Обучение с учителем. Регрессия. Регуляризация в регрессии. Разные подходы. Feature extraction and Feature selection

Михайлов Дмитрий и Смирнов Иван

19 сентября 2023 г.

Содержание

1 Про обучение с учителем

Линейную регрессию мы уже использовали ранее. При вызове функции мы передавали модель и обучающий датасет, для каждого индивида которого уже было определено значение, которое мы хотим предсказать. Такой вид обучения модели называется "обучение с учителем". По сути таковым является любой алгоритм машинного обучения, где мы выдаем алгоритму как ему нужно с нашей точки зрения правильно поступить.

2 Регрессия

Пусть имеем матрицу $\mathbf{X}$ состоящую из n индивидов $\mathbf{x}_i \in \mathbb{R}^p$. Предполагаем, что имеем функциональную завимость между $\mathbf{X}$ и вектором ответов $\mathbf{y}$, состоящего из y_i:

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

Где ϵ_i — ошибка обучения или остаток. Предполагаем что они независимые, что $\mathbb{E}\varepsilon = 0$, и $\mathbb{D}\varepsilon = \sigma^2$.

Задача: найти $\hat{f}$ (оценку функции регрессии f), чтобы для каждого индивида $y_i \approx \hat{f}(\mathbf{x}_i)$.

2.1 Задача оптимизации

Пусть задана модель регрессии — параметрическое семейство функций $f(\mathbf{x}, \boldsymbol{\beta})$, где $\boldsymbol{\beta} \in \mathbb{R}^p$ — вектор параметров модели.

Один из способов вычислить значения параметров модели является метод наименьших квадратов (МНК / OLS), который минимизирует среднеквадратичную ошибку между реальным значением зависимой переменной и прогнозом, выданным моделью.

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} MSE(\boldsymbol{\beta}) \tag{2}$$

MSE — mean squared error, выводится из (1):

$$MSE(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2 \tag{3}$$

2.2 Случай линейной регрессии

В модели линейной регрессии f представляет собой линейную функцию, в матричном виде это:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

Подставим (4) в (3):

$$MSE(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5}$$

Решение при подстановке в (2):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y}, \qquad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{6}$$

Где $\boldsymbol{X}^{-}$ — обобщённо-обратная матрица.

Матрица $\boldsymbol{X}^{-}$ называется обобщённо-обратной, если:

1. По аналогии с $\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{I} \to \boldsymbol{X}\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}$ и $\boldsymbol{X}^{-1}\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{X}^{-1}$ выполняется $\boldsymbol{X}^{-}\boldsymbol{X}\boldsymbol{X}^{-} = \boldsymbol{X}^{-}$, $\boldsymbol{X}\boldsymbol{X}^{-}\boldsymbol{X} = \boldsymbol{X}$

2. (Псевдообратная по Муру-Пенроузу) если по аналогии с $\boldsymbol{X}^{-1} = \boldsymbol{X}^T \to (\boldsymbol{X}^{-1}\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^{-1})^T = \boldsymbol{X}^{-1}\boldsymbol{X}$, дополнительно выполняется $\boldsymbol{X}^{-}\boldsymbol{X} = (\boldsymbol{X}^{-}\boldsymbol{X})^T, \boldsymbol{X}\boldsymbol{X}^{-} = (\boldsymbol{X}\boldsymbol{X}^{-})^T$

У обобщенно-обратных матриц есть следующие важные для нас свойства:

1. Если столбцы $\boldsymbol{X}$ линейно-независимы, то существует $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ и $\boldsymbol{X}^{-} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$

2. Пусть ищем решение $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ относительно $\boldsymbol{\beta}$, если решение не единственно, то $\boldsymbol{\beta} = \boldsymbol{X}^{-}\boldsymbol{y}$ есть решение с минимальной нормой.

Вычислительная проблема: при плохой обусловленности матрицы $\boldsymbol{X}^T\boldsymbol{X}$ вычисление обратной к ней матрицы крайне нежелательно. Поэтому на практике лучше избегать прямого использования формул (6).

