







# Байесовские методы. Лекция 8. Обобщённые линейные модели (GLM)

Целищев М.А.

МГУ им. М. В. Ломоносова  
Факультет вычислительной математики и кибернетики  
Кафедра математической статистики

весна 2021

## Список литературы

-  J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
-  R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan, Second Edition. Chapman and Hall CRC, 2020.
-  O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
-  K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
-  A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
-  S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

## Задача регрессии

По вектору признаков  $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$  хотят предсказать значение целевой переменной  $t \in \mathbb{R}$ .

В наличии исследователя имеется тренировочная выборка

$$D_{\text{tr}} = \{(t_i, \mathbf{x}_i)\}_{i=1}^n, \quad t_i \in \mathbb{R}, \quad \mathbf{x}_i \in \mathbb{R}^d,$$

то есть конечный набор пар (целевая переменная, вектор признаков).

Требуется построить алгоритм, который по новому вектору признаков  $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$  будет делать прогноз на целевую переменную  $t_{\text{new}}$ .

Обозначим  $t = (t_1, \dots, t_n)^T$  — вектор целевых переменных,  
 $X = (x_{ij}) \in \mathbb{R}^{n \times d}$  — матрица признаков (строка  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$  которой соответствует целевой переменной  $t_i$ ).

## Линейная регрессия

Вспомним, что в линейной регрессии предполагалась следующая вероятностная модель:

$$t \sim \mathcal{N}(X\mathbf{w}, \sigma^2 I),$$

где  $\mathbf{w} \in \mathbb{R}^d$  — вектор весов,  
или, что равносильно,

$$t = X\mathbf{w} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

где  $\varepsilon$  — шум в наблюдениях, подчиняющийся нормальному закону с нулевым средним и дисперсией  $\sigma^2$ .

В классической статистике ищут оценки макс. правдоподобия:

$$\mathbf{w}_{\text{ML}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|t - X\mathbf{w}\|^2 = (X^T X)^{-1} X^T t, \quad \sigma_{\text{ML}}^2 = \frac{1}{n} \|t - X\mathbf{w}_{\text{ML}}\|^2,$$

либо в случае проблем с обратимостью, ищут псевдорешение

$$\mathbf{w}_{\lambda}^* = (X^T X + \lambda I)^{-1} X^T t,$$

с параметром регуляризации  $\lambda > 0$ .

## Линейная регрессия, байесовский подход

Затем мы выяснили, откуда берётся регуляризация — из априорного распределения параметров при использовании байесовского подхода. Байесовская модель выглядит так:

$$\begin{cases} t|X, \mathbf{w}, \sigma \sim \mathcal{N}(X\mathbf{w}, \sigma^2 I), & \sigma \in \mathbb{R}_+, \\ \mathbf{w} \sim \mathcal{N}(\mu_0, \Sigma_0), & \mu_0 \in \mathbb{R}^d, \Sigma_0 \in \mathbb{R}^{d \times d}, \end{cases}$$

При этом, апостериорное распределение весов  $\mathbf{w}$  имеет явный (в силу сопряжённости априорного распределения и модели) вид:

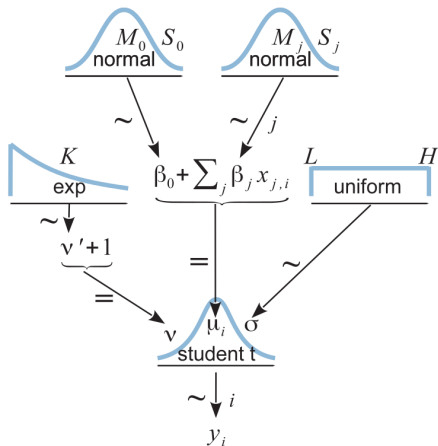
$$\mathbf{w}|t, X, \sigma \sim \mathcal{N}(\mu_n, \Sigma_n),$$

где  $\Sigma_n = (\Sigma_0^{-1} + \frac{1}{\sigma^2} X^T X)^{-1}$ ,  $\mu_n = \Sigma_n (\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} X^T t)$ .

