Машинное обучение, ФКН ВШЭ
Семинар №24

1 Обобщенные линейные модели

§1.1 Предпосылки

Многие рассмотренные в курсе модели так или иначе предполагают некоторое распределение целевой переменной $p(y \mid x)$ при фиксированном объекте x — в частности, для квадратичной функции потерь матожидание этого распределения в каждой точке x является оптимальным прогнозом. В других случаях прогноз может строиться иначе, но он по-прежнему может базироваться на нахождении указанного распределения.

Будем предполагать, что распределение $p(y \mid x)$ при фиксированном объекте x относится к некоторому параметрическому семейству и зависит от некоторого параметра $\theta(x)$, т.е. может быть представлено как $p(y \mid \theta(x))$. Записав правдоподобие выборки, можем получить оптимизационную задачу для нахождения $\theta_i = \theta(x_i)$ и соответствующих условных распределений:

$$\sum_{i=1}^{\ell} \log p(y_i \mid \theta_i) \to \max_{\{\theta_i\}_{i=1}^{\ell}} \tag{1.1}$$

Тем не менее, после решения этой задачи у нас не будет возможности восстановить распределение $p(y \mid \theta(x))$ для произвольного объекта x, не входящего в обучающую выборку. Чтобы решить эту проблему, попробуем как-то восстанавливать параметр для произвольного объекта, положив, к примеру, $\theta(x) = \langle w, x \rangle$, и будем вместо параметров θ_i для ограниченного числа объектов настраивать вектор w, позволяющий восстановить параметр $\theta(x)$ распределения $p(y \mid x)$ для произвольного объекта x:

$$\sum_{i=1}^{\ell} \log p(y_i \mid \theta(x_i)) = \sum_{i=1}^{\ell} \log p(y_i \mid \langle w, x_i \rangle) \to \max_{w} \tag{1.2}$$

После решения этой оптимизационной задачи у нас будет возможность восстанавливать распределение $p(y \mid x)$ для любых объектов и, тем самым, строить прогнозы для них на основании этого распределения. Рассмотрим на примерах, как работает данный подход.

1. Нормальное распределение: $p(y \mid \theta(x)) = \mathcal{N}(\theta(x), \sigma^2)$.

   Полагая, как было описано выше, $\theta(x) = \langle w, x \rangle$ и записывая задачу метода максимизации правдоподобия, приходим к задаче линейной регрессии для среднеквадратичной ошибки:

$$\sum_{i=1}^{\ell} -(y_i - \langle w, x_i \rangle)^2 \to \max_w \tag{1.3}$$

2. Распределение Бернулли: $p(y \mid \theta(x)) = \theta(x)^y (1 - \theta(x))^{1-y}$.

   Для применения описанного подхода нам необходимо предположить, что вероятность успеха может быть представлена как $\theta(x) = \langle w, x \rangle$, однако мы не можем так сделать из-за ограничений, накладываемых на параметр ($\theta(x) \in [0; 1]$). Чтобы разрешить эту ситуацию, применим к этому представлению сигмоиду: $\theta(x) = \sigma(\langle w, x \rangle)$. Позднее мы покажем, что применение метода максимального правдоподобия к такому представлению эквивалентно оптимизационной задаче логистической регрессии.

3. Распределение Пуассона: $p(y \mid \theta(x)) = e^{-\theta(x)} \frac{\theta(x)^y}{y!}$

   Как и в предыдущем случае, положить $\theta(x) = \langle w, x \rangle$ не представляется возможным, поскольку на параметр накладывается ограничение $\theta(x) > 0$. Вместо этого можем положить $\theta(x) = \exp(\langle w, x \rangle)$ и вновь решать задачу максимизации правдоподобия.

Как мы видим, описанный способ восстановления плотности при фиксированном объекте может быть применен не всегда из-за ограничений на параметры распределения. В некоторых случаях мы можем придумать индивидуальное решение проблемы, но нам хотелось бы уметь находить корректное представление $\theta(x)$ для любого параметрического распределения $p(y \mid x)$, которое может представлять для нас интерес. Такой подход предлагают *обобщенные линейные модели* (*generalized linear models*, *GLM*).

