Классическое машинное обучение

Минкин Даниэль

9 февраля 2025 г.

Содержание

1 Линейные модели

Линейные модели — класс моделей которые используют линейное преобразование для вектора входных фичей.

1.1 Регрессия

Формализуем задачу регрессии, пусть у нас есть вектор $\bar{x} \in \mathrm{R}^n$. Тогда предсказание может быть сделано с помощью такой формулы:

$$y_{pred} = \bar{x} \cdot \bar{w} + w_0 \tag{1}$$

Т.е мы ищем такой вектор $\bar{w} \in \mathrm{R}^{n+1}$, который будет выдавать наиболее близкие y_{pred} к y_{true}.

1.2 Классификация

В случае решения задачи классификации через линейные модели сначала делается предсказание как в случае регрессии, а после к результату предсказания применяется разделяющее правило (например в случае бинарной классификации правило может задаваться так: если $y_{pred} > 0$ мы относим объект к положительному классу — если нет, то к отрицательному)

В случае бинарной классификации с разделяющим правилом из примера уравнение регрессии задает гиперплоскость, которая разделяет исходное

пространство: i.e $\sum_{i=1}^{n} w_i \cdot x_i + c > 0$ — плоскость в n-мерном пространстве, которая делит пространство на положительный и отрицательные классы

1.3 Общее

Несколько фактов:

- **При использовании OneHot Encoding-а мы можем избавиться от одной encoded фичи**. Все просто: пусть у нас есть веса для каждой закодированной фичи $w_1, w_2 ... w_n$, также у нас добавляется константа c к предсказанию по фичам. Давайте удалим последнюю фичу, тогда в случае если $x_1 = 0$ ... $x_{n-1} = 0$ нам нужно добавить к константе еще и w_n, иначе результат изменится, следовательно константа в новой модели должна быть равна $w_n + c$. Однако тогда нам нужно внести поправку в веса в случае если один из $x_i \mid i < n$ равен 1, чтобы результат остался таким же. Мы можем просто вычесть из старых весов w_n, за счет того, что мы добавили его к const ничего не изменится. Таким образом мы успешно исключили одну encoded фичу, оставив результаты предсказаний без изменений. **Важно заметить, что если мы работаем в модели без константы, то тогда данный подход не сработает**
- **Для более сложных зависимостей необходимо использовать новые фичи которые являются функциями от старых**. Т.е мы включаем в модель фичи задаваемые как $f(x_1, x_2 ... x_n)$
- Если между признаками есть приближённая линейная зависимость, коэффициенты в линейной модели могут совершенно потерять физический смысл

1.4 Оценка по МНК

1.4.1 Общее

В первую очередь опустим свободный член, так как можно считать, что у нас просто есть доа признак, который всегда равен 1. Пусть функция потерь задается как Евклидова норма между предсказанными и истинными значениями. I.e мы решаем следующую задачу:

$$\|Xw - b\|_{euclidean} \to \min_{w} \tag{2}$$

где X — матрица размера (N, k), N – размер выборки, а k — кол-во фичей, т.е это матрица где в строки записаны вектора, по которым нужно сделать предсказания.

Однако нам также нужно сделать поправку на размер выборки, чтобы значения функции потерь можно было сравнивать между собой для разных выборок. Получается задача выгладит так:

$$\frac{\|Xw - b\|_{euclidean}}{N} \to \min_{w} \tag{3}$$

Функция потерь является функционалом, так как принимает на вход три значения — матрицу наблюдений, true значения предсказываемой перменной и функцию, которая возвращает некое значение по вектору наблюдений

Значение коэффициентов может быть получено через псевдообратную матрицу, по ее свойству:

$$\|XX^{+}b - b\|_{euclidean} \leq \|Xw - b\|_{euclidean} \tag{4}$$

для любого w

Так как деление на константу — монотонное преобразование, данное решение минимизирует нашу функцию потерь

> **Важно заметить:** псевдообратная матрица может быть использована и для других норм, однако тогда нам нужно перейти в евклидово пространство A с новым скалярным произведением, тогда псевдообратная матрица будет минимизировать норму задаваемую как $\sqrt{\langle u, u \rangle_A}$. Об этом будет рассказано позже