Если же по каким-то причинам многомерное нормальное распределение не подходит в качестве априорного для  $\mathbf{w}$ , то можно воспользоваться и любым другим, но апостериорное в явном виде выразить не получится...

## Иерархические модели

... зато получится насэмплировать выборку из апостериорного распределения параметров с помощью MCMC! При этом правдоподобие можно брать не обязательно нормальным.



Например, для представленной слева модели используется распределение Стюдента, у которого толще хвосты (т.е. оно лучше описывает выбросы)

## Номинальные признаки

А что делать, если один из признаков оказался *номинальным*?

То есть признак описывает группу, к которой относится наблюдение. Такие признаки ещё называются *категориальными*.

Скажем,  $j$ -ый признак  $x_{ij}$  для всех  $i$  принимает значения от 1 до  $K$  и описывает группу, к которой относится  $i$ -ое наблюдения.

Пример: предсказание изменения цены нефти за день, номинальный признак — день недели.

Ясно, что использовать его как есть (т.е. значения  $1 \dots K$ ) и работать как с обычным метрическим признаком **нельзя**, потому что при этом на него накладывается метрическая структура (четверг «в два раза больше» вторника), что является бессмыслицей.

Для исправления ситуации применяют процедуру *one-hot-encoding*:

$$x_{ij} = k \quad \longrightarrow \quad x_{ij} = (0, \dots, 0, \underbrace{1}_k, 0, \dots, 0) \in \{0, 1\}^K.$$

## Номинальные признаки

$$x_{ij} = k \quad \longrightarrow \quad x_{ij} = (0, \dots, 0, \underbrace{1}_k, 0, \dots, 0) \in \{0, 1\}^K.$$

Тогда линейным признаком для целевой переменной  $t_i$  будет

$$a_i = \dots + \sum_{k=1}^K w_j[k] \cdot x_{i,j}[k] + \dots$$

Иными словами, увеличится размерность признаков  $X$ , а в остальном линейная модель останется неизменной.

А ещё можно сделать преобразование

$$\beta_{j,0} := \frac{1}{K} \sum_{k=1}^K w_j[k], \quad \beta_{j,k} := w_j[k] - \beta_{j,0}, \quad k = 1 \dots K,$$

и тогда  $a_i = \dots + \beta_{j,0} + \sum_{k=1}^K \beta_{j,k} \cdot x_{i,j}[k] + \dots$ , а новые веса  $\beta_{j,k}$  будут показывать

вклад  $k$ -го класса в значение целевой переменной (заметим, что  $\sum_{k=1}^K \beta_{j,k} = 0$ ).



## Задача классификации

А что, если номинальными будут не признаки, а целевая переменная  $t_i$ ? Иными словами, надо по вектору признаков  $\mathbf{x}_i \in \mathbb{R}^d$  отнести объект к одному из  $K$  классов.

Для простоты рассмотрим случай двух классов, т.е.  $t_i \in \{0, 1\}$ .

Пример: предсказать, выживет ли человек, заражённый вирусом.

Признаками могут служить характеристики человека (пол, вес, возраст, наличие сопутствующих болезней и пр.)

Технически никто не мешает использовать уже рассмотренную линейную модель

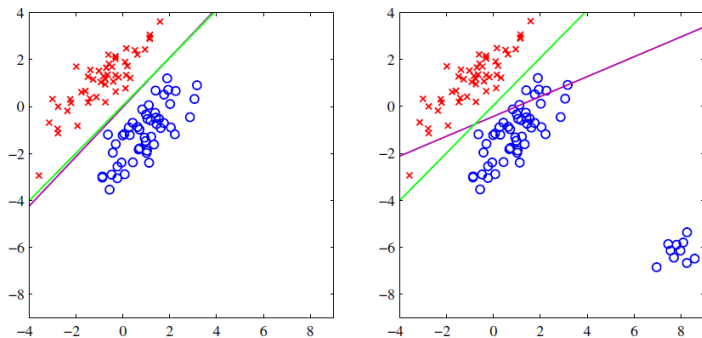
$$t \approx X\mathbf{w}$$

и построить оценку методом наименьших квадратов:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|t - X\mathbf{w}\|^2 = (X^T X)^{-1} X^T t.$$

# Проблема

Картинка из «Pattern Recognition and Machine Learning», Bishop:



**Figure 4.4** The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

Почему так получилось? Потому что метод наименьших квадратов штрафует за предсказания, далёкие от нуля (для 0-го класса) или от 1 (для 1-го класса), даже если они находятся «с правильной стороны» от 0 или 1.

