Математические основы искусственного интеллекта. Регрессионный анализ

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

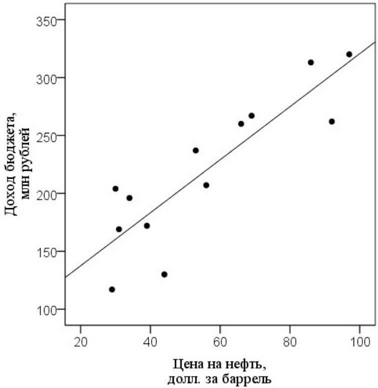
Март 2022

Пусть даны наблюдения за двумя случайными величинами X и Y, которые мы будем трактовать как цену на нефть в долларах за баррель и доход в бюджет в миллионах рублей.

X	30	69	86	56	44	97	53
Y	204	267	313	207	130	320	237

X	66	39	29	34	31	92
Y	260	172	117	196	169	262

Постановка задачи



После того как методами корреляционного анализа установлено, что существует линейная связь между признаками, естественно возникает вопрос, как описать эту связь в виде формулы.

Постановка задачи

Пусть установлено, что цены на нефть влияют на доходную часть бюджета. Требуется узнать:

1. на сколько увеличивается доход бюджета при увеличении цены на нефть на один доллар за баррель;
2. какой ожидается доход бюджета, если цена на нефть установится на уровне 80 долл. за баррель.

План лекции

В рамках теоретико-вероятностного подхода рассматриваем систему линейно зависимых случайных величин Y и X, распределения которых известны, и описываем связь между ними в виде уравнения линейной регрессии — по сути, находим условное математическое ожидание Y по X.

Рассматриваем систему линейно зависимых случайных величин Y и X, но распределения Y и X неизвестны, есть лишь набор из n наблюдений (x_i, y_i), $i = 1, 2, \dots, n$, за X и Y. По этому набору строим выборочное уравнение линейной регрессии.

Строим множественную линейную регрессию, решаем проблемы связанные с увеличением числа предикторов.

Рассматриваем величины, связь которых близка к линейной, но таковой не является, например, $Y = X^{1.2} + \varepsilon$. Ситуация осложняется тем, что точный вид связи априори неизвестен. По набору из n наблюдений (x_i, y_i), $i = 1, 2, \dots, n$, строим выборочное уравнение линейной регрессии, которая лишь упрощенно описывает нелинейную связь.

Следуем теоретико-вероятностному подходу. Не работаем с выборками и не делаем оценок. Рассматривается система случайных величин, описанных функцией или плотностью совместного распределения. Затем на основе этих данных строится функция линейной регрессии Y по X.

Пусть $f_{X,Y}(x, y)$ — плотность совместного распределения случайных величин X и Y. Тогда маргинальная плотность распределения случайных величин X и Y определяется следующим образом:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy, \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dx.$$

Для каждого фиксированного значения x случайной величины X и значения y случайной величины Y условные распределения Y по X и X по Y соответственно определяются по формулам:

$$f_Y(y|X=x) = \frac{f_{X,Y}(x,y)}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy}, \quad f_X(x|Y=y) = \frac{f_{X,Y}(x,y)}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx}.$$

Условные математические ожидания Y при фиксированном x и X при фиксированном y:

$$M[Y|X=x]=\frac{\int\limits_{-\infty}^{+\infty} y\, f_{X,Y}(x,y)\,dy}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy},\quad M[X|Y=y]=\frac{\int\limits_{-\infty}^{+\infty} x\, f_{X,Y}(x,y)\,dx}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx}.$$

Эти соотношения соответственно определяют регрессии Y по X и X по Y (кривые регрессии).

Первое выражает зависимость среднего значения величины Y от x. Данная зависимость, вообще говоря, нелинейная, является функциональной, а не статистической.

