ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург
06 октября 2017 г.

Random facts:

- 6 октября 1917 г. *Literary Digest* впервые назвал музыку, «вызывающую желание трястись и корчиться», словом «джаз»
- а 6 октября 1927 г. на экраны вышел первый в истории звуковой фильм, *The Jazz Singer*
- 6 октября 2010 г. был запущен новый сервис по обмену фотографиями под названием *Instagram*

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Итак, мы рассмотрели логистический сигмоид:

$$p(\mathcal{C}_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1+e^{-a}} = \sigma(a),$$

$$\text{где } a = \ln \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)}, \qquad \sigma(a) = \frac{1}{1+e^{-a}}.$$

- Вывели из него LDA и QDA, обучили их методом максимального правдоподобия, а потом отвлеклись на naive Bayes.

ДВА КЛАССА

- Два класса, и апостериорное распределение – логистический сигмоид на линейной функции:

$$p(\mathcal{C}_1 \mid \phi) = y(\phi) = \sigma(\mathbf{w}^\top \phi), \quad p(\mathcal{C}_2 \mid \phi) = 1 - p(\mathcal{C}_1 \mid \phi).$$

- *Логистическая регрессия* – это когда мы напрямую оптимизируем $\mathbf{w}$.

ДВА КЛАССА

- Для датасета $\{\phi_n, t_n\}$, $t_n \in \{0,1\}$, $\phi_n = \phi(\mathbf{x}_n)$:

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}, \quad y_n = p(\mathcal{C}_1 \mid \phi_n).$$

- Ищем параметры максимального правдоподобия, минимизируя $-\ln p(\mathbf{t} \mid \mathbf{w})$:

$$E(\mathbf{w}) = -\ln p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^{N} \left[t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right].$$

ДВА КЛАССА

- Пользуясь тем, что $\sigma' = \sigma(1-\sigma)$, берём градиент (похоже на перцептрон):
$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n.$$
- Если теперь сделать градиентный спуск, получим как раз разделяющую поверхность.
- А ещё лучше – IRLS, который мы обсуждали в прошлый раз.

НЕСКОЛЬКО КЛАССОВ

- В случае нескольких классов

$$p(\mathcal{C}_k \mid \phi) = y_k(\phi) = \frac{e^{a_k}}{\sum_j e^{a_j}} \text{ для } a_k = \mathbf{w}_k^\top \phi.$$

- Опять выпишем максимальное правдоподобие; во-первых,

$$\frac{\partial y_k}{\partial a_j} = y_k \left([k = j] - y_j \right).$$

- Теперь запишем правдоподобие – для схемы кодирования 1-of-K будет целевой вектор $\mathbf{t}_n$ и правдоподобие

$$p(\mathbf{T} \mid \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathcal{C}_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

  для $y_{nk} = y_k(\phi_n)$; берём логарифм:

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T} \mid \mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}, \text{ и}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N} \left(y_{nj} - t_{nj}\right) \phi_n.$$

НЕСКОЛЬКО КЛАССОВ

- Оптимизировать опять можно по Ньютону-Рапсону; гессиан получится как

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk} \left([k = j] - y_{nj} \right) \phi_n \phi_n^\top.$$

- А что если у нас другая форма сигмоида?
- Мы по-прежнему в той же постановке: два класса, $p(t = 1 \mid a) = f(a)$, $a = \mathbf{w}^\top \phi$, f – функция активации.
- Давайте установим функцию активации с порогом θ: для каждого ϕ_n, вычисляем $a_n = \mathbf{w}^\top \phi_n$, и

$$\begin{cases} t_n = 1, & \text{если } a_n \geq \theta, \\ t_n = 0, & \text{если } a_n < \theta. \end{cases}$$

ПРОБИТ-РЕГРЕССИЯ

- Если θ берётся по распределению $p(\theta)$, это соответствует

$$f(a) = \int_{-\infty}^{a} p(\theta) \mathrm{d}\theta.$$

- Пусть, например, $p(\theta)$ – гауссиан с нулевым средним и единичной дисперсией. Тогда

$$f(a) = \Phi(a) = \int_{-\infty}^{a} \mathcal{N}\left(\theta \mid 0, 1\right) \mathrm{d}\theta.$$

