Пусть задана модель регрессии — параметрическое семейство функций $f(\mathbf{x}, \boldsymbol{\beta})$, где $\boldsymbol{\beta} \in \boldsymbol{B}$ — вектор параметров модели, $\boldsymbol{B} \subset \mathbb{R}^p$ — пространство параметров, $f : \mathbb{R}^p \times \boldsymbol{B} \to \mathbb{R}$ — фиксированная функция.

Выберем в качестве функционала качества Q аппроксимации целевой зависимости на выборке X_n среднеквадратическую ошибку:

$$\text{MSE}_{\text{train}} = Q(\boldsymbol{\beta}, X_n) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i, \boldsymbol{\beta}) - y_i)^2. \tag{1}$$

Обучение по методу наименьших квадратов (МНК) состоит в нахождении такого вектора параметров $\hat{\boldsymbol{\beta}}$, при котором достигается минимум среднего квадрата ошибки на заданной обучающей выборке X_n:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} Q(\boldsymbol{\beta}, X_n).$$

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 ($\text{MSE}_{\text{test}} \gg \text{MSE}_{\text{train}}$), говорят об эффекте переобучения (overtraining) или переподгонки (overfitting).

3 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;
- $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ — матрица данных (детерминированная — для простоты предполагаем, что случайность в модели происходит только от вектора шума);
- $\boldsymbol{\beta} \in \mathbb{R}^p$ — 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:

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

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

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

Задача оптимизации (с квадратичной функцией потерь):

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

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):[1]

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

Математическое ожидание полученной оценки:

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

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

Ковариационная матрица полученной оценки:

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

Теорема Гаусса–Маркова утверждает, что $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ имеет наименьшую дисперсию среди всех несмещённых оценок (best linear unbiased estimate — BLUE).

Наличие явного вида решения крайне удобно в вычислительном плане. Оценка вычисляется достаточно быстро посредством применения сингулярного разложения матрицы данных $\boldsymbol{X}$.

3.1 Вычисление МНК-оценки: 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,[2] $\boldsymbol{D}$ — 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 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$;
- $\boldsymbol{D} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, $\lambda_j \geqslant 0$ — 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

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{МНК}} = \\ = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \\ = (\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{D}^{-2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \\ = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y},\end{aligned} \tag{7}$$

где $\boldsymbol{D}^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})$.

$$\hat{\boldsymbol{\beta}}_{\text{МНК}} = \sum_{j=1}^{p}\frac{1}{\sqrt{\lambda_j}}U_j(V_j^{\mathrm{T}}\boldsymbol{y}) \tag{8}$$

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.

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

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

3.2 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 вычисления $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ близки к нулю. Поэтому в суммах появляются очень большие и неустойчивые слагаемые.

Теряется интерпретируемость оценок коэффициентов, так как коэффициенты могут неоправданно принимать очень большие значения.

Высокая дисперсия $\hat{\boldsymbol{\beta}}_{\text{МНК}}$ приводит к высокой MSE.

Ответы на контрольной выборке неустойчивы (переобучение).

Способы решения проблемы:

- Регуляризация: проблема зарождается в мультиколлинеарности, а проявляется в том, что норма вектора коэффициентов увеличивается. Регуляризация контролирует увеличение нормы вектора.
- Преобразование признаков (feature extraction, feature engineering). Ещё одно решение проблемы мультиколлинеарности заключается в том, чтобы подвергнуть исходные признаки некоторому функциональному преобразованию, гарантировав линейную независимость новых признаков, и, возможно, сократив их количество, то есть уменьшив размерность задачи. В методе главных компонент (principal component analysis, PCA) строится минимальное число новых признаков, по которым исходные признаки восстанавливаются линейным преобразованием с минимальными погрешностями.
- Отбор признаков (feature selection).

