Спецкурс "Регрессия"

Шкляев А.В.

15 января 2019 г.

Задача регрессии

1 Введение

1.1 Общая постановка

Историческая справка: Термин "регрессия" происходит от латинского "regressio" — движение назад. Этот термин был введен Фрэнсисом Гальтоном, который обнаружил при исследовании взаимосвязи роста родителей и их детей то, что у высоких родителей дети в среднем менее высоки. Он назвал это regression towards the mean, то есть возвращение к среднему.

Достаточно общая постановка задачи регрессии может быть сформулирована таким образом: пусть (X, Y) — некоторый случайный вектор (здесь X может быть вектором (мы будем называть их предикторами или регрессорами), а Y — скалярная переменная (мы будем называть ее зависимой или критериальной переменной). Предположим, что мы хотим "предсказать" Y по X с помощью функции $f(X)$. Мы можем использовать такой стандартный подход — выбрать некоторую функцию потерь $L(y, u)$, которая будет показывать наши потери при оценке параметра, равного y, величиной u.

Тогда мы можем подсчитать функцию риска

$$R(f) = \mathbf{E}L(Y, f(X)).$$

Теперь наша задача выбрать f таким образом, чтобы $R(f)$ было минимальным.

Наиболее популярным выбором L является квадратичная функция $L(y, u) = (y - u)^2$. В этом случае мы ищем $f()$ так, чтобы

$$\mathbf{E}(y - f(X))^2$$

было минимально. Иначе говоря, мы минимизируем расстояние между Y и $f(X)$ в $L^2(P)$, то есть ищем проекцию Y на пространство функций от X, имеющих конечное математическое ожидание. Ответом в таком случае будет $f(X) = \mathbf{E}(Y|X)$, поскольку это одно из определений условного математического ожидания.

Другой популярной функцией потерь является абсолютная $L(y, u) = |y - u|$. В этом случае штраф за большую ошибку будет более маленьким и функция будет не столь чувствительна к отдельным большим отклонениям. В этом случае ответ также известен: $f(X) = MEDF_{Y|X}$, где $F_{Y|X}$ — условная ф.р. Y при условии X.

В задачах классификации, в которых y принимает конечное число значений $\{y_1, ..., y_k\}$, нередко используют дискретную функцию потерь $L(y, u) = I_{y \neq u}$. В этом случае ответ также несложно выписывается: $f(X)$ равен тому из y_i, при котором $\mathbf{P}(Y = y_i|X)$ максимально.

1.2 Основная задача

Ближайшее время мы будем заниматься случаем квадратичных потерь и формулировать результаты для него. Итак, в этом случае мы хотим решить задачу поиска функции f, такой что

$$\mathbf{E}(Y|X) = f(X).$$

При этом $Y = f(X) + \varepsilon$, где $\mathbf{E}\varepsilon = 0$, $\mathbf{E}\varepsilon g(X) = 0$ при любой измеримой функции g, $\mathbf{E}g(X)^2 < \infty$. Действительно, из свойств условного математического ожидания

$$\mathbf{E}(Y - f(X)) = \mathbf{E}Y - \mathbf{E}(\mathbf{E}(Y|X)) = \mathbf{E}Y - \mathbf{E}Y = 0,$$
$$\mathbf{E}(Y - f(X))g(X) = \mathbf{E}(\mathbf{E}((Y - f(X))g(X)|X)) = \mathbf{E}g(X)(\mathbf{E}(Y|X) - f(X)) = 0.$$

Необходимым и достаточным условием является условие $\mathbf{E}(\varepsilon|X) = 0$, $\mathbf{E}\varepsilon < \infty$. Действительно, достаточность вытекает из

$$\mathbf{E}\left(\mathbf{E}(\varepsilon|X)\right) = 0,\ \mathbf{E}\varepsilon g(X) = \mathbf{E}(\mathbf{E}(\varepsilon|X)g(X)) = 0.$$

Необходимость следует из соотношения

$$\mathbf{E}(\varepsilon|X) = \mathbf{E}(Y - f(X)|X) = \mathbf{E}(Y|X) - f(X) = 0.$$

Таким образом, можно трактовать задачу следующим образом: найти функцию $f(x)$, такую что $Y = f(X) + \varepsilon$, где $\mathbf{E}\varepsilon < \infty$, $\mathbf{E}(\varepsilon|X) = 0$.

