Обобщённые линейные модели

5 апреля 2018 г.

Мы уже изучили множество моделей машинного обучения и научились применять их на практике. Здесь мы найдём общий подход к конструированию новых алгоритмов. С его помощью получим теоретическое обоснование классических методов машинного обучения, таких как линейная и логистическая регрессии. Их обобщения на нейросети и многомерные случаи ответов: $y \in \mathbb{R}^k$ или $y \in \{0, 1\}^k$. Выведем функции сигмоиды и софтмакса.

В литературе данный раздел математики представлен как обобщённые линейные модели (GLM). Однако все идеи легко переносятся и активно используются в моделях сколь угодно сложных и нелинейных.

Материал во многом опирается на конспект[1] Andrew Ng курса[2] CS229 Стенфорда по машинному обучению и даже скорее является его вольным переводом.

1 Введение

Рассматривая произвольную статистическую модель машинного обучения, будь то логистическая регрессия, свёрточная нейросеть или SVM, мы неявно предполагаем некоторое вероятностное распределение $p(x, y)$ в пространстве всевозможных объектов и ответов $\mathbb{X} \times \mathbb{Y}$. Для некоторых моделей распределения выписываются аналитически, равно как и по аналитической записи распределения можно строить модель. Вероятностная постановка даёт большой простор для исследований, оценивания рисков, нахождения взаимосвязей. Позволяет лучше разобраться в методах.

Совместное распределение объектов и ответов $p(x, y)$ определяет *генеративную* модель. Восстанавливая его, мы получаем возможность сэмплировать новые объекты – создавать картинки и рукописный текст, синтезировать речь и новые молекулы для лекарств. Однако восстановить совместное $p(x, y)$ вовсе не просто. А для решения задачи прогнозирования достаточно условного распределения ответов $p(y|x)$ для данного объекта. Такие модели называются *дискриминативными*. Упрощая задачу дальше, мы будем рассматривать только параметризованные распределения $p(y|x, \theta(X)) = p(y|\theta)$. Параметры θ которых будем настраивать по обучающей выборке X. Например, распределение Бернулли возвращает вероятность θ выпадения орла y и вероятность $(1 - \theta)$ выпадения решки $(1 - y)$.

$$\mathcal{B}(y|\theta) = \theta^y (1 - \theta)^{1-y}, \quad \theta \in [0, 1], \quad y \in \{0, 1\}$$

[1] http://cs229.stanford.edu/notes/cs229-notes1.pdf

[2] http://cs229.stanford.edu/syllabus.html

Здесь среднее значение равно $\mathbb{E}y = \theta$ и совпадает с параметром модели. Равно как и нормальное распределение $\mathcal{N}(y|\mu, \sigma^2)$ параметризуется своим средним $\mathbb{E}y = \mu$ и дисперсией $\mathbb{D}y = \sigma^2$. Нас главным образом будет интересовать среднее, поскольку *оптимальный байесовский алгоритм* в частном случае сопоставляет предсказываемые объекты их математическому ожиданию: $y^\star(x) = \mathbb{E}[y|x]$.

Итак, нам нужно восстановить параметр θ – вероятность выпадения орла y. В самом простом случае, мы можем сказать, что это будет отношение числа выпадений орлов к общему числу бросков монеты: $\theta = \frac{N(y)}{N}$. Сейчас мы занимаемся машинным обучением, и у нас наверняка найдётся выборка, матрица объектов-признаков X. Поэтому решение в виде среднего значения отклика $\theta = \frac{1}{N}\sum\limits_n y_n$ нам уже не подойдёт.

Мы будем смотреть на признаковое описание объекта: кто бросает монетку, в какое время суток, какой рукой, с закрытыми ли глазами, и что у него при этом получается: орёл или решка, y или $1 - y$. Дальше скажем, что у нас линейная модель, что признаки объекта x сворачиваются с обучаемыми весами w через скалярное произведение: $\langle w, x\rangle$, а итоговую вероятность будем доставать сигмоидой: $\theta = \sigma(\langle w, x\rangle)$. Остаётся только записать многомерную задачу оптимизации методом максимального правдоподобия, обучить с её помощью веса w по всем людям X и по результатам их бросков Y. И тогда мы сможем для каждого конкретного человека x уверенно сказать, с какой вероятность он всё-таки получит своего долгожданного орла y.