§1.2 Экспоненциальное семейство распределений

Везде далее мы будем считать, что распределение $p(y \mid x)$ принадлежит экспоненциальному семейству, т.е. может быть представлено в виде

$$p(y \mid \eta(x)) = \frac{h(y)}{Z(\eta(x))} \exp(\langle \eta(x), s(y) \rangle),$$

где $\eta(x) \in \mathbb{R}^N$ — вектор **натуральных** параметров распределения, $Z(\eta(x))$ — нормировочная константа, $s(y)$ — некоторая функция. Кроме того, нам будет достаточно рассматривать только те распределения, для которых $s(y) \equiv y$.

Отметим, что натуральные параметры распределения не обязаны совпадать с параметрами распределения в его «классической» записи (и, как правило, не совпадают) и могут принимать любое значение из $\mathbb{R}$, а потому мы можем приблизить их линейной формой: $\eta(x) = \langle w, x \rangle$.

Таким образом, записав распределение в экспоненциальной форме и решив задачу максимизации прадоподобия для вектора параметров w, для любого нового объекта x мы можем вычислить значение натурального параметра $\eta(x)$ распределения $p(y \mid x)$. Для восстановления искомого распределения в классическом виде нам осталось лишь найти функцию, связывающую натуральные и классические параметры распределения: $\eta(x) = \psi(\theta(x))$. Функция ψ называется *функцией связи*. В этом нам поможет доказанное на лекции утверждение.

Утверждение. В описанных выше предположения матожидание распределения $p(y \mid \eta(x))$ может быть вычислено следующим образом:

$$\frac{\partial}{\partial \eta(x)} \log Z(\eta(x)) = \mathbb{E}[y \mid \eta(x)].$$

Тогда, если параметр распределения может быть выражен через его матожидание (а этом верно для почти всех интересующих нас распределений), то мы получили искомую функцию связи. Заметим также, что для получения явного выражения для функции связи необязательно вычислять указанную выше производную — в случае, если изначально распределение записывается не в экспоненциальном виде (т.е. с использованием только классических параметров), то в процессе приведения распределения к нужному виду явно выписать функцию преобразования натуральных параметров в классические и наоборот.

Таким образом, построение обобщенной линейной модели заключается в следующих шагах:

1. Выбрать желаемое распределение $p(y \mid \theta(x))$ и записать его в экспоненциальном виде $p(y \mid \eta(x))$.

2. Вычислить функцию связи, получив выражение в процессе приведения распределения к экспоненциальному виду, или согласно утверждению выше.

3. Положить $\eta(x) = \langle w, x \rangle$ и записать оптимизационную задачу (к примеру, метод максимального правдоподобия: $\sum_{i=1}^{\ell} \log(y_i \mid \langle w, x_i \rangle) \to \max_w$), решить ее и получить оптимальное значени w^*.

4. Имея w^*, для любого нового объекта x можем восстановить классический параметр распределения $\theta(x) = \psi^{-1}(\eta(x)) = \psi^{-1}(\langle w^*, x \rangle)$ и, тем самым, распределение $p(y \mid \theta(x))$, позволяющее построить прогноз для этого объекта.

§1.3 Построение обобщенных линейных моделей

Попробуем применить описанный подход на практике для некоторых параметрических семейств распределений — в частности, для распределений, решение для которых мы получили не вполне честно в начале семинара.

Задача 1.1. *Пусть* $p(y \mid x) \sim Ber(y \mid \theta(x)) = \theta^y(1-\theta)^{1-y}$. *Выпишите оптимизационную задачу метода максимизации правдоподобия для обобщенной линейной модели и матожидание распределения* $p(y \mid w)$.