Таким образом решением данного СЛАУ будет

$$w_* = A^{+}b \tag{5}$$

В случае если $N \geq k$

$$w_* = (X^T X)^{-1} X^T b \tag{6}$$

А в противном случае, когда $N < k$

$$w_* = X^T (XX^T)^{-1} b \tag{7}$$

Мы можем считать, что X — матрица полного столбцового или строчного ранга, зачастую у нас не будет полностью ЛЗ столбцов или строк, а даже если они и есть их можно исключить из-за бессмысленности

1.4.2 Невырожденность матрицы $X^T X$

В реальных задачах матрицах $X^T X$ или XX^T являются невырожденными, однако нам нужно оценить насколько они "невырождены", так как у нас есть погрешность при вычислении детерминанта, например мы могли получить неотрицательное значение из-за логики работы чисел с плавающей точкой или же в изначальных данных содержится погрешность.

Число обусловленности

Число обусловленности — максимальное значение отношения относительного изменения функции и относительного измененииия арргмента

$$\mu(f,x) = \max_{\Delta x} \frac{\frac{\|f(x+\Delta x)-f(x)\|}{\|f(x)\|}}{\frac{\|\Delta x\|}{\|x\|}} \tag{8}$$

в малой окрестности Δx

Легко заметить, что чем больше значение $\mu(f,x)$, тем хуже так как наше решение может очень сильно измениться от малого приращения аргумента, а мы бы хотели иметь "стабильное" решение

Покажем на зависимость числа обусловленности задачи $f(A) = (A^TA)^{-1}$ от матрицы корреляции. Мы воспользуемся определением числа обусловленности:

$$\mu(f,A) = \max_{\Delta A} \frac{\frac{\|f(A+\Delta A)-(A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \tag{9}$$

Тут нам понадобится матричное дифференцирование, нам нужно найти дифференциал $(A^TA)^{-1}$, заметим, что это композиция двух функций. Во-первых, вспомним определение дифференциала, пусть задана функция $f : R^n \to R^m$ Тогда ее дифференциал может быть найден так:

$$f(x_0 + \Delta x) - f(x_0) = [D_{x_0} f](\Delta x) + o(\|\Delta x\|) \tag{10}$$

При этом дифференциал $[D_{x_0} f]$ — линейное отображение из R^n в R^m, т.е мы находим линейную часть приращения отображения

Найдем дифференциал $f(X) = X^TX$ где X - матрица (n,m) и $n > m$. Тогда:

$$\begin{aligned} f(X+\Delta X) - f(X) &= (X+\Delta X)^T(X+\Delta X) - X^TX \\ &= (X^T + (\Delta X)^T)(X+\Delta X) - X^TX \\ &= X^TX + X^T\Delta X + (\Delta X)^TX + (\Delta X)^T\Delta X - X^TX \\ &= X^T\Delta X + (\Delta X)^TX + (\Delta X)^T\Delta X \end{aligned}$$

Можно легко показать, что $[D_X f](\Delta X) = X^T\Delta X + (\Delta X)^TX$ является линейным оператором по ΔX. При этом $(\Delta X)^T\Delta X$ и есть $o(\|\Delta X\|)$.

Найдем дифференциал $g(X) = X^{-1}$. Мы будем использовать равенство $E = X^{-1}X$. Продифференцируем выражение с обеих сторон.

$$0 = [D_{X_0}(X^{-1}X)](H) \tag{11}$$

По правилам дифференцирования умножения мы получаем

$$0 = [D_{X_0}(X^{-1})](H) \cdot X_0 + X_0^{-1} \cdot [D_{X_0}(X)](H) \tag{12}$$

Таким образом

$$[D_{X_0}(X^{-1})](H) \cdot X_0 = -X_0^{-1} \cdot [D_{X_0}(X)](H) \tag{13}$$

Т.е все сводится к формуле

$$[D_{X_0}(X^{-1})](H) = -X_0^{-1} \cdot [D_{X_0}(X)](H) \cdot X_0^{-1} \tag{14}$$

