БАЙЕСОВСКИЙ ВЫБОР МОДЕЛЕЙ

Сергей Николенко

СПбГУ — Санкт-Петербург

14 октября 2023 г.

Random facts:

- 14 октября в Грузии — Мцхетоба, христианский государственный праздник, проводящийся в кафедральном храме Светицховели города Мцхета; по легенде, храм основан на том месте, где была захоронена риза Иисуса Христа, привезённая в Грузию грузинскими евреями Элиозом и Лонгинозом, которые присутствовали при распятии
- 14 октября 1843 г. на премьере спектакля «Сон в летнюю ночь» впервые прозвучал «Свадебный марш» Мендельсона, а 14 октября 1860 г. был открыт Мариинский театр
- 14 октября 1892 г. Артур Конан Дойль опубликовал книгу «Приключения Шерлока Холмса», а 14 октября 1926 г. в Лондоне вышла книга Алана Милна «Винни-Пух»
- 14 октября 1943 г. произошло восстание в концлагере Собибор, единственное удачное в истории Третьего рейха
- 14 октября 1964 г. Никита Хрущёв был смещён с поста Первого секретаря ЦК КПСС (на его место пришёл Леонид Брежнев), а Мартин Лютер Кинг был удостоен Нобелевской премии мира
- 14 октября 2012 г. Феликс Баумгартнер прыгнул с парашютом с высоты 39 км и успешно приземлился в окрестностях, что характерно, города Розуэлл, Нью-Мексико

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:

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

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

Эмпирический байес

- Теперь

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

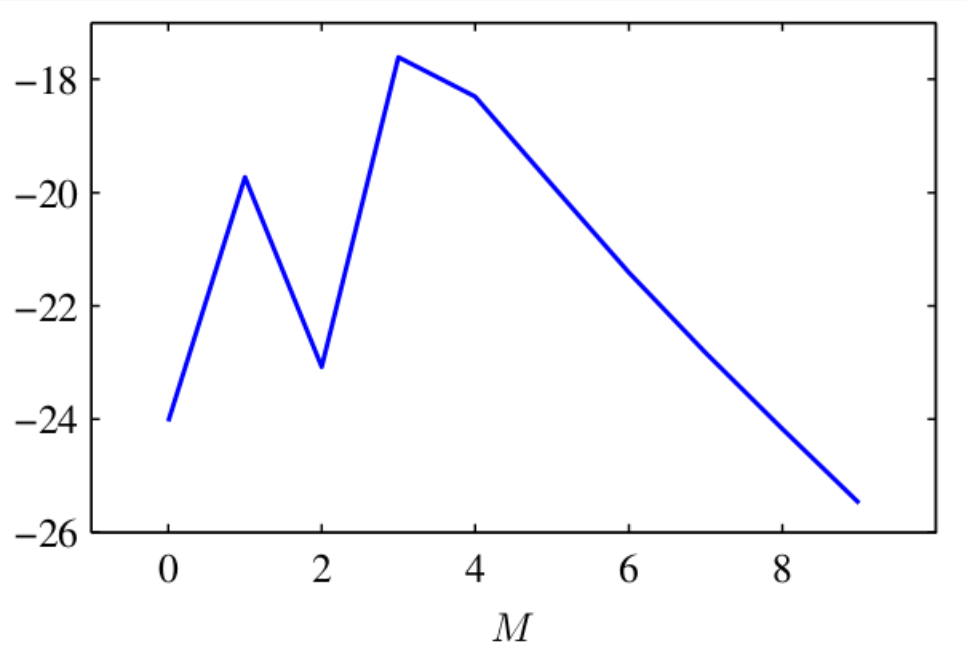
- Получается:

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

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

Эмпирический байес

- Пример графика по числу параметров:



- А как оптимизировать?

Эмпирический байес

- 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 $\alpha + \lambda_i$:

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

Эмпирический байес

- 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

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

Эмпирический байес

- 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

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

Эмпирический байес

- 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

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

Байесовское сравнение моделей

Байесовское сравнение моделей

- Мы говорили о том, что при увеличении числа параметров модели возникает оверфиттинг.
- Как этого избежать? Как сравнить модели с разным числом параметров?
- Теория байесовского вывода предлагает такой выход: давайте будем не точечные оценки параметров модели рассматривать, а тоже интегрировать по параметрам модели.

Байесовское сравнение моделей

- Пусть мы хотим сравнить модели из множества $\{\mathcal{M}_i\}_{i=1}^{L}$.
- Модель – это распределение вероятностей над данными D.
- По тестовому набору D можно оценить апостериорное распределение

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

Байесовское сравнение моделей

- Если знать апостериорное распределение, то можно сделать предсказание:

$$p(t \mid \mathbf{x}, D) = \sum_{i=1}^{L} p(t \mid \mathbf{x}, \mathcal{M}_i, D) p(\mathcal{M}_i \mid D).$$

- *Model selection* (выбор модели) – это когда мы приближаем предсказание, выбирая просто самую (апостериорно) вероятную модель.