Решение. Сначала представим распределение в экспоненциальном виде (везде далее для упрощения записи в выкладках вместо $\theta(x), \eta(x)$ будем писать θ, η соответственно):

$$p(y \mid \theta(x)) = \theta^y(1-\theta)^{1-y} = (e^{\ln\theta})^y \frac{1-\theta}{(1-\theta)^y} = (1-\theta)\frac{e^{y\ln\theta}}{e^{y\ln(1-\theta)}} = (1-\theta)e^{y\ln\frac{\theta}{1-\theta}}. \quad (1.4)$$

Отсюда получаем $h(y) = 1$ и выражение для функции связи $\eta(x) = \psi(\theta(x)) = \ln\frac{\theta(x)}{1-\theta(x)}$. В данном случае нам удалось обойтись без дифференцирования нормировочной константы распределения, поскольку в процессе приведения к экспоненциальному виду нам удалось выписать функцию связи в явном виде, — в случае, если распределение было бы приведено сразу в экспоненциальном виде, мы могли бы получить то же выражение для функции связи путем дифференцирования.

Выразим классический параметр распределения через натуральный:

$$e^\eta = \frac{\theta}{1-\theta} \Leftrightarrow e^\eta - \theta e^\eta = \theta \Leftrightarrow \theta(1+e^\eta) = e^\eta \Leftrightarrow \theta = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}} = \sigma(\eta).$$

Напомним, что в начале семинара для этого распределения мы выбрали именно сигмоиду для преобразования скалярного произведения в параметр распределения — сейчас мы получили тот же результат при помощи GLM.

Запишем оптимизационную задачу ММП:

$$\sum_{i=1}^{\ell} \log p(y_i \mid \theta(x_i)) = \sum_{i=1}^{\ell} \log \theta(x_i)^{y_i}(1-\theta(x_i))^{1-y_i} = \sum_{i=1}^{\ell} y_i \log\theta(x_i) + (1-y_i)\log(1-\theta(x_i)) =$$

$$\sum_{i=1}^{\ell} y_i \log\sigma(\eta(x_i)) + (1-y_i)\log(1-\sigma(\eta(x_i))) =$$

$$\sum_{i=1}^{\ell} y_i \log\sigma(\langle w, x_i\rangle) + (1-y_i)\log(1-\sigma(\langle w, x_i\rangle)) \to \max_w.$$

Заметим, что полученная задача совпадает с оптимизационной задачей логистической регрессии. Решив её, мы получим оптимальное значение w^*, при помощи которого мы сможем строить прогнозы для новых объектов, предсказывая, к примеру, матожидание распределения $p(y \mid x)$:

$$a(x) = \mathbb{E}[y \mid \theta(x)] = \theta(x) = \sigma(\langle w^*, x\rangle).$$

■

Задача 1.2. *Пусть решается задача классификации на* K *классов, и целевая переменная является* K*-мерным бинарным вектором, у которого ровно одна компонента*

равна 1. Рассмотрим мультиномиальное распределение для такой целевой переменной:

$$p(y \mid x) = B(y \mid \theta(x)) = \prod_{j=1}^{K} \theta_j(x)^{y_j},$$
$$y \in \{0,1\}^K, \quad \sum_{j=1}^{K} y_j = 1,$$
$$\theta(x) \in [0;1]^K, \quad \sum_{j=1}^{K} \theta_j(x) = 1.$$

Приведите распределение к экспоненциальному виду и выпишите формулу для параметров распределения $\theta_j(x)$ с использованием оптимальных значений w_j при фиксированном объекте x.

Решение. Заметим, что $y_K = 1 - \sum_{j=1}^{K-1} y_j$ и представим распределение в экспоненциальном виде:

$$B(y \mid \theta(x)) = \prod_{j=1}^{K-1} \theta_j^{y_j} \theta_K^{y_K} = \exp\left(\sum_{j=1}^{K-1} y_j \log\theta_j + \left(1 - \sum_{j=1}^{K-1} y_j\right)\log\theta_K\right) =$$
$$\exp\left(\sum_{j=1}^{K-1} y_j(\log\theta_j - \log\theta_K) + \log\theta_K\right) = \exp(\log\theta_K)\exp\left(\sum_{j=1}^{K-1} y_j \log\frac{\theta_j}{\theta_K}\right) =$$
$$\left\{\log\frac{\theta_K}{\theta_K} = 0\right\} = \theta_K \exp\left(\sum_{j=1}^{K} y_j \log\frac{\theta_j}{\theta_K}\right) = \theta_K \exp\left(\left\langle y, \left\{\log\frac{\theta_j}{\theta_K}\right\}_{j=1}^{K}\right\rangle\right).$$