Многомерное нормальное распределение

Многомерное нормальное распределение вектора $\mathbf{X}$ с математическим ожиданием $\mathbf{m} \in \mathbb{R}^n$ и ковариационной матрицей Σ размерности $n \times n$ с плотностью распределения

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\top}\Sigma^{-1}(\mathbf{x}-\mathbf{m})}, \quad \mathbf{x} \in \mathbb{R}^n,$$

где $|\Sigma|$ — определитель матрицы Σ; Σ^{-1} — матрица, обратная к Σ.

Двумерное нормальное распределение

При $n = 2$ плотность двумерного невырожденного (если коэффициент корреляции r по модулю не равен единице) нормального распределения можно записать в виде:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - r\frac{2(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]\right\}.$$

Условное математическое ожидание X_2 имеет вид

$$M[X_2|X_1 = x_1] = m_2 + r\frac{\sigma_2}{\sigma_1}(x_1 - m_1),$$

т. е. выражается линейной функцией.

Когда регрессия имеет линейный вид

так что

$$\mu_{2v} = \mu'_{2v} - (\mu'_{1v})^2 = (1-\rho^2)\left\{2\rho^2 v + \frac{1}{2} n (1-\rho^2)^2\right\}. \quad (28.23)$$

Соотношения (28.22) и (28.23) показывают, что регрессии среднего и дисперсии величины u по v линейны.

Критерии линейности регрессии

28.5 Пусть $\psi(t_1, t_2) = \log \varphi(t_1, t_2)$ — совместная п. ф. с. величин x и y. Сейчас мы докажем следующий факт: если регрессия величины y по x линейна, так что

$$\mu'_{1x} = \mathrm{M}(y \mid x) = \beta_0 + \beta_1 x, \quad (28.24)$$

то

$$\left[\frac{\partial \psi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i\beta_0 + \beta_1 \frac{\partial \psi(t_1, 0)}{\partial t_1}; \quad (28.25)$$

и наоборот, если выполнено некоторое условие полноты, то (28.25) не только необходимо, но и достаточно для (28.24).

Используя (28.24), из (28.9) при $r = 1$ находим

$$\left[\frac{\partial \varphi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i \int_{-\infty}^{\infty} \exp(it_1 x)\, g(x)(\beta_0 + \beta_1 x)\, dx = \quad (28.26)$$

$$= i\beta_0 \varphi(t_1, 0) + \beta_1 \frac{\partial}{\partial t_1} \varphi(t_1, 0). \quad (28.27)$$

Полагая в (28.27) $\psi = \log \varphi$ и деля обе части на $\varphi(t_1, 0)$, получаем (28.25).

Обратно, если имеет место соотношение (28.25), то, используя (28.9), перепишем его в виде

$$i \int_{-\infty}^{\infty} \exp(it_1 x)(\beta_0 + \beta_1 x - \mu'_{1x})\, g(x)\, dx = 0. \quad (28.28)$$

Теперь видим, что соотношение (28.28) влечет тождественно по x

$$\beta_0 + \beta_1 x - \mu'_{1x} = 0, \quad (28.29)$$

если только семейство $\exp(it_1 x) g(x)$ полно. Следовательно, мы получили (28.24).

Для упрощения математической стороны изложения мы наложим ограничения на случайные величины X и Y. Пусть условное распределение Y относительно своего среднего (которое, как и раньше, является функцией от x) одно и то же для любого x, т. е. только среднее значение Y изменяется при изменении x. Говорят, что Y имеет «однородные ошибки»[1]. Таким образом, существует случайная величина ε такая, что

$$Y = M[Y|X = x] + \varepsilon.$$

В частности, когда регрессия линейная, имеем

$$Y = \beta_1 X + \beta_0 + \varepsilon.$$

[1] Кендалл, Стьюарт Статистические выводы и связи, М. : Наука 1973, стр. 467

В прикладных задачах априори известен только вид уравнения, а конкретные значения коэффициентов β_0 и β_1 неизвестны, естественно выбрать их так, чтобы, зная значение, которое в эксперименте приняла величина X, наиболее точно спрогнозировать значение, которое примет величина Y.
Постановка задачи построения линейной регрессии:

$$M[Y - \beta_1 X - \beta_0]^2 \xrightarrow[\beta_0,\,\beta_1]{} \min .$$

Теорема. *Линейная среднеквадратическая регрессия Y на X имеет вид:*

$$f(x) = m_Y + r\frac{\sigma_Y}{\sigma_X}(x - m_X),$$

где m_Y, m_X, σ_Y, σ_X — математические ожидания и средние квадратические отклонения случайных величин Y и X соответственно; r — коэффициент корреляции случайных величин Y и X.