2.3 МНК и SVD

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

$$\mathbf{X}^{-} = \mathbf{U}\mathbf{D}^{-1}\mathbf{V}^T = \sum_{j=1}^{p}\frac{1}{\sqrt{\lambda_j}}U_jV_j^T.$$

Вектор МНК-решения:

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^{-}\boldsymbol{y} = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j (V_j^{\mathrm{T}} \boldsymbol{y}). \tag{7}$$

Оценка $\boldsymbol{y}$:

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \sum_{j=1}^{p} V_j (V_j^{T}). \tag{8}$$

Норма вектора коэффициентов:

$$||\hat{\boldsymbol{\beta}}||^2 = \sum_{j=1}^{p} \frac{1}{\lambda_j} (V_j^{\mathrm{T}} \boldsymbol{y})^2. \tag{9}$$

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

3 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 $\mathrm{MSE}_{\mathrm{test}}$ 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 $\mathrm{MSE}_{\mathrm{test}}$.

Регуляризация — метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение, которое возникает, когда наша модель

сильно подстраивается под данные, мы боремся с этим добавляя некоторый штраф за большие значения коэффициентов у линейной модели. Тем самым запрещаются слишком "резкие"изгибы, и предотвращается переобучение.

Соответственно, необходимо добавить в целевую функцию штраф за слишком большие значения параметров. Наиболее часто используемые виды регуляризации — L_1 (lasso regularization или регуляризация через манхэттенское расстояние) и L_2 (ridge regularization или регуляризация Тихонова).

3.1 Ridge

Добавляем к матрице $\boldsymbol{X}^T\boldsymbol{X}$ параметр L_2 регуляризации τ:

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \tau\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{10}$$

Пользуясь SVD:

$$\hat{\boldsymbol{\beta}}_{ridge} = \mathbf{X}^{-}\boldsymbol{y} = \sum_{j=1}^{p} \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_j} + \tau} U_j(V_j^{\mathrm{T}}\boldsymbol{y}). \tag{11}$$

За счет положительного параметра $\tau > 0$ при малых λ мы можем отделить знаменатель от нуля. Значение параметра τ выбирают на основании кросс-валидации (минимизируя ошибку на валидационной выборке).

Задача в явном виде записывается как:

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \qquad ||\boldsymbol{\beta}||_2^2 \leq \tau \tag{12}$$

3.2 Lasso

В Lasso используется L_1-регуляризатор. Задача в явном виде записывается как:

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \qquad ||\boldsymbol{\beta}||_1 \leq \tau \tag{13}$$

Lasso чаще всего обнуляет значения некоторых параметров, что в случае с линейными моделями приводит к отбору признаков.