Основной вопрос заключается в том, как найти подходящую *функцию связи* ψ, отображающую объекты x в параметр распределения θ. И что нам даёт правильность этого выбора. Недолго думая, для Бернулли мы решили взять сигмоиду: $\psi = \sigma$. А теперь нам предстоит решить эту непростую задачу в общем случае. Напрямую мы её решить не можем, нам потребуется перейти в новое пространство *натуральных параметров* η, и доказать в нём несколько замечательных свойств.

2 Параметризация распределений

§2.1 Экспоненциальная форма

В курсе математического анализа мы рисовали графики функций и считали площади, описываемые их фигурами. И часто нам удобней оказывалось переходить из декартовой системы координат (x, y) в полярную (ρ, φ), в то время как сама функция по сути оставалась прежней. Новое пространство обладало хорошими свойствами, которые позволяли легко решить задачу. И здесь имеют место схожие идеи.

Распределение вероятностей как и любую другую функцию можно параметризовывать по разному. За основу, *стандартные* параметры θ, часто берут моменты случайной величины: среднее и дисперсию. Мы же будем работать с *натуральными* параметрами η. Они возникают из *экспоненциальной формы* распределения.

Опр. 2.1. *Параметрическое семейство $p(y|\eta)$ принадлежит экспоненциальному классу распределений $\Leftrightarrow \exists$ факторизация:* $p(y|\eta) = \dfrac{h(y)}{g(\eta)} e^{\langle \eta,\, s(y)\rangle}$

$$h(y) \geq 0, \quad s(y) - \forall \text{ функция}, \quad g(\eta) = \int h(y) e^{\langle \eta, s(y)\rangle} dy$$

В данных обозначениях η будут *натуральными* параметрами, $g(\eta)$ – хотя и является функцией, но называется *нормировочной константой* поскольку не зависит от случайной величины y и осуществляет нормировку вероятности, $h(y)$ – *свободный множитель*, и $s(y)$ называется *достаточной статистикой*.

Задача 2.1. *Найдите экспоненциальную форму распределение Бернулли:*

$$B(y|\theta) = \theta^y(1-\theta)^{1-y}$$

Решение. Нам нужно преобразовать функцию $B(y|\theta)$ к виду $B(y|\eta) = \frac{h(y)}{g(\eta)}e^{\eta s(y)}$. Первым делом занесём переменные в экспоненту:

$$B(y|\theta) = (1-\theta)e^{y \ln \frac{\theta}{1-\theta}}$$

Тогда необходимые функции выписываются следующим образом:

Достаточная статистика: $s(y) = y$

Натуральный параметр модели: $\eta = \ln \frac{\theta}{1-\theta}$ и обратно: $\theta = \frac{1}{1+e^{-\eta}}$

Нормировочная константа: $g(\eta) = 1 + e^{\eta}$ и свободный множитель: $h(y) = 1$

Сопоставляя полученные функции, получаем экспоненциальную форму:

$$B(y|\eta) = (1+e^{\eta})^{-1}e^{\eta s(y)}$$

■

Параметры θ и η связаны друг с другом сигмоидой: $\theta = \sigma(\eta)$. Дальше нам предстоит выяснить, почему так получается, и вывести результат для общего случая.

Не для всех распределений существует экспоненциальное представление. А если и существует, то оно не единственно. Невозможно представить равномерную случайную величину y и параметр θ через их произведение в степени экспоненты:

$$\mathcal{U}(y|\theta) = \frac{1}{\theta}[0 \le y \le \theta] = \frac{1}{\theta}e^{\ln[0 \le y \le \theta]}$$

Необходимым условием существования экспоненциальной формы является независимость области определения случайной величины от параметра, что нарушается в $[0 \le y \le \theta]$. Однако распределения можно представлять и в более общем виде:

$$p(y|\theta) = \frac{h(y)}{g(\theta)}f(\theta, s(y))$$

где f – произвольная измеримая, быть может недифференцируемая функция.

§2.2 Достаточные статистики и критерий факторизации

Опр. 2.2. *Часто рассматривают не одну случайную величину y, а целую выборку: $Y = (y_1 ... y_n)$. Статистикой $s(Y)$ назовём произвольную функцию от выборки Y. Что также допускает рассматривать* статистику *как функцию одной случайной величины $s(y)$.*

Опр. 2.3. *Статистика $s(y)$ для оценки параметра θ является достаточной, если любая другая статистика случайной величины y не добавляет никакой новой информации о значении параметра θ. Формально в терминах байесовского подхода это утверждение выглядит следующим образом: $p(\theta|y, s(y)) = p(\theta|s(y))$*

Утв. 1. Критерий факторизаци (Рональд Фишер[3])

Функция $s(y)$ является *достаточной статистикой* параметра θ для данного распределения $p(y|\theta) \Leftrightarrow \exists$ факторизация: $p(y|\theta) = \dfrac{h(y)}{g(\theta)} f(\theta, s(y))$

Доказательство.

