Обобщенные линейные модели

К настоящему моменту мы познакомились с задачами регрессии и классификации. В случае с регрессией мы строили модель случайной величины $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$, а в задаче классификации – случайной величины $y|x;\theta \sim Bernoulli(\phi)$ с соответствующими параметрами μ и σ, заданными как функции от x и θ. Сейчас же мы с вами покажем, что оба эти примера являются частными случаями более широкого семейства моделей, называемых Обобщенными линейными моделями (ОЛМ[1]). Мы также покажем, что в рамках этого подхода можно вывести и другие модели для решения задач регрессии и классификации.

1. Экспоненциальное семейство распределений

Для начала определим экспоненциальное семейство распределений. Будем говорить, что класс распределений вероятностей принадлежит экспоненциальному семейству, если его можно записать в следующем виде:

$$p(y;\eta) = b(y)\exp\big(\eta^T T(y) - a(\eta)\big)^{2}. \quad (1)$$

Здесь η называется **естественным параметром** (а также **каноническим параметром**) распределения; $T(y)$ – это **достаточная статистика** (для распределений, с которыми мы будем иметь дело, это как правило будет $T(y) = y$), $a(\eta)$ – это **нормирующий коэффициент** распределения. Величина $e^{-a(\eta)}$ необходима, чтобы распределение $p(y;\eta)$ суммировалось/ интегрировалось по y в единицу.

Конкретный выбор T, a и b определяет множество (или класс) распределений, параметризованных η; меняя значение η, мы получаем различные конкретные распределения в этом классе.

Покажем теперь, что распределения Бернулли и Гаусса являются примерами классов[3] распределений, принадлежащих экспоненциальному семейству.

1.1 Распределение Бернулли

Распределение Бернулли с параметром ϕ, записываемое как $Bernoulli(\phi)$, задает распределение случайной величины, которая может принимать только два значения $y \in \{0,1\}$, следующим образом:

$$p(y=1;\phi) = \phi, \qquad p(y=0;\phi) = 1-\phi.$$

Мы можем записать функцию распределения в одну строку:

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}.$$

Меняя значения ϕ, мы получаем конкретные распределения Бернулли с различными математическими ожиданиями положительного исхода случайного эксперимента. Давайте теперь покажем, что можно подобрать такие значения T, a и b, что равенство (1) будет в точности соответствовать функции распределения Бернулли.

Выполним следующие эквивалентные преобразования:

---

[1] Generalized Linear Model, GLM.

[2] Обратите внимание на точку с запятой в функции p – она означает параметризацию распределения, т.е. таким образом мы записываем распределение вероятностей случайной величины y, параметризованное η. Также напомним, что запись $\exp(x)$ означает e^x.

[3] Обычно используют термин *семейство нормальных распределений*, в котором конкретное распределение задается конкретными значениями параметров μ и σ, но, чтобы не путаться в этих многочисленных «семействах», мы будем пока использовать синонимичное слово «класс».

$$p(y;\phi) = \phi^y(1-\phi)^{1-y} = e^{\log(\phi^y(1-\phi)^{1-y})} = e^{y\log\phi+(1-y)\log(1-\phi)} = e^{\log\left(\frac{\phi}{1-\phi}\right)y+\log(1-\phi)}.$$

Таким образом, получаем, что для Бернулли естественный параметр $\eta = \log\left(\frac{\phi}{1-\phi}\right)$. Интересно, что если мы теперь попробуем выразить ϕ через η, то получим $\phi = \frac{1}{1+e^{-\eta}}$. А это есть не что иное, как хорошо знакомая нам сигмоида! К этому факту мы снова вернемся, когда покажем, что логистическая регрессия есть частный случай ОЛМ. Чтобы завершить пример с распределением Бернулли, выпишем:

$$\begin{aligned} &T(y) = y, \\ &a(\eta) = -\log(1-\phi) = \log(1+e^{\eta})\ [4], \\ &b(y) = 1. \end{aligned}$$

Этим мы формально доказали, что распределение Бернулли принадлежит экспоненциальному семейству.

