

Лекция II

Регрессия

Т. Е. Абильдаев

Кафедра статистического моделирования
Матмех СПбГУ

2025-03-01

Регрессия: вероятностный бэкграунд

Вспомним немного теории вероятностей для вдохновения.

Пусть X , Y – некоторые случайные величины. Как мы можем помнить, условное математическое ожидание $\mathbb{E}(Y|X)$ – это, грубо говоря, усреднение Y по σ -алгебре, порождаемой X , которое при этом измеримо относительно этой σ -алгебры. По теореме о представлении случайной величины, измеримой относительно другой случайной величины, существует такая борелевская функция f , что

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

Какой смысл несёт f ?

Регрессия: вероятностный бэкграунд

Вспомним немного теории вероятностей для вдохновения.

Пусть X, Y – некоторые случайные величины. Как мы можем помнить, условное математическое ожидание $\mathbb{E}(Y|X)$ – это, грубо говоря, усреднение Y по σ -алгебре, порождаемой X , которое при этом измеримо относительно этой σ -алгебры. По теореме о представлении случайной величины, измеримой относительно другой случайной величины, существует такая борелевская функция f , что

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

Какой смысл несёт f ?

Такой: если у Y конечная дисперсия, то f решает оптимизационную задачу

$$\mathbb{E}|Y - g(X)|^2 \longrightarrow \inf_{g \in \mathcal{B}(\mathbb{R})}.$$

Величину $\mathbb{E}(Y|X)$ также называют *регрессией Y на X* , а функцию $f : \mathbb{R} \rightarrow \mathbb{R}$ – *функцией регрессии*.

Что если нам интересно приближение в другом классе функций и не в смысле L_2 ?

Определение регрессии. Линейная регрессия

Определение

Регрессией переменной (случайной величины) Y на переменные (случайные величины) $X_1, \dots, X_m$ мы будем называть приближение в некотором смысле величины Y величинами $X_1, \dots, X_m$.

Упрощая, регрессия — это про нахождение функциональной зависимости между переменными.

Перейдём к конкретным моделям и начнём с самой известной из них.

Пусть X — матрица, j -ый столбец которой есть вектор из m наблюдений переменной X_j , $j \in \{1, m\}$, а нулевой — вектор длины m из единиц. Ища оптимальную в среднеквадратическом смысле линейную зависимость Y от X , мы приходим к задаче *линейной регрессии*:

$$\|Y - X\beta\|_2^2 = \sum_{k=1}^n \left[Y_k - \sum_{j=0}^m X_{kj} \beta_j \right]^2 \longrightarrow \inf_{\beta_0, \dots, \beta_m}.$$

При чём здесь вероятность и случайные величины?

Линейная регрессия (прод.)

При том, что задачу регрессии, конечно же, можно ставить и решать для любого класса функций и любого loss'a, но будет ли в таком произволе толк – вопрос важный.

Предположим, что $X_1, \dots, X_m, Y$ распределены нормально. Тогда

- а) класс линейных функций оправдан, так как гауссовская регрессия является линейной;
- б) L_2 -loss оправдан, так как по теореме Гаусса-Маркова он доставляет оценку параметров с наименьшей среди линейных несмещённых оценок дисперсией.

Линейная регрессия (прод.)

При том, что задачу регрессии, конечно же, можно ставить и решать для любого класса функций и любого loss'a, но будет ли в таком произволе толк – вопрос важный.

Предположим, что $X_1, \dots, X_m, Y$ распределены нормально. Тогда

- а) класс линейных функций оправдан, так как гауссовская регрессия является линейной;
- б) L_2 -loss оправдан, так как по теореме Гаусса-Маркова он доставляет оценку параметров с наименьшей среди линейных несмещённых оценок дисперсией.

Вообще, прежде, чем говорить о регрессии чего-то на что-то, полезно понимать, как математически подходить к понятиям *причины* и *следствия* (см. слайды по **causal inference**).

Какие ещё бывают регрессии, или, вернее, как можно обобщить линейную регрессию?

Обобщённая линейная модель

В сеттинге линейной регрессии мы имеем

$$\mathbb{E}(Y|X) = X\beta.$$

Понятно, что в общем случае на линейность рассчитывать не приходится. Однако мы можем предположить, что величина $\mathbb{E}(Y|X)$ линеаризуется некоторой функцией g , то есть

$$g(\mathbb{E}(Y|X)) = X\beta.$$

Очевидное требование – g должна быть биекцией.