Байесовское сравнение моделей

- Если модель определена параметрически, через $\mathbf{w}$, то

$$p(D \mid \mathcal{M}_i) = \int p(D \mid \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} \mid \mathcal{M}_i) d\mathbf{w}.$$

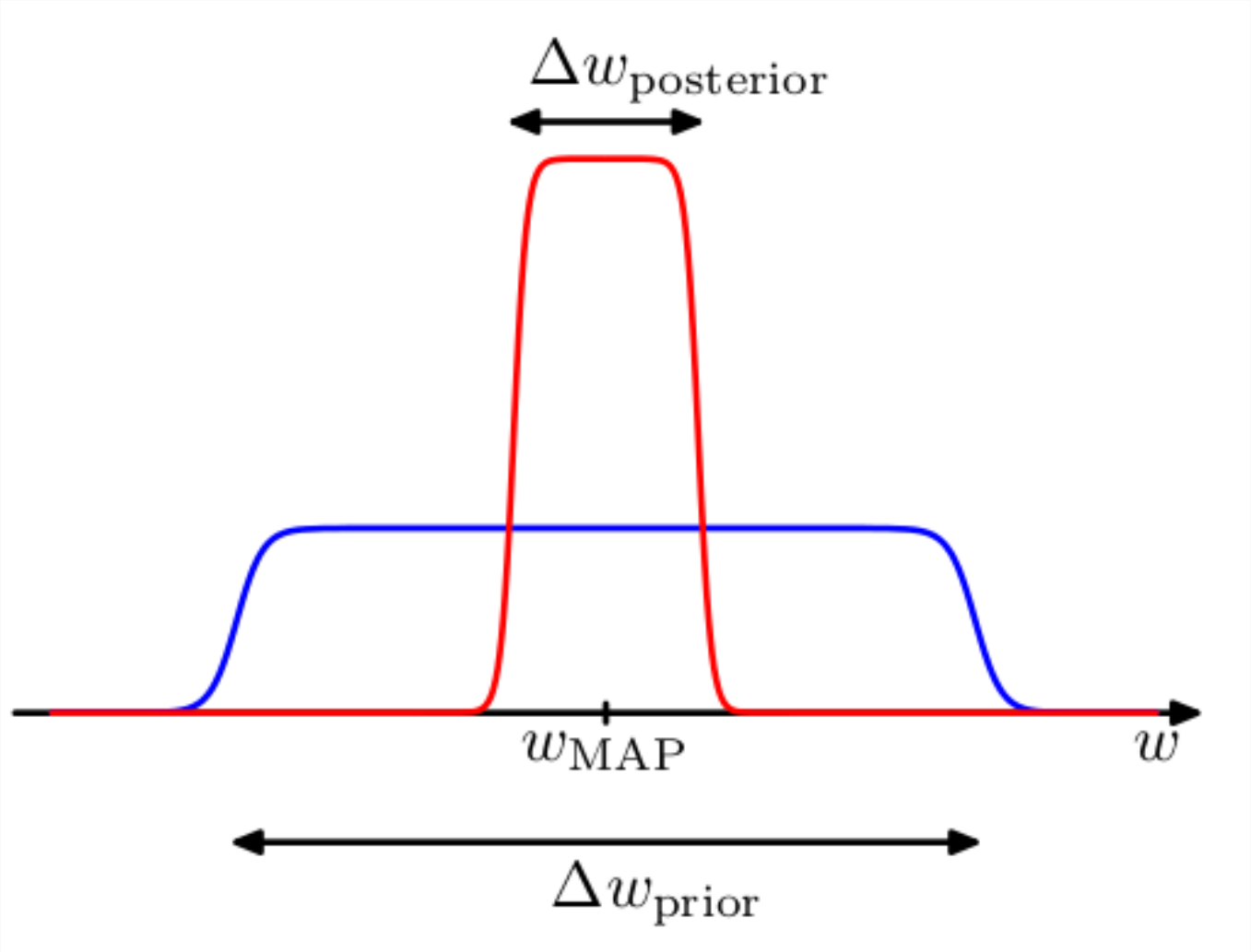
- Т.е. это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению, а потом накидывать данные.
- Это, кстати, в точности знаменатель из теоремы Байеса:

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

Байесовское сравнение моделей

- Предположим, что у модели один параметр w, а апостериорное распределение – это острый пик вокруг w_{MAP} шириной $\Delta w_{\mathrm{posterior}}$.
- Тогда можно приблизить $p(D) = \int p(D \mid w)p(w)dw$ как значение в максимуме, умноженное на ширину.
- Предположим ещё, что априорное распределение тоже плоское, $p(w) = \frac{1}{\Delta w_{\mathrm{prior}}}$.