Таким образом, $\eta_j(x) = \psi(\theta_j(x)) = \log\frac{\theta_j}{\theta_K}$. Выразим классические параметры через натуральные:

$$e^{\eta_j} = \frac{\theta_j}{\theta_K} \Leftrightarrow \theta_K e^{\eta_j} = \theta_j \mid \sum_{j=1}^{K}$$
$$\theta_K \sum_{j=1}^{K} e^{\eta_j} = \sum_{j=1}^{K} \theta_j = 1 \Leftrightarrow \theta_K = \frac{1}{\sum_{j=1}^{K} e^{\eta_j}}$$
$$\Rightarrow \theta_j = \theta_K e^{\eta_j} = \frac{e^{\eta_j}}{\sum_{j=1}^{K} e^{\eta_j}}$$

Записав оптимизационную задачу и решив её, получим оптимальные значения w_j^* (в данном случае для каждого из K параметров будем иметь свой вектор), при помощи которых можем записать формулу для классических параметров распределения:

$$\theta_j(x) = \frac{\exp(\langle w_j, x\rangle)}{\sum_{j=1}^{K}\exp(\langle w_j, x\rangle)},\ j = \overline{1, K-1},$$
$$\theta_K(x) = \frac{1}{\sum_{j=1}^{K}\exp(\langle w_j, x\rangle)}.$$

■

Заметим, что преобразование объекта x в натуральные параметры распределения $p(y \mid \eta(x))$ в рассмотренном случае можно рассматривать как полный слой нейросети без использования нелинейности — эту идею можно развить и использовать для некоторого параметра любое другое приближение вместо $\langle w, x\rangle$. Именно это представление отвечает за линейность в названии подхода, но можно также рассматривать и другие представления для усложнения модели.

Задача 1.3. *Пусть целевая переменная при фиксированном объекте имеет распределение Пуассона* $p(y \mid x) = \mathcal{P}(y \mid \theta(x)) = e^{-\theta(x)}\frac{\theta(x)^y}{y!}$. *Запишите оптимизационную задачу метода максимального правдоподобия при использовании GLM и вычислите градиенты полученного функционала.*

Решение.

Как всегда, сначала приведем распределение к экспоненциальному виду:

$$p(y \mid x) = e^{-\theta}\frac{\theta^y}{y!} = \frac{e^{-\theta}}{y!}e^{y\log\theta} = \frac{\frac{1}{y!}}{e^{\theta}}\exp(y\log\theta).$$

Отсюда получаем $h(y) = \frac{1}{y!}$ и выражение для функции связи $\eta(x) = \psi(\theta(x)) = \log\theta(x)$, откуда $\theta(x) = \exp(\eta(x))$. Заметим, что в этом случае мы вновь получили преобразование, которое мы использовали в случае распределения Пуассона, теперь при помощи GLM. С учетом полученных преобразований, распределение в экспоненциальном виде принимает следующий вид:

$$p(y \mid \eta(x)) = \frac{\frac{1}{y!}}{e^{e^{\eta(x)}}}\exp(y\eta(x)).$$

Запишем оптимизационную задачу, полагая $\eta(x) = \langle w, x\rangle$ и опуская слагаемые, не зависящие от w:

$$Q(w, X) = \sum_{i=1}^{\ell}\left(\log\frac{1}{y_i!} - e^{\eta(x_i)} + y_i\eta(x_i)\right) = \sum_{i=1}^{\ell}(-\exp(\langle w, x_i\rangle) + y_i\langle w, x_i\rangle \to \max_w$$

$$\sum_{i=1}^{\ell}(\exp(\langle w, x_i\rangle) - y_i\langle w, x_i\rangle \to \min_w$$

Вычислим градиент полученного функционала:

$$\nabla_w Q(w, X) = \sum_{i=1}^{\ell}\exp(\langle w, x_i\rangle)x_i - \sum_{i=1}^{\ell}y_i x_i = \sum_{i=1}^{\ell}\left(\exp(Xw)_i - y_i\right)x_i = X^T\left(\exp(Xw) - y\right).$$

■