Основатель современной статистики Рональд Фишер, при достаточно сильных ограничениях на исследуемую функцию, обосновывал своё утверждение следующим образом: при поиске максимума правдоподобия, в получившемся уравнении $\frac{d}{d\theta} \ln p(y|\theta) = 0$ все параметры θ будут выражаться только через некоторую функцию от $s(y)$. А сами элементы выборки y в оценке параметров участия принимать уже не будут. Проделаем эти несложные выкладки:

$$\ln p(y|\theta) = \ln h(y) - \ln g(\theta) + \ln f(\theta, s(y))$$

Максимальное правдоподобие: $\ln p(y|\theta) \to \max\limits_{\theta}$

Вогнутость и дифференцируемость: $\dfrac{d}{d\theta} \ln p(y|\theta) = 0$

$$\frac{d}{d\theta} \ln g(\theta) = \frac{d}{d\theta} \ln f(\theta, s(y))$$

Тогда статистика $s(y)$ содержит в себе всю информацию о параметрах модели θ.

■

Возвращаясь к равномерному распределению $\mathcal{U}(y|\theta) = \frac{1}{\theta}[0 \le y \le \theta]$, получаем:

$$h(y) = 1,\ g(\theta) = \tfrac{1}{\theta} \text{ и } f(s(y), \theta) = [0 \le y \le \theta] \Rightarrow s(y) = y$$

Задача 2.2. *Найдите достаточную статистику выборки из равномерных величин.*

Решение.

$$\mathcal{U}(Y|\theta) = \prod_{j=1}^{n} \mathcal{U}(y_n|\theta) = \prod_{j=1}^{n} \frac{1}{\theta}[0 \le y_n \le \theta] = \frac{1}{\theta^n}[0 \le y_{\min}][y_{\max} \le \theta]$$

$$h(y) = [0 \le y_{\min}],\ g(\theta) = \tfrac{1}{\theta^n} \text{ и } f(s(y), \theta) = [0 \le y_{\max} \le \theta] \Rightarrow s(y) = y_{\max}$$

■

Мы смогли существенно упростить задачу оценки параметра θ. Вместо целой выборки $Y = (y_1 ... y_n)$, где n может быть очень велико, достаточно знать всего один её элемент $y_{\max}$. И хотя даже функция плотности $p(Y|\theta)$ недифференцируема из-за индикатора, мы можем оценить методом максимального правдоподобия: $\hat{\theta} = y_{\max}$. В этом красота достаточных статистик.

[3] https://en.m.wikipedia.org/wiki/Ronald_Fisher

Заметим, что для экспоненты $f(y,\eta) = e^{\eta s(y)}$ из оценки ММП на η получается:

$$\frac{d}{d\eta}\ln g(\eta) = \frac{d}{d\eta}\ln f(\eta, s(y)) = \frac{d}{d\eta}\ln e^{\eta s(y)} = \frac{d}{d\eta}\eta s(y) = s(y)$$

И производная логарифма нормировочной константы $\frac{d}{d\eta}\ln g(\eta) = \hat{s}(y)$ – оценка правдоподобия натурального параметра η совпадает с его достаточной статистикой $s(y)$. Существует гораздо более сильный результат, определяющий равенство условному матожиданию: $\frac{d}{d\eta}\ln g(\eta) = \mathbb{E}[s(y)|\eta]$. Но прежде чем мы к нему перейдём, нам необходимо обозначить несколько важных результатов.

§2.3 Оптимальный байесовский алгоритм

Напомним, что *функционалом среднего риска* $R(a)$ называется матожидание функции потерь $\mathcal{L}(y, a(x))$ по всем парам (x, y) при использовании алгоритма $a(x)$:

$$R(a) = \mathbb{E}\mathcal{L}(y, a(x)) = \int_{\mathbb{Y}}\int_{\mathbb{X}} \mathcal{L}(y, a(x))p(x,y)dxdy$$

Оптимальный байесовский алгоритм минимизирует функционал среднего риска:

$$y^\star = a^\star(x) = \arg\min_a R(a) \quad \forall x \in \mathbb{X}$$

Ранее, на частных примерах мы показали, что функции потерь $\mathcal{L}(y,\mu)$ тесно связаны с распределениями ответов $p(y|\mu)$. Так, минимизация квадратичной функции соответствует максимизации правдоподобия нормального распределения:

$$\min_\mu \sum_n (y_n - \mu(x_n))^2 \Leftrightarrow \max_\mu \prod_n N(y_n|\mu(x_n), \sigma)$$

.