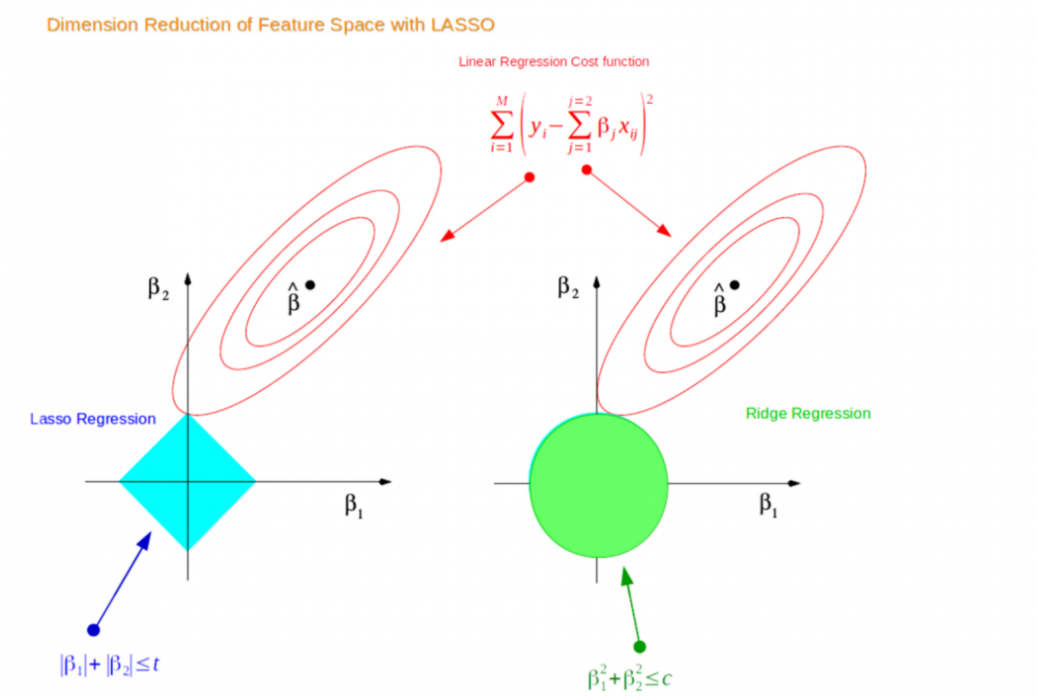

Figure 3: Why LASSO can reduce dimension of feature space? Example on 2D feature space. Modified from the plot used in 'The Elements of Statistical Learning' by Author.

Рис. 1: Сравнение lasso и ridge в случае линейной регрессии с двумя признаками

В случае обычной регрессии без штрафов точка $\hat{\boldsymbol{\beta}}$ была бы решением задачи, но ограничение на слишком большие коэффициенты не дает нам этого сделать, вместо этого берется точка пересечения исходной модели с плоскостью ограничений. Тут мы видим, что у lasso точка пересечения $(0; \beta_2^{lasso})$, а у ridge $(\beta_1^{ridge}; \beta_2^{ridge})$, соответственно lasso позволяет нам таким образом упростить модель и повысить её интерпретируемость.

Однако могут быть крайние случаи в которых lasso не обнулит коэффиценты, а ridge занулит, поэтому оптимальный метод выбиается на основе кросс-валидации на конкретных данных.

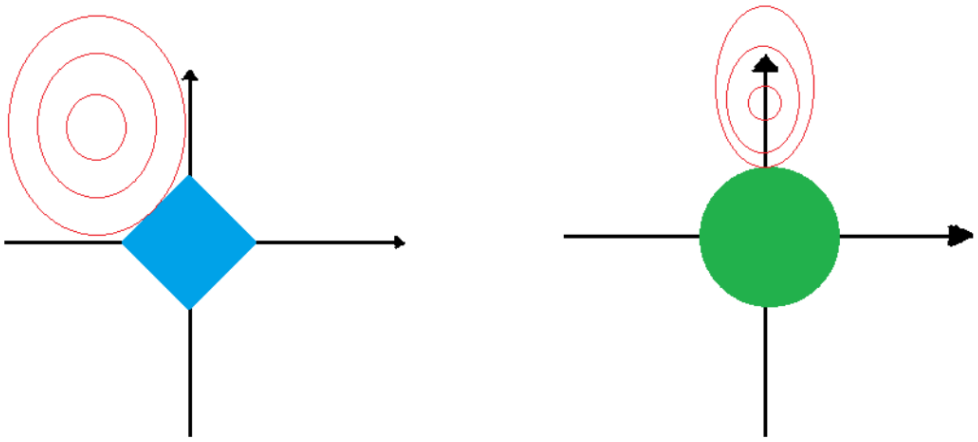

Рис. 2: Случай, когда ridge лучше lasso

4 Работа с признаками

4.1 Feature Extraction. Анализ главных компонент

Отличительной особенностью большинства наборов данных является то, что они содержат большое количество признаков, и, кроме того, эти признаки требуют много вычислительных ресурсов для их обработки. В этом случае их извлечение может быть полезно при выборе конкретных признаков, а также при объединении некоторых связанных признаков, что может уменьшить объем данных.

Мы ранее уже пользовались АГК для уменьшения размерности матрицы данных, и мы также можем применить это к регрессии, чтобы улучшить модель. В методе главных компонент строится минимальное число новых признаков, по которым исходные признаки восстанавливаются линейным преобразованием с минимальными погрешностями. Этот метод также можно использовать для обнаружения аномалий и выбросов, что также улучшает качество регрессии.

4.2 Feature Selection. Информационные критерии

Как отмечалось ранее, 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*.

AIC и BIC также используются при отборе признаков. Эти критерии показывают, какая модель лучше, но не утверждают, что лучшая модель является правильной.

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

$$\mathrm{AIC}_i = 2p_i - 2\ln \mathcal{L}(\boldsymbol{X}; \mathcal{M}_i);$$

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

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.