ЭМПИРИЧЕСКИЙ БАЙЕС

Сергей Николенко

СПбГУ — Санкт-Петербург

08 ноября 2022 г.

Random facts:

- 8 ноября отмечается как День радиологии, ведь именно 8 ноября 1895 г. Конрад Рентген открыл X-лучи
- 8 ноября 1519 г. Эрнан Кортес впервые вошёл в Теночтитлан, и ацтеки торжественно приветствовали его; летом 1520 года Кортес был вынужден бежать, но через год вернулся, разрушил город и убил императора Монтесуму II
- 8 ноября 1923 г. в огромном пивном зале Bürgerbräukeller собрались около 3000 человек; после речи Густава фон Кара стоявший в дверях с кружкой пива Адольф Гитлер выбежал в середину зала, выстрелил в потолок и прокричал: «Национальная революция началась!»
- 8 ноября 1961 г. состоялся дебютный эфир телепередачи «Клуб Весёлых и Находчивых», и с 2001 года день первого телеэфира отмечается как «Международный день КВН»
- 8 ноября 1999 г. Джерри Кэссэдей убил Брюса Миллера на его свалке близ города Флинт, штат Мичиган; за убийство была осуждена жена Брюса, Шери Миллер, которая убедила своего онлайн-любовника Джерри убить своего мужа; это стало первым в мире убийством, совершённым через Интернет

Эмпирический байес

Эмпирический байес

- Откуда берутся гиперпараметры?
- Оказывается, их тоже можно оптимизировать!
- У линейной регрессии, например, два гиперпараметра: $\beta = \frac{1}{\sigma^2}$ и α (точность регуляризатора, пусть гребневого).
- Давайте просто попробуем оптимизировать $p(D \mid \alpha, \beta)$ (marginal likelihood).

Эмпирический байес

- Получается:

$$p(D \mid \alpha, \beta) = \int p(\mathbf{w}) p(D \mid \mathbf{w}) \mathrm{d}\mathbf{w},$$

$$\ln p(D \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \int e^{-\frac{\beta}{2}\|\mathbf{y} - X\mathbf{w}\|^2 - \frac{\alpha}{2}\mathbf{w}^\top \mathbf{w}} \mathrm{d}\mathbf{w}.$$

- Выделяем полный квадрат так же, как раньше:

$$A = \beta X^\top X + \alpha \mathbf{I},$$

$$\mu_N = \beta A^{-1} X^\top \mathbf{y}.$$

- Теперь

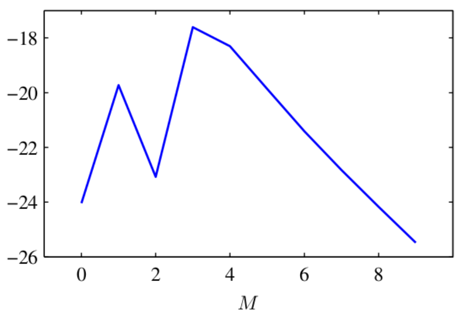
$$\int e^{-\frac{1}{2}(\mathbf{w}-\mu_N)^\top A(\mathbf{w}-\mu_N)}\mathrm{d}\mathbf{w} = (2\pi)^{\frac{d}{2}}\sqrt{\det A^{-1}}.$$

- Получается:

$$\ln p(D \mid \alpha, \beta) =$$
$$\frac{d}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\mathbf{y} - X\mu_N\|^2 - \frac{\alpha}{2}\mu_N^\top\mu_N - \frac{1}{2}\ln\det A - \frac{N}{2}\ln(2\pi).$$

- Это теперь надо максимизировать по α и β, а можно и разные d перебирать, если речь идёт о том, как выбрать оптимальное число признаков.

- Пример графика по числу параметров:



- А как оптимизировать?

- Обозначим через λ_i собственные числа матрицы $\beta\mathbf{X}^\top\mathbf{X}$:

$$\left(\beta\mathbf{X}^\top\mathbf{X}\right)\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

 для некоторых собственных векторов $\mathbf{u}_i$.
- Тогда $\mathbf{u}_i$ будут являться и собственными векторами матрицы $\mathbf{A}$, с собственными числами $\alpha + \lambda_i$:

$$\mathbf{A}\mathbf{u}_i = \left(\beta\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I}\right)\mathbf{u}_i = \lambda_i\mathbf{u}_i + \alpha\mathbf{u}_i.$$

- Теперь будем оптимизировать логарифм маргинального правдоподобия $\log p\left(\mathbf{y}|\mathbf{X},\alpha,\beta\right)$, взяв производную по α:

$$\frac{\partial \log \det \mathbf{A}}{\partial \alpha} = \frac{\partial \log \prod_i (\alpha+\lambda_i)}{\partial \alpha} = \sum_i \frac{\partial \log(\alpha+\lambda_i)}{\partial \alpha} = \sum_i \frac{1}{\alpha+\lambda_i},$$

  и вся производная будет равна

$$\frac{\partial \log p\left(\mathbf{y}|\mathbf{X},\alpha,\beta\right)}{\partial \alpha} = \frac{M}{2\alpha} - \frac{1}{2}\mu_N^\top \mu_n - \frac{1}{2}\sum_i \frac{1}{\alpha+\lambda_i}.$$

- Приравняем производную нулю:

$$\alpha\mu_N^\top\mu_N = M - \alpha\sum_i \frac{1}{\alpha+\lambda_i} = \sum_i \frac{\lambda_i}{\alpha+\lambda_i}.$$

- Получается итеративный процесс

$$\alpha^{(k+1)} = \frac{1}{\mu_N\left(\alpha^{(k)}\right)^\top \mu_N\left(\alpha^{(k)}\right)}\sum_i \frac{\lambda_i}{\alpha^{(k)}+\lambda_i}.$$

- Аналогично по β: можно заметить, что

$$\frac{\partial\lambda_i}{\partial\beta} = \frac{\lambda_i}{\beta}.$$

- Значит,

$$\frac{\partial \log \det \mathbf{A}}{\partial \beta} = \frac{\partial \log \prod_i (\alpha + \lambda_i)}{\partial \beta} = \sum_i \frac{\partial \log(\alpha + \lambda_i)}{\partial \beta} = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i}, \text{ и}$$

$$\frac{\partial \log p\left(\mathbf{y} | \mathbf{X}, \alpha, \beta\right)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mu_n^\top \mathbf{x}_n\right)^2 - \frac{1}{2\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

- Итого по β получаем

$$\beta^{(k+1)} = \frac{N - \sum_i \frac{\lambda_i}{\alpha^{(k)} + \lambda_i}}{\sum_{n=1}^{N} \left(t_n - \mu_n \left(\alpha^{(k)}, \beta^{(k)}\right)^\top \mathbf{x}_n\right)^2}.$$

Спасибо!

Спасибо за внимание!