При этом $[D_{X_0}(X)](H) = H$, таким образом итоговая формула имеет вид

$$[D_{X_0}(X^{-1})](H) = -X_0^{-1} \cdot H \cdot X_0^{-1} \tag{15}$$

Мы имеем дело с функцией $g(f(X))$, дифференциал такой функции представим как

$$[D_{f(X_0)}(g)]([D_{X_0}(f)](\Delta X)) \tag{16}$$

Следовательно $[D_{X_0}((X^TX)^{-1})](\Delta X)$, может быть найдено так

$$(X_0^TX_0)^{-1} \cdot (X_0^T\Delta X + (\Delta X)^TX_0) \cdot (X_0^TX_0)^{-1} \tag{17}$$

Таким образом в изначальной формуле числа обусловленности мы можем записать приращение функции как

$$\mu(A) = \max_{\Delta A} \frac{\frac{\|(A^TA)^{-1} \cdot (A^T\Delta A + (\Delta A)^TA) \cdot (A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \tag{18}$$

Оценим наше выражение сверху, сделаем несколько предположений на будущее, **пусть мы используем одно и тоже семейство норм для входного и выходного пространства и для этого семейства выполняется свойство субмультипликативности**, тогда для $(A^TA)^{-1}$ и $(A^T\Delta A + (\Delta A)^TA)$ можно будет использовать следующее мажорирование

$$\|(A^TA)^{-1}(A^T\Delta A + (\Delta A)^TA)(A^TA)^{-1}\| \leq \|(A^TA)^{-1}\|^2\|A^T\Delta A + (\Delta A)^TA\| \tag{19}$$

В итоге мы можем мажорировать наше выражение так:

$$\frac{\frac{\|(A^TA)^{-1} \cdot (A^T\Delta A + (\Delta A)^TA) \cdot (A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$\frac{\|(A^TA)^{-1}\| \cdot \|A^T\Delta A + (\Delta A)^TA\|}{\frac{\|\Delta A\|}{\|A\|}}$$

Продолжая мажорирование мы получим следующую ситуацию

$$\frac{\|(A^TA)^{-1}\| \cdot \|A^T\Delta A + (\Delta A)^T A\|}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$\frac{\|(A^TA)^{-1}\| \cdot (\|A^T\Delta A\| + \|(\Delta A)^T A\|)}{\frac{\|\Delta A\|}{\|A\|}}$$

Предположим, что норма выходного пространства устойчива к транспонированию, также вспомним условие про субмультипликативность, тогда наше выражение примет вид

$$\frac{\|(A^TA)^{-1}\| \cdot (\|A^T\Delta A\| + \|(\Delta A)^T A\|)}{\frac{\|\Delta A\|}{\|A\|}} =$$
$$2 \cdot \frac{\|(A^TA)^{-1}\| \cdot \|A^T\Delta A\|}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$2 \cdot \frac{\|(A^TA)^{-1}\| \cdot \|A^T\| \cdot \|\Delta A\|}{\frac{\|\Delta A\|}{\|A\|}} =$$
$$2 \cdot \frac{\|(A^TA)^{-1}\| \cdot \|A^T\|}{\frac{1}{\|A\|}} =$$
$$2 \cdot \|(A^TA)^{-1}\| \cdot \|A\|^2$$

Таким образом, при определенных требованиях к семейству норм мы получаем, что

$$\max_{\Delta A} \frac{\frac{\|(A^TA)^{-1} \cdot (A^T\Delta A + (\Delta A)^T A) \cdot (A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$2 \cdot \|(A^TA)^{-1}\| \cdot \|A\|^2$$

Вместо точного максимума мы можем использовать оценку сверху для всех значений.