Доказательство теоремы

Рассмотрим функцию $F(\beta_0, \beta_1) = = M[Y - \beta_1 X - \beta_0]^2$. Пользуясь формулами

$$M[X^2] = M^2[X] + D[x]$$

$$M[XY] = M[X]M[Y] + \mathrm{cov}(X, Y) = M[X]M[Y] + r\sigma_X\sigma_Y,$$

получим

$$F(\beta_0, \beta_1) = \sigma_Y^2 + \beta_1^2\sigma_X^2 - 2r\sigma_X\sigma_Y\beta_1 + (m_Y - \beta_1 m_X - \beta_0)^2.$$

Исследуем функцию F на минимум:

$$\frac{\partial F}{\partial \beta_0} = -2(m_Y - \beta_1 m_X - \beta_0) = 0,$$

$$\frac{\partial F}{\partial \beta_1} = 2\beta_1\sigma_X^2 - 2r\sigma_X\sigma_Y = 0.$$

Откуда получаем

$$\beta_0 = m_Y - r\frac{\sigma_Y}{\sigma_X}m_X, \qquad \beta_1 = r\frac{\sigma_Y}{\sigma_X}.$$

Замечание по терминологии

Функцию $f: \mathbb{R} \to \mathbb{R}$ называют функцией регрессии.

Аргументом и значением функции регрессии являются числа, а не случайные величины.

Если функция f имеет вид $\beta_1 x + \beta_0$, то регрессию называют (парной) линейной; случай множественной линейной регрессии аналогичен.

Независимые переменные иначе называют предикторами, регрессорами или факторами.

Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Выборочное уравнение линейной регрессии

Рассмотрим случай линейной связи.

Аналитическая теория регрессии требует точного знания функции распределения рассматриваемой системы случайных величин, а потому интересна для теории вероятностей, но не для прикладной статистики.

Мы продолжаем рассматривать систему линейно зависимых случайных величин Y и X, но на этот раз распределения Y и X неизвестны, есть лишь набор из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, за X и Y. Предполагается, что не все x_i равны между собой.

Линейная модель, в рамках которой мы работаем, имеет вид:

$$y_i = b_1 x_i + b_0 + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Выборочное уравнение линейной регрессии

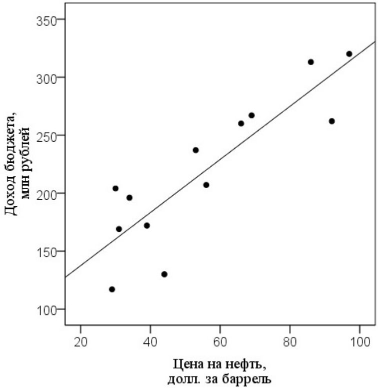
В классической линейной регрессии предполагается, что выполнены следующие условия:

1. факторы и случайные ошибки — независимые случайные величины;
2. случайные ошибки модели гомоскедастичные, т. е. дисперсия ошибок постоянная, не зависит от значений предикторов;
3. отсутствует корреляция (автокорреляция) случайных ошибок разных наблюдений между собой: $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \; 1 \le i < j \le n$.

Выборочное уравнение линейной регрессии

По набору из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, строим выборочное уравнение линейной регрессии $\hat{y}(x) = b_1 x + b_0$, где параметры b_0 и b_1 будут соответственно выборочными оценками параметров регрессии β_0 и β_1.

Подберем параметры b_1, b_0 так, чтобы прямая $\hat{y}(x) = b_1 x + b_0$ проходила как можно ближе к точкам (x_i, y_i), $i = 1, 2, \ldots, n$.