## Проблема

Итак, метод наименьших квадратов **НЕЛЬЗЯ** применять для задач классификации. Так что же, линейные модели не пригодны для классификации? Не совсем. Классическая линейная регрессия предполагала, что целевая переменная  $t$  распределена нормально, но это точно не случай классификации.

Нужна другая вероятностная модель:

$$t_i \sim \text{Be}(\theta(\mathbf{x}_i)), \quad i = 1 \dots n,$$

где  $\theta(\mathbf{x})$  — вер-ть попасть в класс 1, если признаки равны  $\mathbf{x}$ .

Можно ли в качестве  $\theta(\mathbf{x})$  взять линейную ф-цию от признаков?

$$\theta(\mathbf{x}) \stackrel{?}{=} \sum_{j=1}^d w_j x_j, \quad \mathbf{x} \in \mathbb{R}^d.$$

**Нет**, поскольку  $\theta(\mathbf{x})$  принимает значения из отрезка  $[0, 1]$ , а правая часть ничем не ограничена. Тем не менее, совсем от линейности отказываться не хочется...

## Байесовская классификация

Представим, что нам известна не выборка для каждого из двух классов  $C_0$  и  $C_1$ , а даже сами распределения  $p_0(\mathbf{x})$  и  $p_1(\mathbf{x})$  значений признаков для каждого класса. Теперь нужно для точки  $\mathbf{x}_{\text{new}}$  предсказать её класс.

$$\begin{aligned} p(C_1|\mathbf{x}_{\text{new}}) &= \frac{p(\mathbf{x}_{\text{new}}|C_1) p(C_1)}{p(\mathbf{x}_{\text{new}}|C_0) p(C_0) + p(\mathbf{x}_{\text{new}}|C_1) p(C_1)} = \\ &= \frac{p_1(\mathbf{x}_{\text{new}}) p(C_1)}{p_0(\mathbf{x}_{\text{new}}) p(C_0) + p_1(\mathbf{x}_{\text{new}}) p(C_1)} = \frac{1}{1 + \dots} = \frac{1}{1 + e^{-a(\mathbf{x}_{\text{new}})}}, \end{aligned}$$

где  $p(C_k)$  — априорные вероятности попасть в каждый класс (например, равные 0.5, если нет априорных предпочтений, или доле элементов выборки, попавших в соответствующий класс), а

$$a(\mathbf{x}_{\text{new}}) = \ln \frac{p(C_1)p_1(\mathbf{x}_{\text{new}})}{p(C_0)p_0(\mathbf{x}_{\text{new}})}$$

— логарифм шансов (*log-odds*). Чем больше  $a(\mathbf{x}_{\text{new}})$ , тем более вероятно, что  $\mathbf{x}_{\text{new}}$  из класса  $C_1$ , и наоборот. Заметим также, что  $a(\mathbf{x}_{\text{new}})$  знакопеременна.

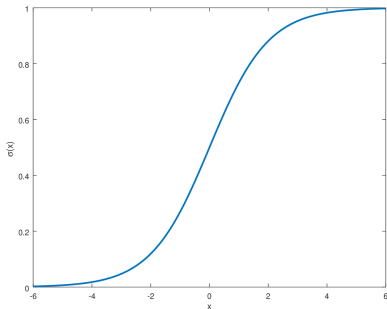
## Логистическая функция

Так давайте с помощью линейной модели предсказывать не вероятность попадания в класс  $C_1$ , а логарифм шансов  $a(\mathbf{x})$ .