4 Регуляризация

Хорошая оценка $\hat{\boldsymbol{\beta}}$ должна иметь низкую среднеквадратическую ошибку

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

Несмещенная МНК-оценка не гарантирует минимизацию всей MSE. Когда матрица $\boldsymbol{X}$ близка к вырожденной (это может произойти из-за наличия мультиколлинеарности или когда число предикторов p почти равно числу наблюдений n), дисперсия $\hat{\boldsymbol{\beta}}$ становится большой и MSE_{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 MSE_{test}.

4.1 Гребневая регрессия (Ridge regression)

4.1.1 Задача гребневой регрессии

Вводится штраф за увеличение нормы вектора $\boldsymbol{\beta}$ и минимизируется следующая функция:

$$Q_\tau(\boldsymbol{\beta}) = ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}||^2 + \tau||\boldsymbol{\beta}||^2 \to \min_{\boldsymbol{\beta}},$$

где τ — неотрицательный параметр регуляризации.

В развернутом виде задача оптимизации записывается так:

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

Решение задачи гребневой регрессии:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \tau\boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

Подход на основе сингулярного разложения позволяет подбирать параметр τ, вычислив SVD только один раз.

Решение гребневой регрессии через SVD:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \sum_{j=1}^{p}\frac{\sqrt{\lambda_j}}{\lambda_j + \tau}U_j(V_j^{\mathrm{T}}Y).$$

4.1.2 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.

Таким образом, необходимо один раз произвести сингулярное разложение матрицы $\boldsymbol{X}$, 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.

4.1.3 Подбор параметра τ

Скользящий контроль:

- выбираем сетку значений τ;
- вычисляем ошибку кросс-проверки для каждого значения τ;
- выбираем τ с наименьшим значением ошибки кросс-проверки;
- перестраиваем модель со всеми наблюдениями с выбранным значением τ.

Эвристика

Скользящий контроль — вычислительно трудоёмкая процедура. Известна практическая рекомендация брать брать τ в отрезке $[0.1, 0.4]$, если столбцы матрицы $\boldsymbol{X}$ заранее стандартизованы.

4.1.4 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.

4.2 Лассо (Lasso)

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

4.2.1 Задача Lasso-регрессии

Метод LASSO решает следующую задачу минимизации:

$$||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \tau||\boldsymbol{\beta}||_1 \to \min_{\boldsymbol{\beta}},$$

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 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.

4.3 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 $\hat{\boldsymbol{\beta}}$. 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.