Поиск коэффициентов

Формализовать эту идею можно следующим образом. Рассмотрим функцию

$$F(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2.$$

Исследуем функцию F на минимум, приравняв частные производные к нулю:

$$\frac{\partial F}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_1 x_i - b_0) = 0,$$

$$\frac{\partial F}{\partial b_1} = 2 \sum_{i=1}^{n} (y_i - b_1 x_i - b_0) x_i = 0.$$

Поиск коэффициентов

Имеем оценки коэффициентов регрессии:

$$b_0 = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2},$$

$$b_1 = \frac{n \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}.$$

Из этих формул, в частности, следует, что для линейных моделей МНК-оценки являются линейными.

Свйоства МНК оценок

МНК-оценки для классической линейной регрессии являются несмещенными, состоятельными и наиболее эффективными оценками в классе всех линейных несмещенных оценок.

Требования к модели и свойства оценок устанавливает теорема Гаусса–Маркова. В англоязычной литературе иногда употребляют аббревиатуру BLUE (Best Linear Unbiased Estimator) — наилучшая линейная несмещенная оценка.

Пример построения парной линейной регрессии

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,870[a]	,758	,736	32,902

a. Predictors: (Constant), X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37215,432	1	37215,432	34,378	,000[a]
	Residual	11907,798	11	1082,527		
	Total	49123,231	12			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	91,696	23,636		3,879	,003
	X	2,289	,390	,870	5,863	,000

a. Dependent Variable: Y

Единственное важное число в таблице ANOVA — это величина p-value в столбце Sig., вероятность ошибки первого рода. В данном примере Sig. < 0.001, следовательно, на уровне значимости 0.001 нулевую гипотезу об отсутствии связи между случайными величинами X и Y отвергаем в пользу альтернативной, т. е. считаем связь между X и Y статистически значимой на уровне значимости 0.001.

Если расчетная величина Sig. оказалась больше уровня значимости, например, Sig. $= 0.785$, нет оснований отвергнуть нулевую гипотезу, статистическая связь между переменными не обнаружена, и анализ заканчивают.

Находим явный вид уравнения регрессии:

$$\hat{y}(x) = 2.289x + 91.696.$$

В столбце Sig. приведены результаты тестирования двух нулевых гипотез: о том, что константа и коэффициент при X в уравнении незначимо отличаются от нуля.

Для константы Sig.$= 0.003$, для коэффициента при X Sig. < 0.001.

Если, например, уровень значимости $\alpha = 0.05$, то обе нулевые гипотезы отвергаем как противоречащие экспериментальным наблюдениям.

Точность прогноза

Поскольку точные значения константы и коэффициента при x неизвестны (они были оценены статистическими методами по выборочным данным), возникает вопрос о точности этих оценок.

Стандартные ошибки оценок равны 23.636 и 0.390 соответственно.

Чем менее точки на графике разбросаны относительно прямой и чем больше наблюдений, тем ошибки меньше, соответственно оценки точнее.

После того как модель построена, необходимо ответить на вопрос, насколько эта регрессионная модель точна. В данном примере среднее значение Y равно 219.54, стандартное отклонение — 63.98, упрощенно говоря, можно сказать, что ожидаемое значение Y равно 219.54 ± 63.98. То есть не имея никакой априорной информации, можно сделать предположение о значении Y, при этом «коридор» ошибок составляет 63.98. Если же знать значение предиктора X, то можно сузить этот «коридор» на 75.8 %.

Более строго величина $R^2 = 0.758$, называемая коэффициентом детерминации, характеризует долю объясненной дисперсии. Коэффициент детерминации меняется от 0 до 1, чем он больше, тем точнее модель.

Задания

Профессор Р. А. Шамойлова учит студентов проводить регрессионный анализ. На лекции она сказала: «Основной предпосылкой регрессионного анализа является то, что только результативный признак подчиняется нормальному распределению, а факторные признаки могут иметь произвольный закон распределения».

Согласны ли вы с этим утверждением?