Пример 1.1. Достаточным условием для этого является независимость ε и X, но это условие не необходимо. Например, если $\varepsilon = XZ$, где Z независима от X и имеет нулевое математическое ожидание, то

$$\mathbf{E}(\varepsilon|X) = X\mathbf{E}(Z|X) = X\mathbf{E}Z = 0.$$

1.3 Параметрическая и непараметрическая постановка

Дальнейшее исследование этой задачи существенно зависит от ограничений, которые мы накладываем на возможные функции f.

1. Параметрическая регрессия.
   В этом случае мы предполагаем, что $f(x)$ имеет заданный параметрически вид $f(x;\theta)$, в котором фигурируют какие-то неизвестные параметры θ, но вид $f(\cdot;\cdot)$ нам известен заранее.
   В такие задачи входят, в частности: линейная регрессия, в которой $f(\cdot;\theta)$ линейная функция x с параметрами θ; логистическая регрессия, когда Y принимает два значения $\{0,1\}$ (это не самая общая постановка задачи логистической регрессии), а функция f, задающая вероятность $Y = 1$, имеет вид $\exp(g(x,\theta))/(1+\exp(g(x,\theta))$, где $g(\cdot;\theta)$ — линейная функция.

2. Непараметрическая регрессия.
   В этом случае мы делаем лишь общие предположения о виде f. При этом возникает проблема "переобучения", поскольку по конечной выборке (X_i, Y_i) мы можем подобрать функцию f так что $f(X_i) = Y_i$ при всех i. В связи с этим задача, по сути своей, сводится к параметрической, поскольку f заменяется на конечное количество коэффициентов (например, коэффициентов ее разложения в ряд). О непараметрической регрессии мы поговорим в конце курса.

2 Практические аспекты

2.1 Причинность и зависимость

Представим, что дети пишут языковой тест, y — их оценка, x — вес ребенка. Пусть мы обнаружили, что y в целом больше, когда больше x. Можно ли говорить, что больший вес детей влечет лучшую успеваемость?
А что, если дети разных возрастов от 5 до 15 лет?
А что, если среди детей есть дети из двух разных стран, причем в одной стране дети в целом крупнее, чем в другой?
А что, если родители кормят детей конфетами, если те хорошо учатся?
Таким образом, исследование причинности достаточно сложно. С помощью регрессии мы можем обнаружить зависимость переменных и ее направление, но истоки этой зависимости могут быть различными. Причинно-следственная связь может идти от чего-то третьего, она может быть "дискретной" (как в примере с двумя странами) или может быть и вовсе быть обратной.

2.2 Целесообразность использования регрессии

1. Мы можем делать предсказание.
   С помощью регрессии можно прогнозировать Y, глядя на X.

2. Мы можем исключать неудобные переменные.
   Если Z зависит от X и Y, то мы можем устранить влияние X, рассмотрев регрессию, скажем, $Y = aX + b$, оценив коэффициенты $\widehat{a}$, $\widehat{b}$, а затем сделать регрессию Z по $Y - \hat{a}X - \hat{b}$. Тем самым, мы устраняем влияние X на Y и исследуем зависимость Z от Y без учета X.
   Например, можно устранить фактор возраста из задачи про языковой тест, убрав влияние возраста на вес.

3. Мы можем количественно измерить влияние факторов на результат с помощью коэффициентов регрессии.

2.3 Проблемы и ошибки

1. Коллинеарность предикторов.
   Представим себе крайний случай, когда один из предикторов встречается дважды. Тогда мы не можем однозначно восстановить коэффициенты при первом и втором. Мы можем получить различные ответы (являющиеся правильными) и сделать вывод, что модель плохо подходит к данным (поскольку ответ неустойчив). Такая же картина будет появляться, если предикторы не тождественны, но близки друг к другу.

2. Зависимость результата от отрезка, на котором мы изучаем данные.
   Варьируя отрезок, на котором мы рассматриваем данные, мы получаем разные ответы.

3. Зависимость результата от выбранных переменных.
   Если мы хотим оценить вклад одного элемента в результат, то он будет существенно зависеть от того, какие вообще переменные мы включили в модель. Скажем, если $y = 2x_1 + 3x_2$, то при регрессии $y(x_1, x_2)$ коэффициент при x_1 равен 2, а при регрессии $y(x_1, x_3)$, где $x_3 = x_1 + x_2$, он равен -1.