ПРОБИТ-РЕГРЕССИЯ

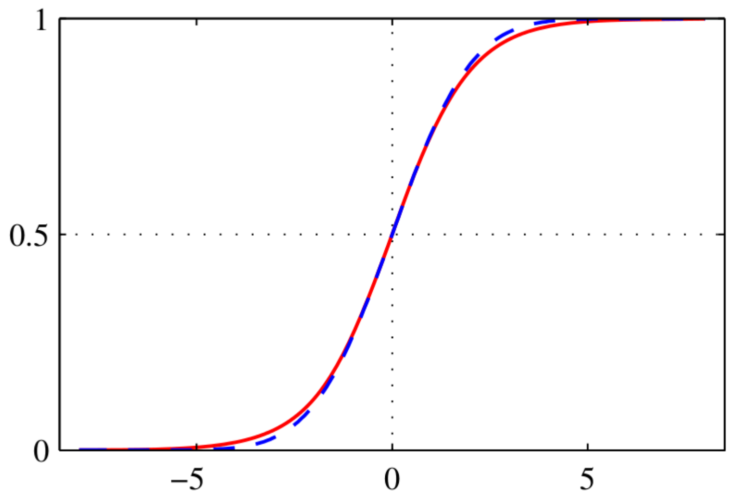
- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-\frac{\theta^2}{2}} \mathrm{d}\theta :$$

$$\Phi(a) = \frac{1}{2}\left[1 + \frac{1}{\sqrt{2}}\mathrm{erf}(a)\right].$$

- Пробит-регрессия – это модель с пробит-функцией активации.

σ и Φ



ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ И БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Небольшое лирическое отступление: как приблизить сложное распределение простым?
- Например, как приблизить гауссианом возле максимума? (естественная задача)
- Рассмотрим пока распределение от одной непрерывной переменной $p(z) = \frac{1}{Z} f(z)$.

ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ

- Первый шаг: найдём максимум z_0.
- Второй шаг: разложим в ряд Тейлора

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}A(z - z_0)^2, \text{ где } A = -\frac{d^2}{dz^2}\ln f(z)\mid_{z=z_0}.$$

- Третий шаг: приблизим

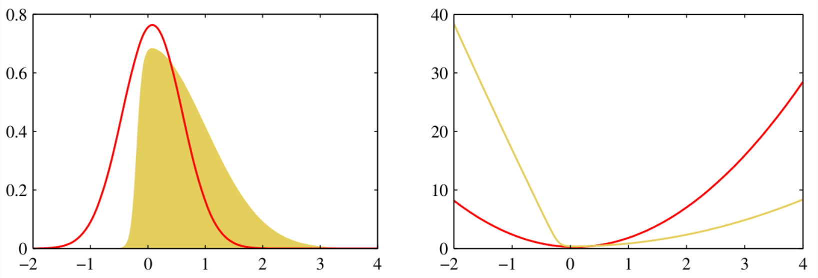
$$f(z) \approx f(z_0)e^{-\frac{A}{2}(z-z_0)^2},$$

и после нормализации это будет как раз гауссиан.

- Это можно обобщить на многомерное распределение $p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$:
$$f(\mathbf{z}) \approx f(\mathbf{z}_0) e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A}(\mathbf{z}-\mathbf{z}_0)},$$
$$\text{где } \mathbf{A} = -\nabla\nabla \ln f(\mathbf{z}) \mid_{z=z_0} .$$

Упражнение. Какая здесь будет нормировочная константа?

ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ



Сравнение моделей по Лапласу

- Вооружившись лапласовской аппроксимацией, давайте применим её сначала к выбору моделей.
- Напомним: чтобы сравнить модели из множества $\{\mathcal{M}_i\}_{i=1}^L$, по тестовому набору D оценим апостериорное распределение

$$p(\mathcal{M}_i \mid D) \propto p(\mathcal{M}_i)p(D \mid \mathcal{M}_i).$$

- Если модель определена параметрически, то $p(D \mid \mathcal{M}_i) = \int p(D \mid \theta, \mathcal{M}_i)p(\theta \mid \mathcal{M}_i)d\theta$.
- Это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению; знаменатель из теоремы Байеса:

$$p(\theta \mid \mathcal{M}_i, D) = \frac{p(D \mid \theta, \mathcal{M}_i)p(\theta \mid \mathcal{M}_i)}{p(D \mid \mathcal{M}_i)}.$$

СРАВНЕНИЕ МОДЕЛЕЙ ПО ЛАПЛАСУ

- Мы раньше приближали фактически кусочно-постоянной функцией.
- Теперь давайте гауссианом приблизим; возьмём интеграл:

$$Z = \int f(\mathbf{z})d\mathbf{z} \approx \int f(\mathbf{z}_0)e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A}(\mathbf{z}-\mathbf{z}_0)}d\mathbf{z} = f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}.$$

- А у нас $Z = p(D)$, $f(\theta) = p(D \mid \theta)p(\theta)$.