Тогда

$$p(C_1|\mathbf{x}) = \frac{1}{1 + e^{-a(\mathbf{x})}} = \sigma(a(\mathbf{x})),$$

где  $\sigma(x) = \frac{1}{1 + e^{-x}}$ ,  $x \in \mathbb{R}$ , — логистическая функция (сигмоид).



ДЗ: показать, что

$$\sigma(-x) = 1 - \sigma(x),$$

$$\sigma'(x) = \sigma'(-x) = \sigma(x)(1 - \sigma(x))$$

## Логистическая регрессия

Итак, вероятностная модель для классификации (с 2 классами):

$$\begin{cases} t_i \sim \text{Be}(\theta(\mathbf{x}_i)), & i = 1, \dots, n, \\ \theta(\mathbf{x}) = \sigma(a(\mathbf{x})) = \sigma(\mathbf{w}^T \mathbf{x}), \end{cases}$$

где  $\mathbf{w} \in \mathbb{R}^d$  — вектор весов (вектор параметров модели).

Как его оценить? Методом максимального правдоподобия!

$$\mathcal{L}(D_{\text{tr}}; \mathbf{w}) = \prod_{i=1}^n \theta(\mathbf{x}_i)^{t_i} (1 - \theta(\mathbf{x}_i))^{1-t_i}$$

$$\ln \mathcal{L}(D_{\text{tr}}; \mathbf{w}) = \sum_{i=1}^n \left[ t_i \ln \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - t_i) \ln (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \right]$$

$$\frac{\partial}{\partial w_j} \ln \mathcal{L} = \sum_{i=1}^n \left[ t_i (1 - \sigma(\dots)) - (1 - t_i) \sigma(\dots) \right] x_{ij} = \sum_{i=1}^n \left[ t_i - \sigma(\mathbf{w}^T \mathbf{x}_i) \right] x_{ij},$$

## Логистическая регрессия

$$\nabla_{\mathbf{w}} \ln \mathcal{L}(D_{\text{tr}}; \mathbf{w}) = \sum_{i=1}^n \left[ t_i - \sigma(\mathbf{w}^T \mathbf{x}_i) \right] \mathbf{x}_i$$

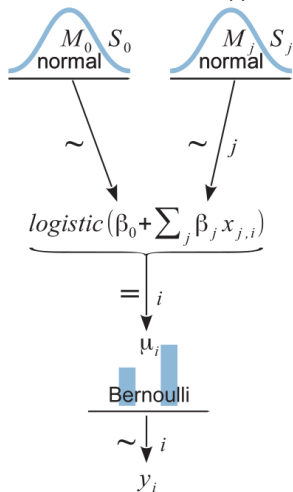
Увы, в явном виде решить уравнение  $\nabla_{\mathbf{w}} \ln \mathcal{L} = 0$  нельзя.

Поэтому оценку максимального правдоподобия  $\mathbf{w}_{\text{ML}}$  находят численными методами оптимизации первого (градиентный подъём) или второго (метод Ньютона-Рапсона) порядков.

Но это точечная оценка, а хотелось бы что-то большее. Почему бы не применить байесовскую науку и здесь?

## Логистическая регрессия

Байесовская модель логистической регрессии:



$$\begin{cases} t_i | \mathbf{x}_i, \mathbf{w} \sim \text{Be}(\theta(\mathbf{x}_i)), & i = 1, \dots, n, \\ \theta(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}), \\ w_j \sim \mathcal{N}(M_j, S_j), & j = 1 \dots d. \end{cases}$$

Тогда апостериорное распределение

$$p(\mathbf{w} | X, t) \propto \text{Likelihood} \times \text{Prior} = \mathcal{L}(D_{\text{tr}}; \mathbf{w}) p(\mathbf{w}).$$

Теперь можно насэмплировать сколько угодно точек  $\mathbf{w}_{[1]}, \dots, \mathbf{w}_{[N]}$  из этого апостериорного распределения с помощью техники MCMC.