1.2 Нормальное распределение

Перейдем теперь к рассмотрению нормального распределения. Вспомним, что мы выводили линейную регрессию из предположения о следующей модели зависимой переменной y:

$$y|x;\theta \sim \mathcal{N}(\theta^T x, 1), \tag{2}$$

т.е. σ^2 никак не влияло на θ и окончательный вид гипотезы $h_\theta(x) = \theta^T x$. Поэтому, чтобы упростить последующие выкладки, мы можем предположить, что $\sigma^2 = 1$[5]. Имеем:

$$p(y;\eta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\cdot\exp\left(\mu y - \frac{\mu^2}{2}\right).$$

Получаем, что нормальное распределение принадлежит экспоненциальному семейству с

$$\begin{aligned} &\eta = \mu, \\ &T(y) = y, \\ &a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}, \\ &b(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right). \end{aligned}$$

Много других классов распределений принадлежит экспоненциальному семейству: категориальное, Пуассона (используемое для моделирования случайной величины y, принимающей натуральные значения), гамма и экспоненциальное (для моделирования непрерывных неотрицательных случайных величин, как например, временны́х промежутков), бета и Дирихле (для распределений над отрезком [0,1]), и так далее. В следующем разделе мы опишем общий «рецепт» построения обобщенной линейной модели машинного обучения, в которой y (для заданных x и θ) моделируется с помощью одного из этих классов распределений.

---

[4] Мы просто подставили в эту формулу выражение для $\phi = \frac{1}{1+e^{-\eta}}$ и выполнили ряд тривиальных алгебраических преобразований.

[5] Иными словами, мы сейчас докажем, что не весь класс нормальных распределений, а только подкласс $\mathcal{N}(\mu, 1)$ принадлежит экспоненциальному семейству. Мы конечно же можем оставить σ в качестве параметра и рассмотреть более общий случай. Тогда η будет уже вектором, содержащим как μ, так и σ^2. В этом случае мы должны будем воспользоваться более общей формой экспоненциального распределения $p(y;\eta,\tau) = b(a,\tau)\exp((\eta^T T(y) - a(\eta))/c(\tau))$, где τ называется **параметром дисперсии** и для нормального распределения мы будем иметь $c(\tau) = \sigma^2$.

2. Построение ОЛМ

Предположим, мы хотим построить модель, позволяющую оценить (предсказать) количество клиентов (y), посещающих ваш интернет-магазин в течение часа, на основе определенных признаков (x), таких как акции, активность рекламной кампании, погода, день недели и т.д. Мы знаем, что распределение Пуассона как раз удобно использовать для моделирования *количества* посещений (в общем случае – некоторых событий) в среднем за период. Используя этот факт, как мы можем построить адекватную модель для решения нашей задачи? К счастью, распределение Пуассона принадлежит экспоненциальному семейству, поэтому мы можем построить на его основе ОЛМ. В этом разделе мы опишем метод построения ОЛМ моделей для решения подобной и многих других задач.

Сформулируем задачу в более общей постановке – пусть нам необходимо решить задачу регрессии или классификации, в которой требуется прогнозировать значение некоторой случайной величины y в зависимости от значений некоторого набора признаков x. Для построения ОЛМ мы сделаем следующие три предположения относительно нашей модели:

Предположение 1. $y|x;\theta \sim ExponentialFamily(\eta)$. То есть для заданных x и θ распределение случайной величины y принадлежит некоторому классу из экспоненциального семейства с параметром η.