СРАВНЕНИЕ МОДЕЛЕЙ ПО ЛАПЛАСУ

- Получаем

$$\ln p(D) \approx \ln p(D \mid \theta_{\text{MAP}}) + \ln P(\theta_{\text{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|.$$

- $\ln P(\theta_{\text{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|$ – *фактор Оккама.*
- $\mathbf{A} = -\nabla\nabla \ln p(D \mid \theta_{\text{MAP}})p(\theta_{\text{MAP}}) = -\nabla\nabla \ln p(\theta_{\text{MAP}} \mid D)$.

СРАВНЕНИЕ МОДЕЛЕЙ ПО ЛАПЛАСУ

- Получаем

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) + \ln P(\theta_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|.$$

- Если гауссовское априорное распределение $p(\theta)$ достаточно широкое, и $\mathbf{A}$ полного ранга, то можно грубо приблизить (докажите это!)

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) - \frac{1}{2}M\ln N,$$

  где M – число параметров, N – число точек в D, а аддитивные константы мы опустили.
- Это *байесовский информационный критерий* (Bayesian information criterion, BIC), он же *критерий Шварца* (Schwarz criterion).

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Теперь давайте обработаем логистическую регрессию по-байесовски.
- Логистическую регрессию так просто не выпишешь, как линейную – точного ответа из произведения логистических сигмоидов не получается.
- Будем приближать по Лапласу.

- Априорное распределение выберем гауссовским:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mu_0, \Sigma_0).$$

- Тогда апостериорное будет

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}) &\propto p(\mathbf{w}) p(\mathbf{t} \mid \mathbf{w}), \text{ и} \\ \ln p(\mathbf{w} \mid \mathbf{t}) &= -\frac{1}{2}\left(\mathbf{w} - \mu_0\right)^\top \Sigma_0^{-1}\left(\mathbf{w} - \mu_0\right) \\ &\quad + \sum_{n=1}^{N}\left[t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\right] + \mathrm{const}, \\ \text{где } y_n &= \sigma(\mathbf{w}^\top \phi_n). \end{aligned}$$

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Чтобы приблизить, сначала находим максимум $\mathbf{w}_{\mathrm{MAP}}$, а потом матрица ковариаций – это матрица вторых производных

$$\Sigma_N = -\nabla\nabla \ln p(\mathbf{w} \mid \mathbf{t}) = \Sigma_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^\top.$$

- Наше приближение – это

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{w}_{\mathrm{MAP}}, \Sigma_N).$$

- Теперь можно описать байесовское предсказание:

$$p(\mathcal{C}_1 \mid \phi, \mathbf{t}) = \int p(\mathcal{C}_1 \mid \phi, \mathbf{w}) p(\mathbf{w} \mid \mathbf{t}) d\mathbf{w} \approx \int \sigma(\mathbf{w}^\top \phi) q(\mathbf{w}) \mathrm{d}\mathbf{w}.$$

- Заметим, что $\sigma(\mathbf{w}^\top \phi)$ зависит от $\mathbf{w}$ только через его проекцию на ϕ.
- Обозначим $a = \mathbf{w}^\top \phi$:

$$\sigma(\mathbf{w}^\top \phi) = \int \delta(a - \mathbf{w}^\top \phi) \sigma(a) \mathrm{d}a.$$

- $\sigma(\mathbf{w}^\top\phi) = \int \delta(a - \mathbf{w}^\top\phi)\sigma(a)\mathrm{d}a$, а значит,

$$\int \sigma(\mathbf{w}^\top\phi)q(\mathbf{w})d\mathbf{w} = \int \sigma(a)p(a)\mathrm{d}a,$$
$$\text{где } p(a) = \int \delta(a - \mathbf{w}^\top\phi)q(\mathbf{w})\mathrm{d}\mathbf{w}.$$

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ.

