ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ ПО-БАЙЕСОВСКИ

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
3 марта 2017 г.

Random facts:

- 3 марта 1944 г. в Париже дебютировал Ив Монтан, 3 марта 1955 г. Элвис Пресли впервые появился на телевидении, а 3 марта 1994 г. Курт Кобейн впал в кому, приняв валиум с шампанским
- 3 марта 1639 г. основанный в Бостоне колледж получил название в честь священника Джона Гарварда, который завещал колледжу 779 фунтов

СРАВНЕНИЕ МОДЕЛЕЙ ПО ЛАПЛАСУ

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

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

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

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.

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Априорное распределение выберем гауссовским:

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

- Тогда апостериорное будет

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

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

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

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- 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 $a = \mathbf{w}^\top \phi$:

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

БАЙЕСОВСКАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

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

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

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.

- 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:

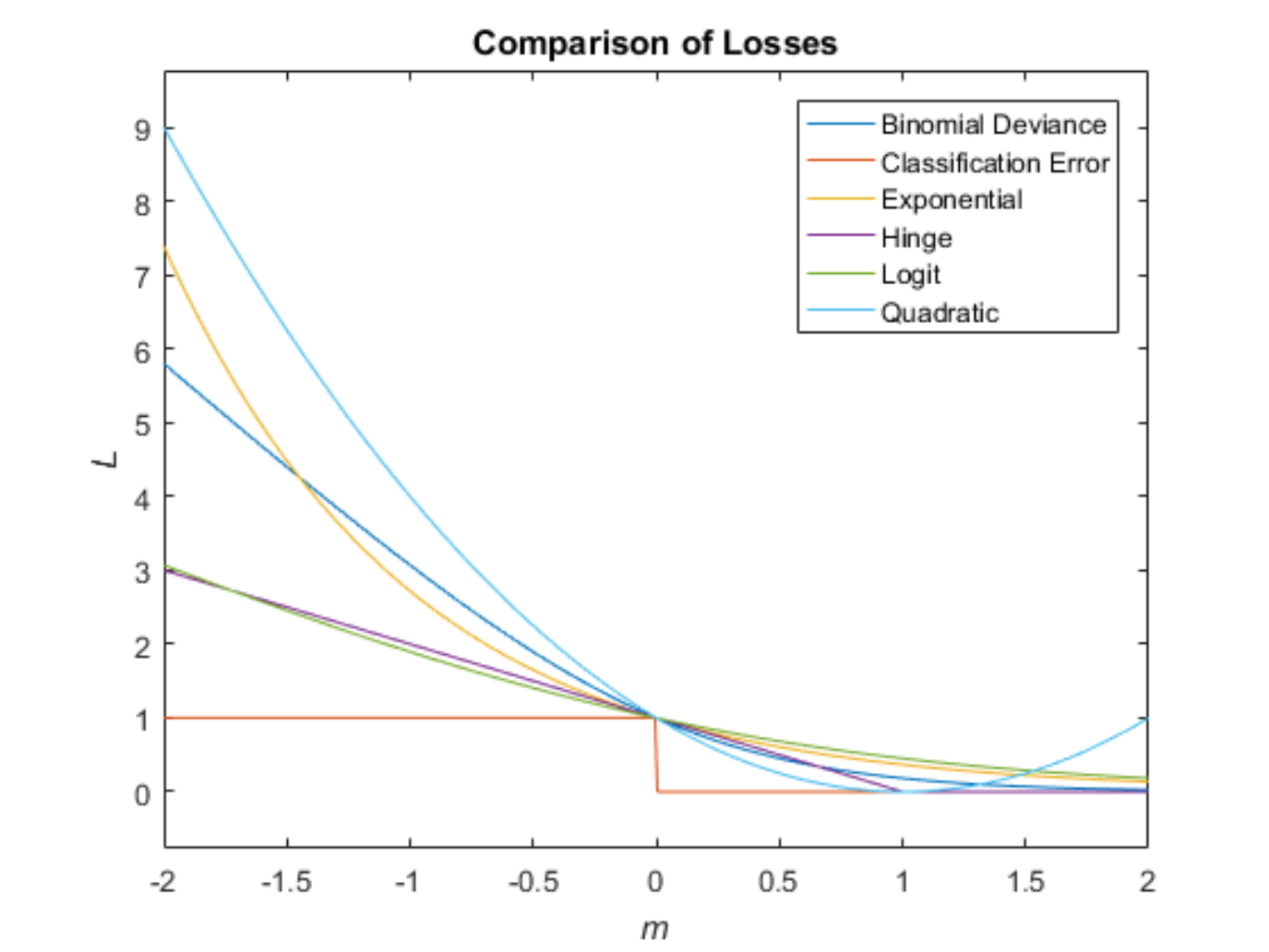
$$
\begin{aligned}
p(\mathcal{C}_1 \mid \phi, \mathbf{t}) =& \sigma\left(\kappa(\sigma_a^2)\mu_a\right), \text{ где} \\
\mu_a =& \mathbf{w}_{\mathrm{MAP}}^\top \phi, \\
\sigma_a^2 =& \phi^\top \Sigma_N \phi, \\
\kappa(\sigma^2) =& \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}.
\end{aligned}
$$

- 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!