Как мы помним, $\mathbb{E}(Y|X)$ – это, вообще говоря, случайная величина. Так вот в сеттинге обобщённой линейной регрессии эта величина принадлежит к некоторому экспоненциальному семейству.

Давайте вспомним, что такое экспоненциальное семейство распределений.

Экспоненциальное семейство распределений

Пусть семейство $\Pi = \{f(y; \theta) \mid \theta \in \Theta \subset \mathbb{R}^d\}$ таково, что

$$f(y; \theta) = h(y)e^{\langle \eta(\theta), T(y) \rangle - B(\theta)}$$

для некоторых функций h , η , T и B (h положительная, η и T векторнозначные).

Тогда Π называется экспоненциальным семейством распределений.

Чем является $T(y)$ со статистической точки зрения?

Экспоненциальное семейство распределений

Пусть семейство $\Pi = \{f(y; \theta) \mid \theta \in \Theta \subset \mathbb{R}^d\}$ таково, что

$$f(y; \theta) = h(y) e^{\langle \eta(\theta), T(y) \rangle - B(\theta)}$$

для некоторых функций h , η , T и B (h положительная, η и T векторнозначные).

Тогда Π называется экспоненциальным семейством распределений.

Чем является $T(y)$ со статистической точки зрения?

$T(y)$ является *достаточной статистикой*, то есть «содержит в себе всю информацию о данных (выборке)». Что это значит на более формальном уровне?

Пусть у нас имеется выборка, то есть последовательность независимых одинаково распределённых случайных величин. Тогда совместная плотность наблюдений есть

$$f(y_1, \dots, y_n; \theta) = h_n(y_1, \dots, y_n) e^{\langle \eta(\theta), T_n(y_1, \dots, y_n) \rangle - B_n(\theta)},$$

где $h_n(y_1, \dots, y_n) = \prod_{k=1}^n h(y_k)$, $T_n(y_1, \dots, y_n) = \sum_{k=1}^n T(y_k)$, $B_n(\theta) = nB(\theta)$.

Таким образом, для того, чтобы оценить θ , достаточно лишь знать $T(y)$.

Примеры экспоненциальных семейств

1. $\text{Ber}(p)$.

Имеем

$$f(y; p) = y^p(1 - y)^{1-p} = \exp \left[\log \left(\frac{p}{1-p} \right) y - \log \left(\frac{1}{1-p} \right) \right],$$

откуда $h(y) \equiv 1$, $\eta(p) = \text{logit}(p)$, $T(y) = y$, $B(p) = \log(1/(1-p))$.

2. $N(a, \sigma^2)$.

3. $\text{LogN}(a, \sigma^2)$.

4. $\text{Bin}(n, p)$ при фиксированном n .

5. $\text{Mult}(n, p_1, \dots, p_k)$ при фиксированном n .

6. $\text{Geom}(p)$.

7. $\text{NBin}(n, p)$ при фиксированном n .

8. $\text{Pois}(\lambda)$.

9. $\text{Exp}(\lambda)$.

10. $\text{Gamma}(\alpha, \lambda)$

11. $\text{Beta}(\alpha, \beta)$.

12. $\chi^2(n)$.

...

Обобщённая линейная модель (прод.)

Напомним, что пока мы имеем

$$g(\mathbb{E}(Y|X)) = X\beta, \quad \mathbb{E}(Y|X) \sim h(y)e^{\langle \eta(\theta), T(y) \rangle - B(\theta)}.$$

Завершающей деталью является *связь* (*link*), выражаемая тождеством

$$g(\mathbb{E}(Y|X)) = X\beta = \theta.$$

Таким образом,

$$\mathbb{E}(Y|X) \sim h(y)e^{\langle \eta(X\beta), T(y) \rangle - B(X\beta)},$$

или на языке наблюдений

$$Y_k \sim h(y) \exp\left(\left\langle \eta\left(\sum_{j=0}^m X_{kj}\beta_j\right), T(y) \right\rangle - B\left(\sum_{j=0}^m X_{kj}\beta_j\right)\right).$$

Параметры $\beta_0, \dots, \beta_m$ оцениваются по методу максимального правдоподобия.

Примеры обобщённых линейных моделей

1. $N(a, \sigma^2)$.