Покажем влияние на данный результат близости матрицы к "вырожденной". Применим сингулярное разложение к матрице A, если она имеет размеры (n, m), где $n > m$

$$A = U_{(n,n)}\Sigma_{(n,m)}V^T_{(m,m)} \tag{20}$$

Σ — диагональная матрица где на диагонали находятся сингулярные числа матрицы, а U и V — унитарные матрицы, тогда A^TA равно
$V\Sigma^TU^TU\Sigma V^T = V\Sigma^TE\Sigma V^T$. Тогда:

$$(A^TA)^{-1} = (V\Sigma^TE\Sigma V^T)^{-1} \tag{21}$$

При этом $\Sigma^TE\Sigma$ — матрица (m, m), если матрица A имеет размер (n, m), где $n > m$. Так как матрица A имеет m сингулярных значений и является

матрицей полного столбцового ранга — матрица $\Sigma^T\Sigma$ является обратимой. Продолжим раскрывать наше выражение

$$(A^TA)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T \tag{22}$$

Мажорируем наше старое выражение, оно будет равно

$$2 \cdot \|(A^TA)^{-1}\| \cdot \|A\|^2 \leq 2 \cdot \|(\Sigma^T\Sigma)^{-1}\| \cdot \|V\|^2 \cdot \|A\|^2 \tag{23}$$

$\Sigma^T\Sigma$ — квадратная матрица на диагонали у которой находятся квадраты сингулярных значений матрицы, следовательно, обратная матрица к ней будет иметь вид $\text{diag}(\frac{1}{\sigma_1^2},\ \frac{1}{\sigma_2^2}\ ...\ \frac{1}{\sigma_m^2})$

$$2 \cdot \|(\Sigma^T\Sigma)^{-1}\| \cdot \|V\|^2 \cdot \|A\|^2 = 2 \cdot \|\,\text{diag}(\frac{1}{\sigma_1^2},\ \frac{1}{\sigma_2^2}\ ...\ \frac{1}{\sigma_m^2})\| \cdot \|V\|^2 \cdot \|A\|^2 \tag{24}$$

При этом сингулярные числа показывают близость матрицы к невырожденной, или же близость к матрице полного столбцового/строкового ранга, если матрица A имеет сильную зависимость между столбцами то $\min\{\sigma_1, \sigma_2...\} \to 0$, что ведет к увеличению числа обусловленности (сингулярные числа рассмотрю позже)

Второй подход к числу обусловленности

Мы можем сделать проще и рассмотреть число обусловленности матрицы A как линейного оператора. Рассмотрим уравнение:

$$Ax = b \tag{25}$$

Тогда решением в общем виде является

$$x = A^+b \tag{26}$$

Рассмотрим число обусловленности

$$\mu(A, b) = \max_{\Delta b} \frac{\frac{\|A^+(b+\Delta b) - A^+b\|}{\|A^+b\|}}{\frac{\|b+\Delta b-b\|}{\|b\|}} = \max_{\Delta b} \frac{\frac{\|A^+\Delta b\|}{\|A^+b\|}}{\frac{\|\Delta b\|}{\|b\|}} \tag{27}$$

Перенеся деление

$$\max_{\Delta b} \frac{\|A^+\Delta b\| \cdot \|b\|}{\|\Delta b\| \cdot \|A^+b\|} = \max_{\Delta b}(\frac{\|A^+\Delta b\|}{\|\Delta b\|}) \cdot (\frac{\|b\|}{\|A^+b\|}) \tag{28}$$

Аналогично прошлому случаю затребуем свойство субмультипликативности от нормы, тогда мы можем мажорировать первый множитель с помощью $\|A^+\|$. Следовательно:

$$\mu(A, b) = \|A^+\|\frac{\|b\|}{\|A^+b\|} = \|A^+\|\frac{\|Ax + \epsilon\|}{\|x\|} \tag{29}$$

где ϵ - вектор отклонений

$$\|A^+\|\frac{\|Ax+\epsilon\|}{\|x\|} \leq \|A^+\|\frac{\|Ax\|+\|\epsilon\|}{\|x\|} \tag{30}$$

Таким образом мы получаем

$$\|A^+\|\frac{\|Ax\|}{\|x\|} + \|A^+\|\frac{\|\epsilon\|}{\|x\|} \leq \|A^+\|\frac{\|A\|x\|}{\|x\|} + \|A^+\|\frac{\|\epsilon\|}{\|x\|} \tag{31}$$

