# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
3 марта 2017 г.

*Random facts*:

- 3 марта в Японии -- Хинамацури, Праздник девочек с тысячелетней историей. Девочки надевают нарядные кимоно с цветочным рисунком, ходят в гости, дарят друг другу подарки, едят хиси моти и хина арарэ, пьют сиродзакэ и рассматривают хина нингё, расставленные на хинакадзари
- 3 марта 1875 г. полностью провалилась премьеры оперы Жоржа Бизе <<Кармен>>

# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

# В ПРОШЛЫЙ РАЗ

- 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

- Возвращаемся к задаче классификации.
- 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.
- Заметим, правда, что если данные действительно разделимы, то может получиться жуткий оверфиттинг: $\|\mathbf{w}\| \to \infty$, и сигмоид превращается в функцию Хевисайда. Надо регуляризовать.

# IRLS

- В логистической регрессии не получается замкнутого решения из-за сигмоида.
- Но функция $E(\mathbf{w})$ всё равно выпуклая, и можно воспользоваться методом Ньютона-Рапсона – на каждом шаге использовать локальную квадратичную аппроксимацию к функции ошибки:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \mathbf{H}^{-1}\nabla E(\mathbf{w}),$$

где $\mathbf{H}$ (Hessian) – матрица вторых производных $E(\mathbf{w})$.

# IRLS

- Замечание: давайте применим Ньютона-Рапсона к обычной линейной регрессии с квадратической ошибкой:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\mathbf{w}^\top \phi_n - t_n\right) \phi_n = \Phi^\top \Phi \mathbf{w} - \Phi^\top \mathbf{t},$$

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi_n \phi_n^\top = \Phi^\top \Phi,$$

и шаг оптимизации будет

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left(\Phi^\top \Phi\right)^{-1} \left[\Phi^\top \Phi \mathbf{w}^{\text{old}} - \Phi^\top \mathbf{t}\right] = \\ = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t},$$

т.е. мы за один шаг придём к решению.

# IRLS

- Для логистической регрессии:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\,\phi_n = \Phi^\top (\mathbf{y} - \mathbf{t}),$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^\top = \Phi^\top R\Phi$$

для диагональной матрицы $R$ с $R_{nn} = y_n(1 - y_n)$.

# IRLS

- Формула шага оптимизации:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top (\mathbf{y} - \mathbf{t}) = \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top R \mathbf{z},$$

  где $\mathbf{z} = \Phi \mathbf{w}^{\text{old}} - R^{-1} (\mathbf{y} - \mathbf{t})$.
- Получилось как бы решение взвешенной задачи минимизации квадратического отклонения с матрицей весов $R$.
- Отсюда название: iterative reweighted least squares (IRLS).

# 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,

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

# НЕСКОЛЬКО КЛАССОВ

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

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

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

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

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

# несколько классов

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

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

# ПРОБИТ-РЕГРЕССИЯ

- 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 $\theta$: для каждого $\phi_n$, 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 $\theta$ 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

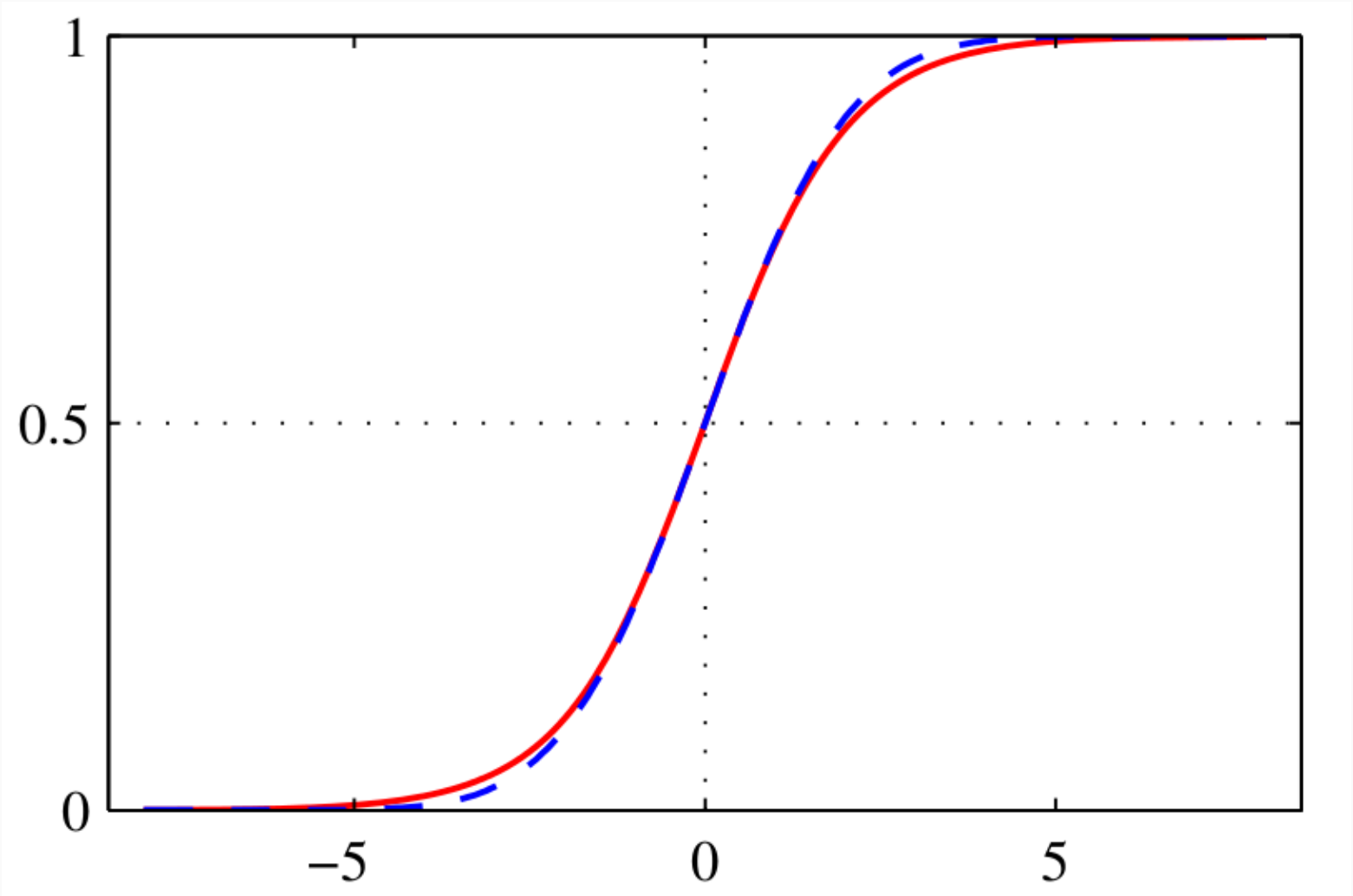
- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