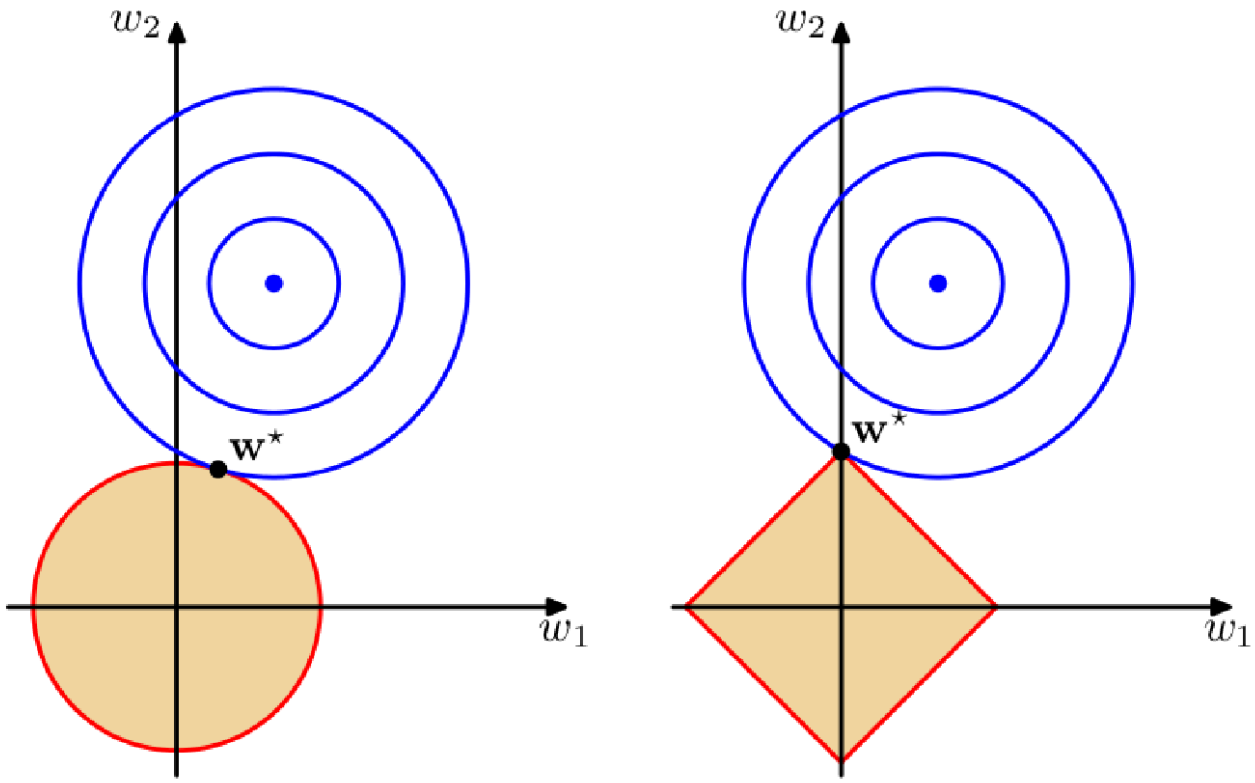

Рис. 1: Синие линии уровня функционала качества (синяя точка — безусловный минимум, который достигается на МНК решении). Оранжевая зона — ограничения, задаваемые L2 и L1-регуляризаторами. Чёрная точка — минимум целевой функции при заданном ограничении.

Замечания:

- Оба метода успешно решают проблему мультиколлинеарности
- Гребневая регрессия использует все признаки

- Лассо производит отбор признаков, что предпочтительнее, если среди признаков есть шумовые или измерения признаков связаны с ощутимыми затратами.
- С помощью кросс-валидации можно определить какой подход лучше для конкретных данных.

4.4 Elastic net regularization

Решается задача оптимизации

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \tau_1||\boldsymbol{\beta}||_1 + \tau_2||\boldsymbol{\beta}||_2^2 \to \min_{\boldsymbol{\beta}}.$$

- 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 (регрессионная модель обладает лучшей предсказательной способностью), 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.
- Если количество признаков p больше, чем количество наблюдений n, Lasso выберет не более n ненулевых предикторов (даже если все p предикторов актуальны), поэтому в случае многомерных данных с малым числом наблюдений предпочтительней использовать Elstic net.
- Elastic net можно свести к SVM, для которого разработано много быстрых решений.

5 Отбор признаков

Меньшее число признаков улучшает интерпретируемость модели и уменьшает время обучения, поэтому целесообразно выбрать модель, имеющую хорошую предсказательную способность при относительно небольшом числе признаков. Модели обычно сравниваются с помощью информационных критериев AIC, BIC (представляют из себя функцию правдоподобия выборки с поправкой-штрафом, зависящей от числа параметров), скорректированного коэффициента детерминации adj.R^2 (доля объясненной дисперсии с некоторым штрафом за размерность пространства параметров).

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*. Рассмотрим ещё несколько подходов к решению задачи отбора признаков: 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. 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. Например, для $p = 20$ получаем 210 моделей — значительно меньше, чем у best subset. Далее можем выбирать на основании желаемого числа признаков или опять же на основании тех же критериев.

Аналогично можно начинать со всех признаков и последовательно удалять по одному, пока не придём к модели с одним признаком (*Backward subset selection*). 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.

6 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 Соколова (ФКН ВШЭ);
- Лекции Larry Wasserman — Statistical Learning;
- All of Statistics — Larry Wasserman.
- https://ml-handbook.ru/