Финальное выражение выглядит так:

$$\|A^+\|\|A\| + \|A^+\|\frac{\|\epsilon\|}{\|x\|} \tag{32}$$

В случае если используется сингулярная норма: Сингулярная норма матрицы — ее максимальное сингулярное значение, а псевдообратная матрица имеет обратные сингулярные значения к исходной матрице (легко показать), таким образом:

$$\|A\|_{spec} = \sigma_{max}(A) \tag{33}$$

и

$$\|A^+\|_{spec} = \frac{1}{\sigma_{min}(A)} \tag{34}$$

Таким образом мы получаем:

$$\mu(A,b) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} + \|A^+\|\frac{\|\epsilon\|}{\|x\|} \tag{35}$$

Если пренебречь вторым слагаемым, то мы получим то, что описано в учебнике ШАДА: “Пожертвовав математической строгостью, мы можем считать, что число обусловленности матрицы X – это корень из отношения наибольшего и наименьшего из собственных чисел матрицы X^TX“.

Пояснение: Сингулярные значения матрицы A — это квадратные корни из собственных значений матрицы A^TA

Intuition проблемы числа обусловленности

Матрица A — линейное отображение из одного пространства в другое, если в новом пространстве есть вектора с сильной линейной связью, то оно становится более сжатым, т.е непохожие вектора в исходном пространстве могут стать сильно более похожими в новом пространстве за счет того, что оно сжато в размерах, таким образом если мы немного изменим целевой вектор (из нового пространства), то вектор который его образовал может сильно отличаться от того, что мы получили в прошлый раз, так как за счет сжатия они лежат рядом в новом пространстве, однако не в новом, таким образом: число обусловленности — лишь следствие того, что новое пространство сжато

Для наглядности можно рассмотреть, двумерный случай, рассмотрим матрицу перехода в новое пространство

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}$$

где $\epsilon > 0$

Мы видим, что $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ отобразится в $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$, а $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ в $\left[\begin{smallmatrix}1\\1+\epsilon\end{smallmatrix}\right]$. При маленьких ϵ матрица остается невырожденной, однако пространство "сжимается" делая $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ и $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ все менее отличимыми в новом пространстве

1.4.3 Переход к новому скалярному произвдениию

Мы хотим перейти в пространство с новым скалярным произведением, однако большинство компьютерных систем заточены под обычное Евклидово скалярное произведение, поэтому нам бы хотелось перейти в такой базис, чтобы операции в этом базисе с использованием "обычного" скалярного произведения были равносильны операциям с использованием "нового" скалярного произведения в исходном пространстве

Заметим, что любое произвольное скалярное произведение может быть задано как квадратичная форма, исходя из свойства линейности.

$$\langle u, v \rangle_G = u^T G v \tag{36}$$

Заметим, что G — положительно определена и симметрична, по свойству скалярного произведения. Мы можем найти матрицу перехода P таким образом, через разложение Холецкого.

$$G = P^T P \tag{37}$$

Тогда вектор k и матрица M в новом пространстве превращаются в Pk и PMP^T соответственно

Пример

Пусть мы работаем в новом пространстве и хотим умножить матрицу на вектор в новом пространстве. У нас есть $k' = Pk$ и $M' = PMP^T$, мы хотим, чтобы итоговый вектор имел вид:

$$Mk = \begin{bmatrix} \langle M_{1,*}, k \rangle_G \\ \langle M_{2,*}, k \rangle_G \\ \vdots \\ \langle M_{n,*}, k \rangle_G \end{bmatrix}$$

это равносильно:

$$Mk = \begin{bmatrix} M_{1,*}^T Gk \\ M_{2,*}^T Gk \\ \vdots \\ M_{3,*}^T Gk \end{bmatrix}$$

следовательно

$$[Mk]_{new\ dot\ product} = MGk \tag{38}$$

Рассмотрим линейный оператор MG, тогда в новом базисе данный оператор может быть представлен как $PMGP^{-1}$

$$PMGP^{-1} = PMP^T PP^{-1} = PMP^T \tag{39}$$

Проверим, что это работает

$$P^{-1}(PMP^T Pk) = P^{-1}(PMGk) = MGk \tag{40}$$

Таким образом

$$P^{-1}M'k' = MGk \tag{41}$$

Покажем, что это сработает и для умножения на вектор справа

$$k'^T M' P^{-1^T} = k^T P^T PMP^T P^{-1^T} = k^T GM \tag{42}$$

Таким образом мы выполнили поставленную перед нами задачу