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ.
- Значит, $p(a)$ – тоже гауссиан; найдём его моменты:

$$\begin{aligned}\mu_a =& \mathrm{E}[a] = \int a p(a) \mathrm{d}a = \int q(\mathbf{w}) \mathbf{w}^\top \phi \mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^\top \phi, \\ \sigma_a^2 =& \int \left(a^2 - \mathrm{E}[a]\right)^2 p(a) \mathrm{d}a = \\ =& \int q(\mathbf{w}) \left[(\mathbf{w}^\top \phi)^2 - (\mu_N^\top \phi)^2\right]^2 \mathrm{d}\mathbf{w} = \phi^\top \Sigma_N \phi.\end{aligned}$$

- Итого получили, что

$$p(\mathcal{C}_1 \mid \mathbf{t}) = \int \sigma(a) p(a) \mathrm{d}a = \int \sigma(a) \mathcal{N}(a \mid \mu_a, \sigma_a^2) \mathrm{d}a.$$

- $p(\mathcal{C}_1 \mid \mathbf{t}) = \int \sigma(a)\mathcal{N}(a \mid \mu_a, \sigma_a^2)\mathrm{d}a$.
- Этот интеграл так просто не взять, потому что сигмоид сложный, но можно приблизить, если приблизить $\sigma(a)$ через пробит: $\sigma(a) \approx \Phi(\lambda a)$ для $\lambda = \sqrt{\pi/8}$.

Упражнение. Докажите, что для $\lambda = \sqrt{\pi/8}$ у σ и Φ одинаковый наклон в нуле.

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- А если мы перейдём к пробит-функции, то её свёртка с гауссианом будет просто другим пробитом:

$$\int \Phi(\lambda a)\mathcal{N}(a \mid \mu, \sigma^2)\mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\frac{1}{\lambda^2} + \sigma^2}}\right).$$

Упражнение. Докажите это.

- В итоге получается аппроксимация

$$\int \sigma(a)\mathcal{N}(a \mid \mu, \sigma^2)\mathrm{d}a \approx \sigma\left(\kappa(\sigma^2)\mu\right),$$

$$\text{где } \kappa(\sigma^2) = \frac{1}{\sqrt{1+\frac{\pi}{8}\sigma^2}}.$$

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- И теперь, собирая всё вместе, мы получили распределение предсказаний:

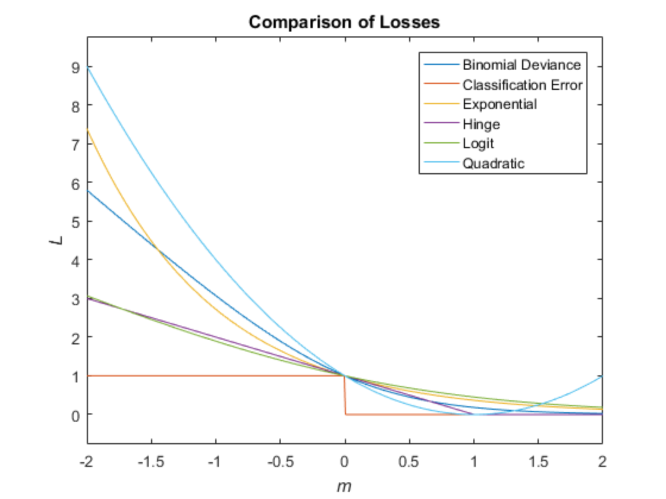
$$\begin{aligned} p(\mathcal{C}_1 \mid \phi, \mathbf{t}) =& \sigma\left(\kappa(\sigma_a^2)\mu_a\right), \text{ где} \\ \mu_a =& \mathbf{w}_{\mathrm{MAP}}^\top \phi, \\ \sigma_a^2 =& \phi^\top \Sigma_N \phi, \\ \kappa(\sigma^2) =& \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}. \end{aligned}$$

- Кстати, разделяющая поверхность $p(\mathcal{C}_1 \mid \phi, \mathbf{t}) = \frac{1}{2}$ задаётся уравнением $\mu_a = 0$, и тут нет никакой разницы с просто использованием $\mathbf{w}_{\mathrm{MAP}}$. Разница будет только для более сложных критериев.

ФУНКЦИИ ПОТЕРЬ В КЛАССИФИКАЦИИ

- И напоследок немножко другой взгляд: разные методы классификации отличаются друг от друга тем, какую функцию ошибки они оптимизируют.
- У классификации проблема с «правильной» функцией ошибки, то есть ошибкой собственно классификации:
    - она и не везде дифференцируема,
    - и производная её никому не нужна.
- Давайте посмотрим на разные функции потерь (loss functions); мы уже несколько видели, но ещё немало осталось.

ФУНКЦИИ ПОТЕРЬ В КЛАССИФИКАЦИИ



СПАСИБО!

Спасибо за внимание!