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

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

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

# $\sigma$ и $\Phi$



# ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ И БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

# ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ

- 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.

# ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ

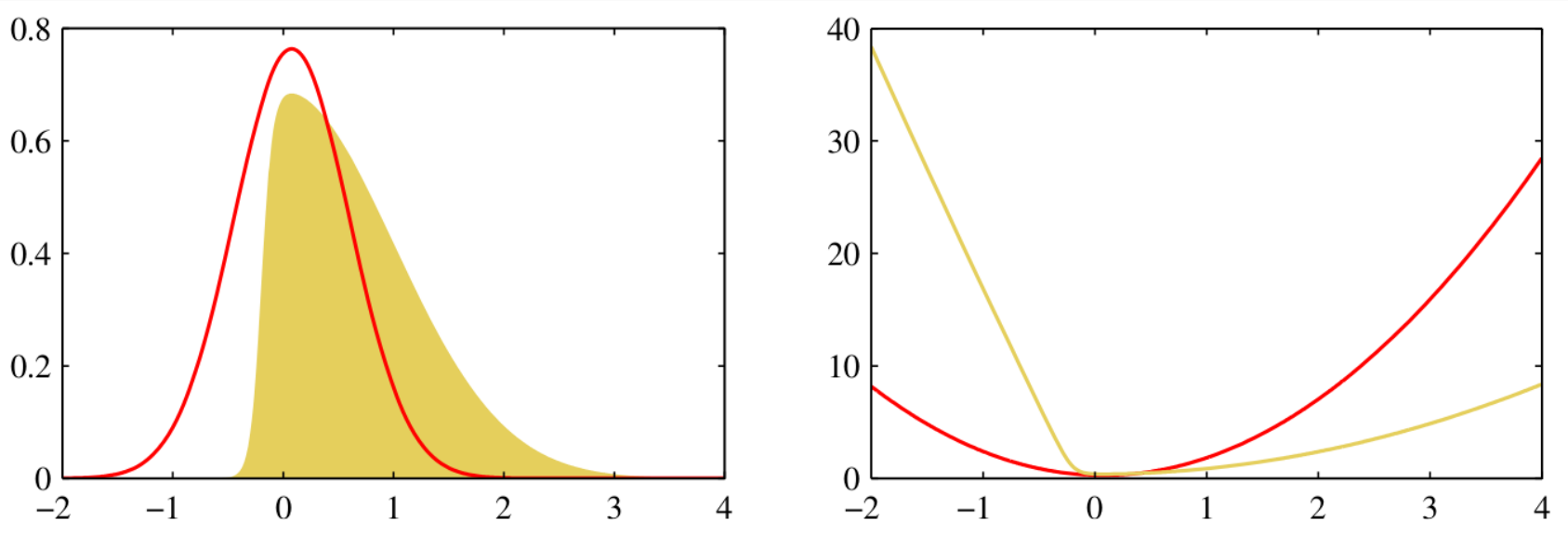
- Это можно обобщить на многомерное распределение $p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$:

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A} (\mathbf{z}-\mathbf{z}_0)},$$

$$\text{где } \mathbf{A} = -\nabla\nabla \ln f(\mathbf{z}) \,|_{z=z_0} \,.$$

**Упражнение.** Какая здесь будет нормировочная константа?

# ЛАПЛАСОВСКАЯ АППРОКСИМАЦИЯ



# СПАСИБО!

Спасибо за внимание!