Мы доказали, что оптимальная байесовская функция регрессии для квадратичной функции потерь $\mathcal{L}(y,\mu) = (y-\mu)^2$ или, что тоже самое, для нормального распределения $p(y|\mu,\sigma)$, имеет вид условного матожидания ответов y для данного объекта x и обученного алгоритма – параметра среднего значения $\mu = \mathbb{E}y$:

$$a^\star(x) = \mathbb{E}[y|x,\theta] = \int_{\mathbb{Y}} yp(y|x,\theta)dy$$

Отталкиваясь от нормального распределения, в общем случае для произвольного $p(y|x,\theta)$ справедлива аналогичная оценка на достаточную статистику $s(y)$.

Утв. 2. *Оптимальным байесовским алгоритмом* $a_\star(x)$ распределения $p(y|x,\eta)$ является матожидание достаточной статистики этого распределения: $\mathbb{E}[s(y)|x,\eta]$

$$a^\star(x) = \mathbb{E}[s(y)|x,\eta] = \int_{\mathbb{Y}} s(y)p(y|x,\eta)dy$$

Здесь есть некоторая сложность, что при построении таких алгоритмов мы не всегда сможем оценивать отклик $y^\star$, а только функцию от него: $s^\star(y)$. И полученная оценка – вовсе не то же самое, что доказанная выше оценка максимального правдоподобия: $\hat{s}(y) = \arg\max\limits_{\eta} p(y|x, \eta)$. Будем различать их крышкой и звёздочкой. Оптимальный байесовский алгоритм работает именно с матожиданием. Далее, обусловленность распределения по x будем опускать, $p(y|x, \eta) = p(y|\eta)$, если только это не приводит к неоднозначностям.

§2.4 Матожидание достаточной статистики

Утв. 3. Производная логарифма нормировочной константы является матожиданием достаточной статистики: $\dfrac{d}{d\eta} \ln g(\eta) = \mathbb{E}[s(y)|\eta]$ для $s(y)$, $\eta \in \mathbb{R}$.
И в многомерном случае: $\dfrac{\partial}{\partial \eta_j} \ln g(\eta) = \mathbb{E}[s_j(y)|\eta_j]$ для $s(y)$, $\eta \in \mathbb{R}^k$.

Доказательство.

Начнём искать частную производную $\ln g(\eta)$ по η_j и получим нужный результат:

$$\frac{\partial \ln g(\eta)}{\partial \eta_j} = \frac{1}{g(\eta)} \frac{\partial}{\partial \eta_j} g(\eta) = \frac{1}{g(\eta)} \frac{\partial}{\partial \eta_j} \int h(y) e^{\langle \eta, s(y) \rangle} dy = \frac{1}{g(\eta)} \int h(y) s_j(y) e^{\langle \eta, s(y) \rangle} dy =$$

$$= \int s_j(y) \frac{h(y)}{g(\eta)} e^{\langle \eta, s(y) \rangle} dy = \int s_j(y) p(y|\eta) dy = \mathbb{E}[s_j(y)|\eta]$$

В векторной форме: $\nabla_\eta \ln g(\eta) = \mathbb{E}[s(y)|\eta]$

■

Тогда, для того чтобы найти оптимальный байесовский алгоритм данного распределения, достаточно привести это распределение в экспоненциальную форму и посчитать производную логарифма нормировочной константы!

Следующую задачу предлагается решить самостоятельно.

Задача 2.3. *Пусть* $s(y) \in \mathbb{R}$. *Докажите, что* $\dfrac{d^2}{d\eta^2} \ln g(\eta) = \mathbb{D}[s(y)|\eta] > 0$

И в векторной форме для случая $s(y) \in \mathbb{R}^k$: $\nabla^2_\eta \ln g(\eta) = \mathrm{Cov}[s(y)|\eta] > O$

Задача 2.4. *Для распределения Бернулли найдите оптимальный байесовский алгоритм как матожидание достаточной статистики и выпишите функционал качества методом максимального правдоподобия.*

Рассмотрите случай линейной модели, когда натуральные параметры задаются линейной функцией объектов и весов: $\eta(x) = \langle w, x \rangle$. *Полученные результаты будут строгим теоретическим обоснованием логистической регрессии.*

Решение. Мы уже нашли экспоненциальную форму распределения в задаче (2.1):

$$B(y|\eta) = (1+e^{\eta})^{-1}e^{\eta s(y)}$$

Достаточная статистика яляется тождественной функцией: $s(y) = y$. Тогда мы можем построить алгоритм, который будет предсказывать значения $a^\star(x) = y^\star$.