Предположение 2. Результатом работы нашей модели, т.е. прогнозом, будем считать ожидаемое значение величины $T(y)$. Так как в большинстве наших задач $T(y) = y$, то данное условие означает, что мы хотели бы, чтобы обученная модель давала нам гипотезу вида $h(x) = E[y|x]$. Заметим, что данное условие выполняется как для линейной, так и для логистической регрессий. Например, для логистической регрессии мы имеем

$$h_\theta(x) = p(y = 1|x;\theta) = 0 \cdot p(y = 0|x;\theta) + 1 \cdot p(y = 1|x;\theta) = E[y|x;\theta].$$

Предположение 3. Естественный параметр η и входные значения x связаны линейным соотношением: $\eta = \theta^T x$ (или, если η – вектор, то $\eta_i = \theta_i^T x$).

Третье из этих предположений может показаться наименее обоснованным из вышеперечисленных, и его лучше рассматривать как "проектное решение", т.е. как придуманный нами способ связать входные значения модели x и естественный параметр распределения. Эти три предположения позволяют нам построить очень элегантный класс алгоритмов машинного обучения, а именно ОЛМ, которые обладают многими желанными свойствами, такими как, например, простота обучения. Кроме того, данный подход часто очень эффективен для моделирования y на основе многих известных нам распределений вероятностей. В частности, мы вскоре покажем, что логистическая и линейная регрессии могут быть выведены из ОЛМ.

2.1 Линейная регрессия

Для того чтобы показать, что линейная регрессия является частным случаем ОЛМ семейства моделей, рассмотрим случай, когда прогнозируемая величина y (частно называемая **переменной отклика** в терминологии ОЛМ) является непрерывной, и мы моделируем условное распределение y при заданном x как $\mathcal{N}(\mu, \sigma^2)$. Здесь μ может зависеть от x. Возьмем экспоненциальную форму записи нормального распределения (выведена в предыдущем разделе). В результате имеем:

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \mu \\ &= \eta \\ &= \theta^T x. \end{aligned}$$

Первое равенство следует из Предположения 2; второе – из того факта, что $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$, следовательно ожидаемое значение $y|x$ есть μ; третье равенство следует из Предположения 1 и связи между η и μ, которую мы вывели в первом разделе: $\mu = \eta$; ну и, наконец, третье равенство следует из Предположения 3.

2.2 Логистическая регрессия

Теперь рассмотрим логистическую регрессию. Так как нас интересует бинарная классификация, то $y \in \{0,1\}$. Следовательно, вполне естественно использовать распределение Бернулли для моделирования случайной величины $y|x$. В нашем выводе распределения Бернулли через экспоненциальное семейство мы получили, что $\phi = \frac{1}{1+e^{-\eta}}$. Далее заметим, что если $y|x;\theta \sim Bernoulli(\phi)$, то $E[y|x;\theta] = \phi$. Поэтому, используя Предположения 1-3, получаем:

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \phi \\ &= \frac{1}{1+e^{-\eta}} \\ &= \frac{1}{1+e^{-\theta^T x}}. \end{aligned}$$

В результате получаем сигмоиду в качестве гипотезы. Это дает нам еще одно объяснение, откуда взялась сигмоида в логистической регрессии: если предположить, что условное распределение y для заданного x[6] есть Бернулли, то данная форма гипотезы естественным образом следует из ОЛМ и определения экспоненциального семейства.

Напоследок познакомимся с часто используемой в ОЛМ терминологией. Функция g, выражающая математическое ожидание распределения как функцию от естественного параметра ($g(\eta) = E[Y(y);\eta]$) называется **канонической функцией отклика**. Обратная к ней функция g^{-1} называется **канонической функцией связи**. Таким образом, канонической функцией отклика для нормального распределения является тождественная функция, а для распределения Бернулли – логистическая функция.

[6] Всюду по тексту мы использовали запись «$y|x$» и фразу «условное распределение y для заданного x» синонимично. По-другому можно сказать, что мы моделируем случайную функцию от x, т.е. для каждого конкретного значения x мы получаем не фиксированный результат y, а некоторую случайную величину, среднее значение которой есть y. Этим, в частности, и объясняется смысл Предположения 2.