## Логистическая регрессия, предсказания

Теперь нужно сделать предсказание для новой точки  $\mathbf{x}_{\text{new}}$ :

$$\begin{aligned} p(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, D_{\text{tr}}) &= \int p(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{w}) p(\mathbf{w} | D_{\text{tr}}) d\mathbf{w} = \\ &= \int \sigma(\mathbf{w}^T \mathbf{x}_{\text{new}}) p(\mathbf{w} | D_{\text{tr}}) d\mathbf{w}. \end{aligned}$$

Поскольку у нас есть выборка  $\mathbf{w}_{[1]}, \dots, \mathbf{w}_{[N]}$  из апостериорного распределения  $p(\mathbf{w} | D_{\text{tr}})$ , то оценкой искомого предсказания будет:

$$\hat{p}(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, D_{\text{tr}}) = \frac{1}{N} \sum_{r=1}^N \sigma(\mathbf{w}_{[r]}^T \mathbf{x}_{\text{new}}),$$

то есть усреднение по сэмплированным из апостериорного распределения параметрам  $\mathbf{w}$ .

По эргодической теореме эта оценка будет несмещённой и состоятельной.

## Случай $K$ классов

А что, если классов не два, а больше? То есть  $t_i \in \{1, \dots, K\}$ .

Хочется построить обобщение логистической регрессии.

Модель данных:  $\mathbf{P}(t_i = k | \mathbf{x}_i) = p_k(\mathbf{x}_i)$ ,  $k = 1 \dots K$ , или

$$t_i | \mathbf{x}_i \sim \text{Categorical}(p_1(\mathbf{x}_i), \dots, p_K(\mathbf{x}_i))$$

Теперь нужно уметь предсказывать вероятности  $p_k(\mathbf{x})$  попадания в каждый класс для объекта с признаками  $\mathbf{x}$ .

Как и в случае двух классов, линейную модель для  $p_k(\mathbf{x})$  использовать нельзя (несовпадение доменов), но и отказываться совсем от линейности не хочется, а хочется, чтобы линейные функции от признаков  $a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} = \sum_{j=1}^d w_{kj} x_j$  были тем больше, чем больше уверенность модели в том, что наблюдение  $\mathbf{x}$  из класса  $k$ . Заметим, что число параметров при этом равно  $K \times d$ .

## Softmax-функция

$$\text{Softmax}(a) := \left( \dots, \frac{e^{a_k}}{e^{a_1} + \dots + e^{a_K}}, \dots \right)^T, \quad a \in \mathbb{R}^K,$$

то есть функция принимает вектор чисел  $a \in \mathbb{R}^K$  и выдаёт вектор неотрицательных чисел той же размерности, в сумме дающих 1 (поэтому их можно интерпретировать как вероятности).

Заметим, что  $\lim_{\tau \rightarrow +\infty} \text{Softmax}(\tau a) = (0, \dots, 0, 1, 0, \dots, 0)^T$ , где единица стоит в позиции  $k^*(a) = \arg \max_k a_k$ , то есть softmax есть сглаженная версия максимума.

Тогда можно предсказывать вектор вероятностей:

$$(p_1(\mathbf{x}), \dots, p_K(\mathbf{x})) = \text{Softmax}(a_1(\mathbf{x}), \dots, a_K(\mathbf{x})),$$

где, как и ранее,  $a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} = \sum_{j=1}^d w_{kj} x_j$ .

При этом выполняется такое свойство: отношение  $\frac{p_k(\mathbf{x})}{p_l(\mathbf{x})} = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\exp(\mathbf{w}_l^T \mathbf{x})}$  зависит только от весов  $\mathbf{w}_k, \mathbf{w}_l$ .