Натуральный параметр модели: $\eta = \ln\dfrac{\theta}{1-\theta}$, обратно: $\theta = \dfrac{1}{1+e^{-\eta}}$

И нормировочная константа: $g(\eta) = (1+e^{\eta}) \Rightarrow \ln g(\eta) = \ln(1+e^{\eta})$

$$y^\star = \mathbb{E}[y|\eta] = \frac{d}{d\eta}\ln g(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}, \quad \text{получаем сигмоиду}$$

Распишем функционал качества:

$$\prod_n B(y_n|\theta(x_n)) = \prod_n \theta(x_n)^{y_n}(1-\theta(x_n))^{1-y_n} \to \max_\theta \qquad \{\text{log-loss}\}$$

$$\sum_n \ln B(y_n|\theta(x_n)) = \sum_n y_n \ln\theta(x_n) + (1-y_n)\ln(1-\theta(x_n)) \to \max_\theta$$

Выражая через натуральный параметр, получаем log-loss (со знаком минус):

$$\theta(x_n) = \frac{1}{1+e^{-\eta(x_n)}} = \sigma(\eta(x_n)) \Rightarrow \sum_n y_n \ln\sigma(\eta(x_n)) + (1-y_n)\ln(1-\sigma(\eta(x_n))) \to \max_\eta$$

Итоговый функционал качества линейной модели – лог-лосс:

$$\eta(x_n) = \langle w, x_n\rangle \Rightarrow \sum_n y_n \ln\sigma(\langle w, x_n\rangle) + (1-y_n)\ln(1-\sigma(\langle w, x_n\rangle)) \to \max_w$$

■

§2.5 Функция связи параметров распределений

Вы можете заметить, что в распределении Бернулли производная логарифма нормировочной константы $\mathbb{E}y = \frac{d}{d\eta}\ln g(\eta) = \frac{1}{1+e^{-\eta}}$ совпадает с обратной функцией натурального параметра: $\mathbb{E}y = \theta = \frac{1}{1+e^{-\eta}}$. Тогда через матожидание $\mathbb{E}y$ наши параметры связываются вместе: $\frac{d}{d\eta}\ln g(\eta) = \theta$. В этом есть свой смысл – существует взаимно однозначное соответствие $\psi(\eta) = \theta$ (например, сигмоида $\psi = \sigma$) между *средним* θ и *натуральным* η, которое мы уже так долго ищем.

Допустим, распределение вероятностей параметризовано *стандартным* параметром, отвечающим за среднее значение: $\theta = \mathbb{E}y$:{например $B(y|\theta) = \theta^y(1-\theta)^{1-y}$}. И также существует экспоненциальная форма с *натуральной* параметризацией η: $\{B(y|\eta) = (1+e^{\eta})^{-1}e^{\eta s(y)}\}$. Соответствие вытекает из следующих простых фактов:

1. В предположении о параметризации распределения: $\frac{d}{d\eta}\ln g(\eta) = \theta = \mathbb{E}y$

2. Показано (задача 2.3), что $\dfrac{d^2}{d\eta^2}\ln g(\eta) = \mathbb{D}[s(y)|\eta] > 0$. И тогда $\ln g(\eta)$ – строго выпуклая функция.

3. Из строгой выпуклости существует взаимно однозначное соответствие между первой производной $\frac{d}{d\eta}\ln g(\eta)$ функции и её аргумента η. Это и есть искомое соответствие $\psi(\eta) = \frac{d}{d\eta}\ln g(\eta)$. Из первого пункта $\psi(\eta) = \theta$

Функция $\psi(\eta) = \theta$, связывающая средние и натуральные параметры, единственная и обратимая: $\psi^{-1}(\theta) = \eta$. В литературе её называют *функцией связи* (link function). Выбор конкретной функции связи определяет свой алгоритм прогнозирования, и, вообще говоря, может быть произвольной. Например, гиперболическим тангенсом и любой другой функцией активации[4].

И если из экспоненциального представления уже найдена функциональная зависимость между средним значением θ и натуральным параметром η, то эта связь единственная. И уже можно не считать производную логарифма нормировочной константы: $\frac{d}{d\eta}\ln g(\eta)$ – обе эти функции будут совпадать. Но так не всегда получается, например, для нормального распределения параметры не выражаются явно, а только через производную.