Это, очевидно, обычная линейная регрессия.

2. $\text{Ber}(p)$.

Имеем

$$f(y; p) = \exp \left[\log \left(\frac{p}{1-p} \right) y - \log \left(\frac{1}{1-p} \right) \right],$$

откуда

$$\log \left(\frac{p}{1-p} \right) = \text{logit}(p) = X\beta \iff p = \sigma(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}.$$

Это логистическая регрессия, или лог-регрессия.

Для того, чтобы найти β , решаем

$$\sum_{k=1}^n \left[Y_k \sum_{j=0}^m X_{kj} \beta_j - \log \left(1 + \exp \left(\sum_{j=0}^m X_{kj} \beta_j \right) \right) \right] \longrightarrow \sup_{\beta_0, \dots, \beta_m}.$$

Примеры обобщённых линейных моделей (прод.)

3. $\text{Pois}(\lambda)$.

Имеем

$$f(y; p) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{1}{y!} e^{\log(\lambda)y - \lambda},$$

откуда

$$\log(\lambda) = X\beta \iff \lambda = \exp(X\beta).$$

Это пуассоновская регрессия, или лог-линейная регрессия.

Для того, чтобы найти β , решаем

$$\sum_{k=1}^n \left[Y_k \sum_{j=1}^{m+1} X_{kj} \beta_j - \exp \left(\sum_{j=0}^m X_{kj} \beta_j \right) \right] \longrightarrow \sup_{\beta_0, \dots, \beta_m}.$$

На практике зачастую переменных много. Если мы добавим в модель их все, скорее всего, произойдёт *переобучение* – модель будет отлично аппроксимировать данные, на которых она обучалась, но будет плоха на новых данных.

Какие бывают подходы к выбору регрессоров?

a) Best subsets regression.

Идея проста. Давайте выберем метрику качества модели и сравним её значения на всевозможных подмножествах множества регрессоров. Выбираем то из них, на котором метрика наилучшая.

Какой у этого подхода критический недостаток?

На практике зачастую переменных много. Если мы добавим в модель их все, скорее всего, произойдёт *переобучение* – модель будет отлично аппроксимировать данные, на которых она обучалась, но будет плоха на новых данных.

Какие бывают подходы к выбору регрессоров?

a) Best subsets regression.

Идея проста. Давайте выберем метрику качества модели и сравним её значения на всевозможных подмножествах множества регрессоров. Выбираем то из них, на котором метрика наилучшая.

Какой у этого подхода критический недостаток?

Множество подмножеств растёт экспоненциально относительно числа регрессоров, поэтому прямой перебор это не самая светлая идея.

О выборе регрессоров (прод.)

б) Stepwise regression.

Давайте оставим метрику качества модели, множество подмножеств регрессоров и прикрутим к ним жадный поиск. Есть два подхода – forward selection и backward selection. В forward'е на каждом шаге добавляем переменную, приращение метрики на которой наибольшее. В backward'е начинаем с полной модели и на каждом шаге отбрасываем переменную, на которой уменьшение метрики наименьшее. Останавливаемся либо на заранее выбранном пороге метрики, либо тогда, когда приращение/уменьшение становится незначительным.

в) LASSO regression.

Аддитивная добавка к целевой функции в виде $\lambda \|\beta\|_1$, $\lambda > 0$, приводит к тому, что в результате оптимизации многие компоненты β зануляются.

г) Ridge regression.

А это уже не про выбор регрессоров, а про их контроль. Аддитивная добавка к целевой функции в виде $\lambda \|\beta\|_2$, $\lambda > 0$, уменьшает абсолютные значения компонентов β .

А как вообще сравнивать модели друг с другом? То есть какие метрики бывают?

- В случае линейной модели у нас есть R^2 , R_{adj}^2 , C_p -статистика Мэллоу. Последнее считается по формуле

$$C_p = \text{MSE} + \frac{2n(d+1)}{\hat{\sigma}^2},$$

- В общем случае, помимо прочего, у нас есть информационные критерии. Акаике:

$$\text{AIC}(\theta) = L_\theta - d,$$

байесовский:

$$\text{BIC}(\theta) = L_\theta - \frac{\log(n)}{2}d.$$

При $n \rightarrow \infty$ выбор по BIC приводит к истинным параметрам, а выбор по AIC приводит к модели с наименьшей дисперсией ошибки.