## Неидентифицируемость модели

Заметим, что если сдвинуть каждый вес  $w_{kj} \mapsto w_{kj} + c_j$ , то вероятность попадания в каждый класс  $k$

$$\frac{e^{(\mathbf{w}_k + \mathbf{c})^T \mathbf{x}}}{e^{(\mathbf{w}_1 + \mathbf{c})^T \mathbf{x}} + \dots + e^{(\mathbf{w}_K + \mathbf{c})^T \mathbf{x}}} = \frac{e^{\mathbf{w}_k^T \mathbf{x}}}{e^{\mathbf{w}_1^T \mathbf{x}} + \dots + e^{\mathbf{w}_K^T \mathbf{x}}}$$

не изменятся. Иначе:  $\text{Softmax}(\mathbf{a} + \delta) = \text{Softmax}(\mathbf{a}) \quad \forall \mathbf{a} \in \mathbb{R}^K, \delta \in \mathbb{R}.$

Это чрезвычайно плохо, поскольку разные значения весов  $W$  дают одну и ту же модель. Такая модель называется *неидентифицируемой*, и её надо лечить.

Чаще всего выбирают один класс  $r$  (называемый *референсным*), для которого принудительно полагают  $\mathbf{w}_r = \mathbf{0} \in \mathbb{R}^d$ .

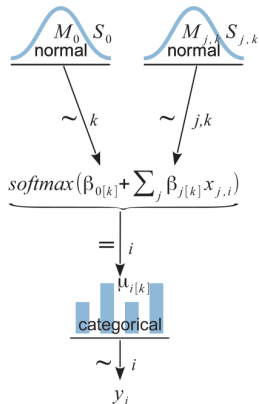
В таком случае модель становится *идентифицируемой*, и можно оценивать её параметры  $W$ .

Кстати, можно проверить, что при таком подходе в случае двух классов Softmax-регрессия превращается в логистическую регрессию.

## Softmax-регрессия

Оценка максимального правдоподобия для параметров  $W$ , как и в случае логистической регрессии, явно не вычисляется, и приходится использовать численные методы оптимизации.

А можно строить и байесовскую модель:



$$\begin{cases} t_i | \mathbf{x}_i, \mathbf{w} \sim \text{Categorical}(\mu(\mathbf{x}_i)), & i = 1, \dots, n, \\ \mu(\mathbf{x}) = \text{Softmax}(\mathbf{w}_1^T \mathbf{x}, \dots, \mathbf{w}_{K-1}^T \mathbf{x}, 0), \\ w_{kj} \sim \mathcal{N}(M_{kj}, S_{kj}), & k = 1 \dots K-1, j = 1 \dots d. \end{cases}$$

Тогда апостериорное распределение

$$p(W|X, t) \propto \text{Likelihood} \times \text{Prior},$$

где правдоподобие равно

$$p(t|X, W) = \prod_{i=1}^n \mu_{t_i}(\mathbf{x}_i),$$

и можно запускать MCMC.

## Ordinal Targets

А что, если целевая переменная  $t_i$  принимает значения от 1 до  $K$ , но при этом эти значения «упорядочены»?

Пример: предсказать рейтинг фильма от 1 (отвратительно) до 7 (превосходно), основываясь на некоторых признаках (год и страна выпуска, жанр, режиссёр и проч.).

В принципе, можно использовать модель Softmax-регрессии для решения этой задачи. Но хотелось бы не отбрасывать структуру порядка целевой переменной.

Предположим, что для каждого наблюдения с признаками  $\mathbf{x}$  есть *латентная* (ненаблюдаемая) переменная  $\eta(\mathbf{x})$  и такие *отсечки* (пороги)  $\theta_1, \dots, \theta_{K-1}$  её значений, что

$$t_i = k \iff \eta(\mathbf{x}_i) \in [\theta_{k-1}, \theta_k), \quad \text{где } \theta_0 = -\infty, \theta_K = +\infty.$$