Задача 2.5. *Найдите линейный прогнозирующий алгоритм как матожидание достаточной статистики для одномерного нормального распределения.*

Решение.

$$N(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}}$$

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}, \quad s_1(x) = x, \quad s_2(x) = x^2$$

$$\text{Тогда } g(\eta) = e^{-\frac{\eta_2^2}{4\eta_1}}\frac{\sqrt{2\pi}}{-2\eta_1} \Rightarrow \frac{\partial}{\partial\eta_1}\ln g(\eta) = -\frac{\eta_2}{2\eta_1} = \mu = y^\star$$

$$\eta = \langle w, x\rangle \Rightarrow y^\star = a(w, x) = \langle w, x\rangle$$

■

Аналогично в многомерном случае получается $y = Wx \in \mathbb{R}^k$, где веса записаны по строкам: $W \in \mathbb{R}^{k\times d}$. Или для всей выборки сразу: $Y = XW \in \mathbb{R}^{n\times k}$, где объекты записаны по строкам, а веса – по столбцам. $X \in \mathbb{R}^{n\times d}, W \in \mathbb{R}^{d\times k}$

3 Конструирование новых алгоритмов

Подняв некоторую теорию, теперь мы можем по заданным распределениям строить новые модели машинного обучения! Для этого нужно только определить, как настраивать параметры: найти функционал качества $L(y, x, w) \to \min\limits_w$.

И определить, как получать ответы с этими параметрами для данных объектов: найти оптимальный байесовский алгоритм прогнозирования $y^\star = a(x, w)$.

В обобщенных линейных моделях мы предпологаем, что натуральные параметры модели η выражаются в виде скалярного произведения: $\eta(x) = \langle w, x\rangle$, если $\eta \in \mathbb{R}$ и $\eta(x) = Wx$, если $\eta \in \mathbb{R}^k$. Здесь $x, w \in \mathbb{R}^d$ – объект выборки и вектор весов,

[4] https://en.wikipedia.org/wiki/Activation_function

$W^{k\times d}$ – матрица весов модели, если записать по строкам веса для каждого класса. Нейронные сети моделируют параметры η композицией линейных преобразований с нелинейными активациями – функциями связи.

Всего несколько несложных шагов к успеху:

1. Привести распределение в экспоненциальную форму: $p(y|\eta) = \dfrac{h(y)}{g(\eta)} e^{\langle \eta, s(y) \rangle}$

2. Найти выражение для параметра среднего значения $\theta = \psi(\eta) = \mathbb{E}[y|\eta]$ через натуральный параметр в явном виде, если это возможно. Тогда: $y^\star = \psi(\eta)$ – так можно будет получать ответы алгоритма.

   Или рассчитать через градиент нормировочной константы матожидание достаточной статистики: $\nabla_\eta \ln g(\eta) = \mathbb{E}[s(y)|\eta]$. Тогда: $s^\star(y) = \nabla_\eta \ln g(\eta)$.

3. Расписать задачу оптимизации методом максимального правдоподобия $\prod\limits_j p(y_j|\eta(x_j)) \to \max\limits_{\eta(x_j)}$ через натуральные параметры $\eta(x_j)$.

4. Выбрать способ извлечения информации из объектов выборки: $\eta(x_j)$

Линейная регрессия Пуассона с параметром интенсивности λ позволяет строить регрессию натурального отклика: $y \in \mathbb{N}$. Классический пример: число звонок на телефонную линию в теории массового обслуживания. А ещё можно предсказывать, сколько людей придёт на пляж в зависимости от солнечной *интенсивности*:

$$\mathcal{P}(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{(y!)^{-1}}{e^\lambda} e^{y \ln \lambda}$$

Единственное жёсткое ограничение метода в том, что у Пуассона среднее и дисперсия равны друг другу $\mathbb{E}[y|\eta] = \mathbb{D}[y|\eta] = \lambda$ и это должно отражаться в данных. В отличие от обычной регрессии с нормальным распределением, где μ и σ^2 независимы.

