Классификаторы III: логистическая регрессия

Сергей Николенко

Академический Университет, 2012

Outline

В прошлый раз

- В прошлый раз 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:

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

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

Два класса

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

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

- 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):

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\Phi}_n.$$

- 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:

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

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

и шаг оптимизации будет

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

$$= \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^\top \mathbf{t},$$

т.е. мы за один шаг придём к решению.

IRLS

- Для логистической регрессии:

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

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

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

IRLS

- Формула шага оптимизации:

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

$$= \left(\mathbf{\Phi}^\top \boldsymbol{R} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \boldsymbol{R} \mathbf{z},$$

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

Несколько классов

- В случае нескольких классов

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

- 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

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

для $y_{nk} = y_k(\boldsymbol{\Phi}_n)$; берём логарифм:

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

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

Несколько классов

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

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

Пробит-регрессия

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

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

Пробит-регрессия

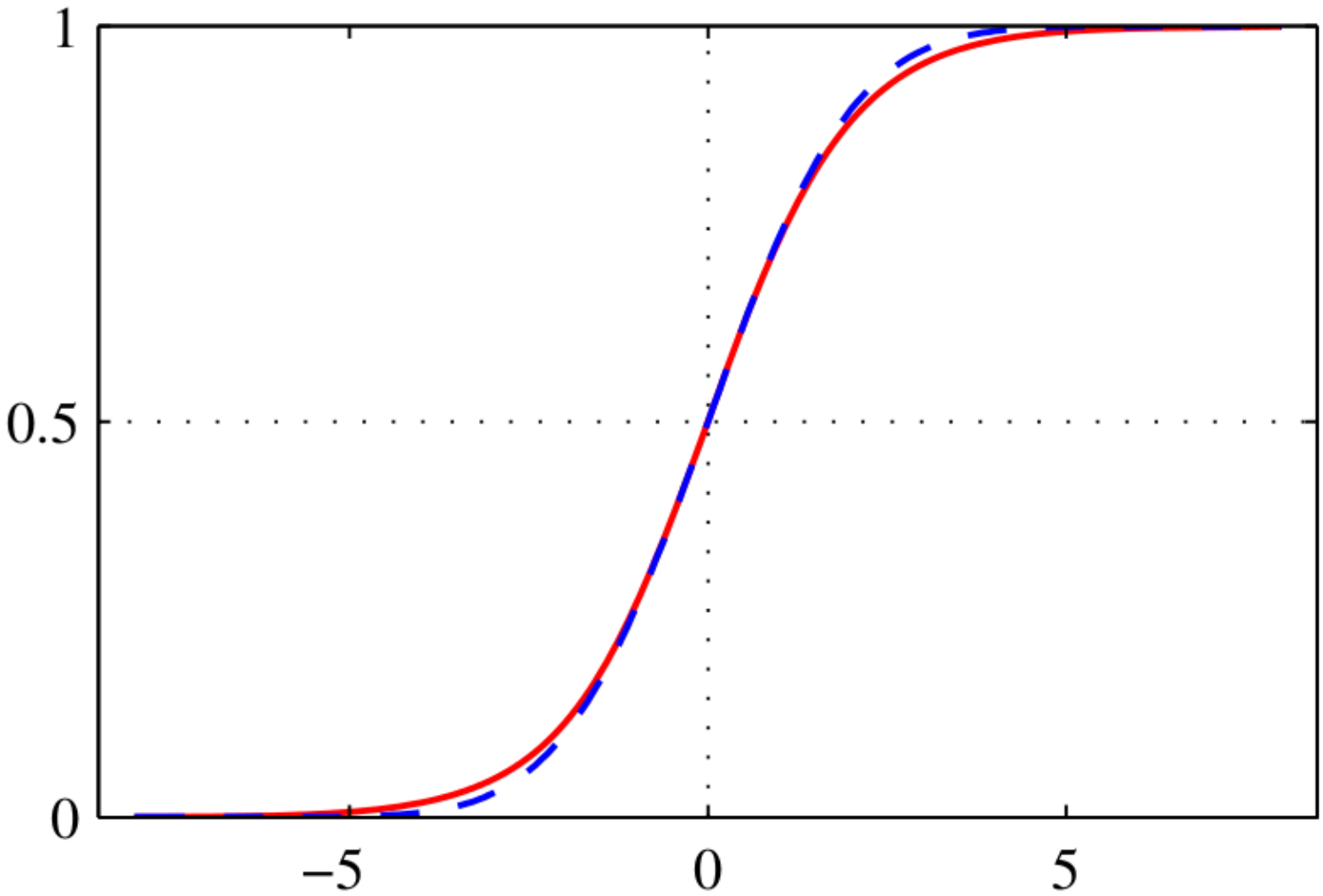
- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