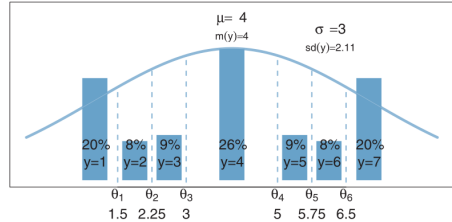
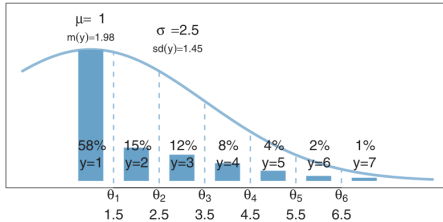
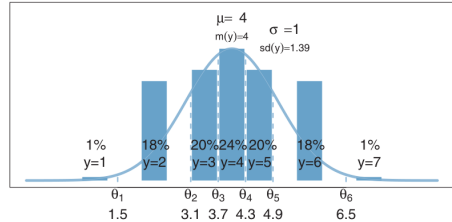
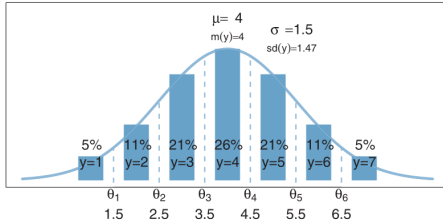
Для простоты будем считать, что  $\eta(\mathbf{x})$  нормально распределена (с параметрами  $\mu$  и  $\sigma$ , зависящими от  $\mathbf{x}$ )

# Ordinal Probit Regression

Если  $\Phi$  — функция станд. норм. распределения, то

$$\mathbf{P}(t_i = k) = \mathbf{P}(\eta \in [\theta_{k-1}, \theta_k)) = \Phi\left(\frac{\theta_k - \mu}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \mu}{\sigma}\right).$$

Возможные ситуации при различных  $\mu$ ,  $\sigma$  и отсечках  $\theta$ :



## Ordinal Probit Regression

Будем предсказывать значение центра распределения латентной переменной  $\eta(\mathbf{x})$  с помощью линейной модели:  $\mu(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ .

Итого полная вероятностная модель выглядит так:

$$\begin{cases} t_i \sim \text{Categorical} \left( q_1(\mathbf{x}_i), \dots, q_K(\mathbf{x}_i) \right), & i = 1 \dots n, \\ q_k(\mathbf{x}) = \Phi \left( \frac{\theta_k - \mu(\mathbf{x})}{\sigma} \right) - \Phi \left( \frac{\theta_{k-1} - \mu(\mathbf{x})}{\sigma} \right), & k = 1 \dots K, \\ \mu(\mathbf{x}) = \mathbf{w}^T \mathbf{x}. \end{cases}$$

В этой модели параметрами будут  $\mathbf{w} \in \mathbb{R}^d$ ,  $\theta_1 < \dots < \theta_{K-1}$  и  $\sigma > 0$ .

Увы, такая модель *неидентифицируема* по двум причинам:

- при умножении всех  $\mathbf{w}$ ,  $\theta_k$ ,  $\sigma$  на константу  $\tau > 0$  получаем ту же модель;
- если первый признак  $x_1$  — константа 1 (чаще всего так и бывает), то при добавлении к весу  $w_1$  и всем отсечкам  $\theta_k$  константы  $\delta \in \mathbb{R}$  модель не изменится.

Поэтому фиксируют какие-нибудь два параметра, например:

$$\theta_1 = 1.5, \quad \theta_{K-1} = K - 0.5.$$



## Ordinal Probit Regression

Итак, идентифицируемая вероятностная модель ordinal probit-регрессии выглядит так:

$$\begin{cases} t_i | \mathbf{x}_i, \mathbf{w}, \theta, \sigma \sim \text{Categorical} \left( q_1(\mathbf{x}_i), \dots, q_K(\mathbf{x}_i) \right), & i = 1 \dots n, \\ q_k(\mathbf{x}) = \Phi \left( \frac{\theta_k - \mu(\mathbf{x})}{\sigma} \right) - \Phi \left( \frac{\theta_{k-1} - \mu(\mathbf{x})}{\sigma} \right), & k = 1 \dots K, \\ \theta_0 = -\infty, \quad \theta_1 = 1.5, \quad \theta_{K-1} = K - 0.5, \quad \theta_K = +\infty, \\ \mu(\mathbf{x}) = \mathbf{w}^T \mathbf{x}. \end{cases}$$

Параметрами будут  $\mathbf{w} \in \mathbb{R}^d$ ,  $\theta_2 < \dots < \theta_{K-2}$  и  $\sigma > 0$ .