Задача 3.1. *Постройте прогнозирующий алгоритм $y^\star = a(w, x) \in \mathbb{N}$ в предположении равенства среднего и дисперсии: $\mathbb{E}[y|\eta] = \mathbb{D}[y|\eta]$ и линейности модели $\eta = \langle w, x \rangle$. Выпишите обновление весов w через градиентный спуск.*

Решение. Выберем $y \sim \mathcal{P}(\lambda)$

достаточная статистика: $s(y) = y$

нормировочная константа: $g(\eta) = e^{e^\eta}$

натуральный параметр модели: $\eta = \ln \lambda$ и $\lambda = e^\eta$

$y^\star = a(x, w) = e^{\langle w, x \rangle}$ – функция прогнозирования

Дальше найдём функционал качества и его градиенты:

$$\prod_j p(y_j|\lambda(x_j)) = h(y) \prod_j e^{-\lambda(x_j)} \lambda(x_j)^{y_j} \to \max_\lambda$$

$$\ln \sum_j p(y_j|\lambda(x_j)) = -\sum_j \lambda(x_j) + \sum_j y_j \ln \lambda(x_j) + c \to \max_\lambda$$

Дальше подставим функцию связи в наш алгоритм:

$$\lambda(x_j) = e^{\langle w, x_j \rangle} \Rightarrow -\sum_j e^{\langle w, x_j \rangle} + \sum_j y_j \langle w, x_j \rangle + c \to \max_w$$

$$L(w, x, y) = \sum_j e^{\langle w, x_j \rangle} - \sum_j y_j \langle w, x_j \rangle + c \to \min_w$$

Найдём градиент:

$$\nabla_w L = \sum_j e^{\langle w, x_j \rangle} x_j - \sum_j y_j x_j = \sum_j e^{(Xw)_j} x_j - \sum_j y_j x_j = X^T(\exp Xw - y)$$

$$w_{k+1} = w_k - \alpha_{k+1} X^T(\exp Xw_k - y) \text{ -- формулы обновления весов через GD}$$

■

GLM'ы широко используются компаниями для оценки рисков и прогнозов. Например, в этих задачах мы явно предполагаем определённый закон распределения ответов $p(y|\eta)$, который отличен от нормального:

- Судоходная компания может использовать регрессию Пуассона: $p(y|\lambda) = \dfrac{e^{-\lambda}\lambda^y}{y!}$ для описания количества повреждений судов нескольких типов, построенных в разное время. Полученная модель помогает определить, суда каких типов чаще всего повреждаются.
- Страховая компания может использовать показательную регрессию $p(y|\lambda) = \lambda e^{-\lambda y}$ для описания страховых исков о повреждении автомобилей. Полученная модель помогает определить факторы, которые дают наибольший размер исков от клиентов компании.

4 Основные задачи

§4.1 Регрессия и классификация

Одномерные регрессия $y \in \mathbb{R}$ и классификация $y \in \{0, 1\}$ выводятся из своих естественных распределений $p(y|\eta)$ – нормального и Бернулли:

- Нормальное распределение:

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(y-\mu)^2}{2\sigma^2}}, \text{ где } \mu = \eta(x), \quad \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{++}$$

- Распределение Бернулли:

$$\mathcal{B}(y|\theta) = \theta^y (1-\theta)^{1-y}, \text{ где } \theta = \frac{1}{1 + e^{-\eta(x)}}, \quad , \quad \theta \in [0, 1]$$

Многомерные задачи регрессии $y \in \mathbb{R}^k$ и классификации $y \in \{0, 1\}^k$ также легко выводятся из своих распределений, многомерных аналогов $\mathcal{N}$ и $\mathcal{B}$:

- Многомерное нормальное распределение:

$$\mathcal{N}(y|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\dfrac{1}{2}\langle \Sigma^{-1}(y-\mu),(y-\mu)\rangle}$$

$\mu = \eta(x)$ – тождественная активация; $x \in \mathbb{R}^d; y, \mu \in \mathbb{R}^k; \Sigma \in \mathbb{R}^{k\times k}$

- Многомерное распределение Бернулли (категориальное):

$$\mathcal{B}(y|\theta) = \prod_{j=1}^{k} \theta_j^{y_j}, \quad \theta_j = \frac{e^{\eta_j(x)}}{\sum\limits_{j=1}^{k} e^{\eta_j(x)}} \quad \text{– softmax активация}$$

или в векторной форме: $\theta = \langle 1_k, \exp\eta(x)\rangle^{-1}\exp\eta(x) \quad \{\langle 1_k, \exp Wx\rangle^{-1}\exp Wx\}$

$\theta \in [0,1]^k$ – вероятности принадлежности своему классу: $\sum\limits_{j=1}^{k}\theta_j = 1$

$y \in \{0,1\}^k$ – бинарный вектор классов с одной единичкой: $\sum\limits_{j=1}^{k} y_j = 1$

Есть ряд простых фактов, существенно облегчающих решение задач машинного обучения. И оправдывающих то, что мы, как правило, не задумываемся о распределении ответов y. Только смотрим – вероятность ли нам нужно предсказать, класс или произвольное вещественное число.