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

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

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

σ и Φ



Outline

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

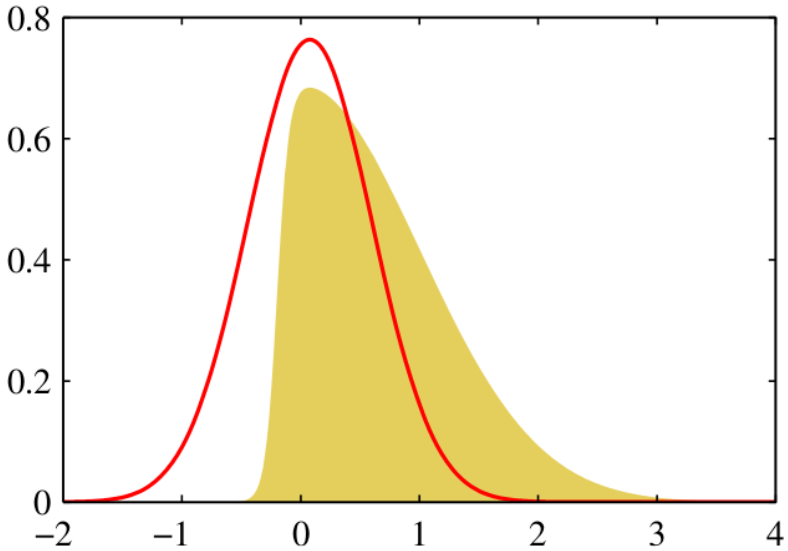
- Это можно обобщить на многомерное распределение $p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$:

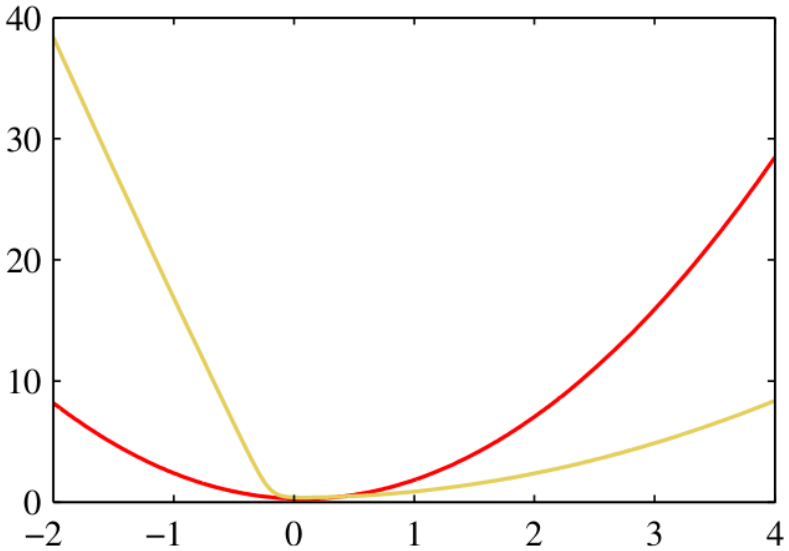
$$f(\mathbf{z}) \approx f(\mathbf{z}_0) e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A} (\mathbf{z}-\mathbf{z}_0)},$$

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

Упражнение. Какая здесь будет нормировочная константа?

Лапласовская аппроксимация





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

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

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

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

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.

Байесовская логистическая регрессия

- Априорное распределение выберем гауссовским:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Тогда апостериорное будет

$$p(\mathbf{w} \mid \mathbf{t}) \propto p(\mathbf{w}) p(\mathbf{t} \mid \mathbf{w}), \text{ и}$$

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbf{w} - \boldsymbol{\mu}_0) + \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)] + \text{const},$$

где $y_n = \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n)$.

Байесовская логистическая регрессия

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

$$\boldsymbol{\Sigma}_N = -\nabla\nabla \ln p(\mathbf{w} \mid \mathbf{t}) = \boldsymbol{\Sigma}_0^{-1} + \sum_{n=1}^{N} y_n(1 - y_n)\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^\top.$$

- Наше приближение – это

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{w}_{\text{MAP}}, \boldsymbol{\Sigma}_N).$$

Байесовская логистическая регрессия

- Теперь можно описать байесовское предсказание:

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

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

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

Байесовская логистическая регрессия

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

$$\int \sigma(\mathbf{w}^\top \boldsymbol{\phi}) q(\mathbf{w}) d\mathbf{w} = \int \sigma(a) p(a) \mathrm{d}a,$$

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

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

Байесовская логистическая регрессия

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

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

- Итого получили, что

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

Байесовская логистическая регрессия

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

Байесовская логистическая регрессия

- 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:

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

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

Thank you!

Спасибо за внимание!