4. Предположение о виде зависимости.
   Мы делаем исходное предположение о виде зависимости, которое достаточно непросто проверить. При этом это предположение существенно влияет на результат.

3 Простая линейная регрессия

Начнем с наиболее простой задачи, когда X скалярный, функция потерь квадратичная, а $f(x) = ax+b$ — линейная.

Более того, предположим для простоты, что $Y = aX + b + \varepsilon$, где ε — н.о.р. величины, не зависящие от X с $\mathbf{E}\varepsilon = 0$. В связи с этим мы можем фиксировать выборку X и рассматривать задачу при ее условии. В этом случае x_i будут константами, а ε останутся с тем же распределением.

3.1 ОМП в нормальной модели

Пусть $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Тогда мы можем использовать для оценки коэффициентов a, b метод максимального правдоподобия.
Функция правдоподобия имеет вид

$$L(y_1, ..., y_n; a, b, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - ax_i - b)^2\right).$$

Тогда

$$\ln L(y_1, ..., y_n; a, b, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - ax_i - b)^2.$$

Чтобы найти максимум по a, b, минимизируем

$$g(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2.$$

Тогда

$$\frac{\partial \ln L}{\partial a} = -2\sum_{i=1}^n (y_i - ax_i - b)x_i = 0,$$
$$\frac{\partial \ln L}{\partial b} = -2\sum_{i=1}^n (y_i - ax_i - b) = 0.$$

Решая полученную систему уравнений, получаем ОМП

$$\widehat{a} = \frac{\overline{yx} - \overline{y}\cdot\overline{x}}{\overline{x^2} - \overline{x}^2},\ \widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

Здесь $\overline{yx} = \sum_{i=1}^n y_i x_i/n$, $\overline{x^2} = \sum_{i=1}^n x_i^2/n$, $\overline{x} = \sum_{i=1}^n x_i/n$, $\overline{y} = \sum_{i=1}^n y_i$. То, что найденная точка действительно точка максимума, следует из выпуклости $g(a, b)$ вниз.

Можно интерпретировать найденную оценку $\widehat{a}$ как $Cov(x, y)/S_X^2$, где $Cov(x, y) = \overline{yx} - \overline{x}\cdot\overline{y}$ — выборочная ковариация, $S_X^2 = \overline{x^2} - \overline{x}^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$ — выборочная дисперсия x_i.

Остается найти ОМП для σ:

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (y_i - ax_i - b)^2,$$

откуда

$$\widehat{\sigma}^2 = \frac{RSS}{n},\ RSS = \sum_{i=1}^n (y_i - \widehat{a}x_i - \widehat{b})^2,$$

где RSS (Residual Sum of Squares или остаточная сумма квадратов) соответствует сумме квадратов

ошибок прогноза для точек (x, y).

3.2 Распределение коэффициентов в нормальной модели

Найдем распределения $\widehat{a}$, $\widehat{b}$.

Воспользуемся представлением $y_i = ax_i + b + \varepsilon_i$:

$$\widehat{a} = \frac{1}{S_X^2}\frac{1}{n}\sum_{i=1}^{n}(ax_i + b + \varepsilon_i)(x_i - \overline{x}) = a\frac{\sum_{i=1}^{n}(x_i^2 - x_i\overline{x})}{nS_X^2} + b\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{nS_X^2} + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i}{nS_X^2}.$$

Остается заметить, что

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - x_i\overline{x}\right) = \overline{x^2} - \overline{x}^2,\ \sum_{i=1}^{n}(x_i - \overline{x}) = 0,\ \frac{1}{nS_X^2}\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i \sim \mathcal{N}\left(0, \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2\sigma^2}{n^2S_X^4}\right).$$

Тем самым,

$$\widehat{a} \sim \mathcal{N}\left(a, \frac{\sigma^2}{nS_X^2}\right).$$

Аналогичным образом,

$$\widehat{b} \sim \mathcal{N}\left(b, \frac{\sigma^2\overline{x}^2}{nS_X^2}\right).$$

Более того,

$$(\widehat{a}, \widehat{b}) \sim \mathcal{N}\left((a, b),\ \frac{\sigma^2}{nS_X^2}\Sigma\right),$$

где

$$\Sigma = \begin{pmatrix} 1 & -\overline{x}, \\ -\overline{x} & \overline{x^2} \end{pmatrix}.$$