Итак, для обычной регрессии, заданной нормальным распределением $N(y|\mu,\sigma^2)$ достаточно только нормального распределения ошибок $\varepsilon \sim N(0,\sigma^2)$, а не самого отклика y. Действительно, в силу линейности $y \sim N(\mu,\sigma) \Leftrightarrow y-\mu = \varepsilon \sim N(0,\sigma^2)$. Тогда ответы выборки y в задаче регрессии могут быть распределены как угодно, важна только *нормальность остатков* ε для спрогнозированных μ. Проверять регрессионные остатки на нормальность всегда полезно. И если это не так, то имеет смысл попробовать другую модель или изменить постановку самой задачи. А для классификации, отвечающей распределению Бернулли важно лишь только, чтобы каждый объект принадлежал к одному своему классу $y \in \{0,1\}^k$, что, как правило, выполняется для большинства задач.

§4.2 Вывод softmax-активации

Задача 4.1. *Если распределение Бернулли соответствует броску монетки, то его многомерным аналогом будет один бросок k-гранного кубика, где k – число классов. Назовём его категориальным. К слову, биномиальное распределение – это n-кратное подбрасывание монетки. А мультиномиальное – n-кратное подбрасывание k-гранного кубика. Ну, давайте начнём.*

Решение.

Функция вероятности: $\mathcal{B}(y|\theta) = \displaystyle\prod_{j=1}^{k}\theta_j^{y_j}$, при условиях:

$$\sum_{j=1}^{k}\theta_j = 1, \quad \sum_{j=1}^{k} y_j = 1, \quad \theta \in [0,1]^k, \quad y \in \{0,1\}^k$$

Выпишем экспоненциальную форму распределения:

$$y_k = 1 - \sum_{j=1}^{k-1} y_j \Rightarrow C(y|\theta) = \theta_k^{1-\sum\limits_{j=1}^{k-1} y_j} \prod_{j=1}^{k-1} \theta_j^{y_j} = e^{\sum\limits_{j=1}^{k-1} y_j \ln \theta_j + (1-\sum\limits_{j=1}^{k-1} y_j) \ln \theta_k} =$$

$$= e^{\sum\limits_{j=1}^{k-1} y_j \ln \theta_j - \sum\limits_{j=1}^{k-1} y_j \ln \theta_k + \ln \theta_k} = e^{\sum\limits_{j=1}^{k-1} y_j \ln \frac{\theta_j}{\theta_k} + \ln \theta_k} = \frac{1}{\theta_k} e^{\sum\limits_{j=1}^{k-1} y_j \ln \frac{\theta_j}{\theta_k}}$$

Тогда натуральные параметры $\eta_j = \ln \dfrac{\theta_j}{\theta_k}$, причём $\eta_k = \ln \dfrac{\theta_k}{\theta_k} = 0$

Для распределения Бернулли параметр соответствует матожиданию: $\theta = \mathbb{E} y$. Поэтому нам достаточно всего лишь перевыразить θ через η:

1. $e^{\eta_j} = \frac{\theta_j}{\theta_k}$
2. $\theta_k e^{\eta_j} = \theta_j$
3. $\theta_k \sum\limits_{j=1}^{k} e^{\eta_j} = \sum\limits_{j=1}^{k} \theta_j = 1$
4. $\theta_k = \frac{1}{\sum\limits_{j=1}^{k} e^{\eta_j}}$
5. $\theta_j = \frac{e^{\eta_j}}{\sum\limits_{j=1}^{k} e^{\eta_j}}$ – profit

∎

5 Заключение

На самом деле, теорию обобщённых линейных моделей (GLM) обычно[5] начинают рассказывать с того, что есть в жизни простая регрессия $y^\star = \langle w, x \rangle$. И что есть непростая – можно ввести некоторое нелинейное отображение, связывающее линейный прогноз с целевой переменной: $y^\star = \psi \langle w, x \rangle$ – функцию связи. Вообще говоря, её можно выбирать как угодно, например, почему бы не взять сигмоиду. Дальше рассказывают, что функция, которая получается дифференцированием логарифма нормировочной константы называется *канонической функцией связи* (canonical link function) распределения. Вот и всё, никакой математики, никакого праздника. Рассказ обычно продолжается в сторону описательной статистики: гипотез, тестов и доверительных интервалов. А мы ведь занимаемся машинным обучением[6], правда?

[5] http://www.machinelearning.ru/wiki/images/4/47/Psad_otherreg_17.pdf
[6] http://www.machinelearning.ru/wiki/images/a/a6/BMMO11_13.pdf