Для подсчёта оценки макс. правдоподобия снова используются численные методы оптимизации.

А если на параметры задать априорные распределения, то можно исследовать апостериорное распределение параметров, насемплировав выборку с помощью MCMC.

Единственная проблема: задать априорное распределение, которое бы «уважало» упорядоченность отсечек  $\theta_2 < \dots < \theta_{K-2}$ .

## Count Targets

А если целевая переменная  $t$  принимает неотрицательные значения  $0, 1, 2, \dots$ , но теоретически не ограничены сверху?

Пример: предсказание числа исполненных заявок на бирже за единицу времени. Признаками могут быть текущий индекс волатильности, время до конца торгового дня и др.

Какое распределение использовать для построения вероятностной модели в таком случае? Подойдёт **распределение Пуассона**:

$$\mathbf{P}(t = k) = \frac{\lambda}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots,$$

где параметр  $\lambda > 0$  называется *интенсивностью* и имеет смысл среднего числа произошедших событий за единицу времени.

Будем считать, что интенсивность зависит от признаков  $\mathbf{x} \in \mathbb{R}^d$ .  $\lambda(\mathbf{x}) \stackrel{?}{=} \mathbf{w}^T \mathbf{x}$  — нельзя, т.к. несовпадение доменов. Тогда имеет смысл моделировать линейной функцией не  $\lambda$ , а  $\ln \lambda$ :

$$\ln \lambda(\mathbf{x}) = \mathbf{w}^T \mathbf{x}.$$

## Poisson Regression

Тогда вероятностная модель будет выглядеть так:

$$\begin{cases} t_i | \mathbf{x}_i, \mathbf{w} \sim \text{Pois}(\lambda(\mathbf{x}_i)), & i = 1 \dots n, \\ \lambda(\mathbf{x}) = e^{\mathbf{w}^T \mathbf{x}} \end{cases}$$

Оценка максимального правдоподобия для весов  $\mathbf{w}$  снова находится численными методами оптимизации.

А если задать априорные распределения на них, например

$$w_j \sim \mathcal{N}(M_j, S_j), \quad j = 1 \dots d,$$

то апостериорное  $p(\mathbf{w} | D_{\text{tr}}) \propto \text{Likelihood} \times \text{Prior}$ , и можно с помощью техники МСМС насемплировать точки  $\mathbf{w}_{[1]}, \dots, \mathbf{w}_{[N]}$  из этого апостериорного распр-ия.

Предсказания для новой точки  $\mathbf{x}_{\text{new}}$ , как всегда, оцениваются усреднением правдоподобий по выборке весов  $\mathbf{w}_{[1]}, \dots, \mathbf{w}_{[N]}$ :

$$\hat{p}(t | \mathbf{x}_{\text{new}}, D_{\text{tr}}) = \frac{1}{N} \sum_{r=1}^N \frac{e^{\mathbf{w}_{[r]}^T \mathbf{x}_{\text{new}}}}{t!} \exp \left( -e^{\mathbf{w}_{[r]}^T \mathbf{x}_{\text{new}}} \right), \quad t = 0, 1, 2, \dots$$

## Bayesian Feature Importance

Ещё одна задача, с которой может помочь справиться Байесовский подход, — это выбор значимых признаков.

Если в обобщённой линейной модели вместо линейной функции

$$a(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i = \sum_{j=1}^d w_j x_{ij} \quad \text{использовать} \quad a(\mathbf{x}_i) = \sum_{j=1}^d \delta_j w_j x_{ij},$$

где  $\delta_j \in \{0, 1\}$  — индикатор того, включать или нет  $j$ -ый признак в модель, то, задавая априорные распределения  $\delta_j \sim \text{Be}(0.5)$ , можно вычислить (точнее, оценить семплированием) апостериорную вероятность того, что  $\delta_j = 1$ .

Все  $d$  признаков после этого можно упорядочить по важности согласно этим оценкам, а самые неважные — отбросить.

That's all Folks!