Приближение $p(D)$



Приближение $p(D)$

- Тогда получится

$$p(D) = \int p(D \mid w)p(w)dw \approx p\left(D \mid w_{\text{MAP}}\right) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}},$$

$$\ln p(D) \approx \ln p\left(D \mid w_{\text{MAP}}\right) + \ln\left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right).$$

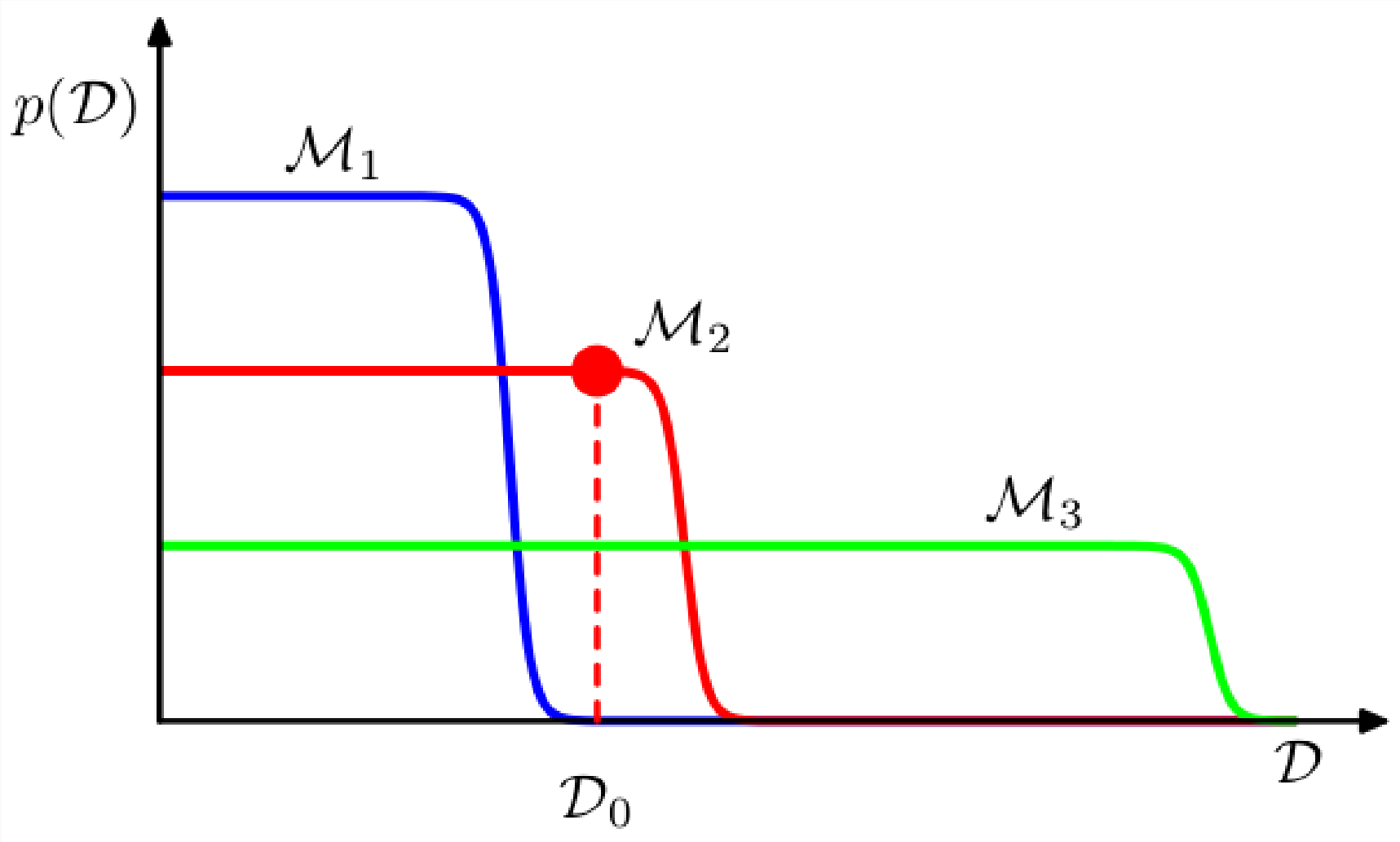
- Это значит, что мы добавляем штраф за «слишком узкое» апостериорное распределение – то есть в точности штраф за оверфиттинг!
- Для модели из M параметров, если предположить, что у них одинаковые $\Delta w_{\text{posterior}}$, получим

$$\ln p(D) \approx \ln p\left(D \mid w_{\text{MAP}}\right) + M \ln\left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right).$$

Другой взгляд

- Другими словами: давайте посмотрим, какие датасеты может генерировать та или иная модель.
- Простая модель (e.g., линейная) генерирует похожие датасеты, «мало» разных датасетов, у неё высокая $p(D \mid \mathcal{M})$.
- Сложная модель (e.g., многочлен девятой степени) генерирует «много» разных датасетов, у неё низкая $p(D \mid \mathcal{M})$.
- Но сложная может хорошо выразить датасеты, которые не может выразить простая; поэтому в сумме надо выбирать «среднюю».

Приближение $p(D)$



Правильный ответ лучше

- Sanity check: тут какие-то штрафы мы навводили; будет ли истинный правильный ответ $p(D \mid \mathcal{M}_{\text{true}})$ всегда оптимальным в этом смысле?
- Конечно, для конкретного датасета может так повезти, что не будет.
- Но если усреднить по всем датасетам, выбранным по $p(D \mid \mathcal{M}_{\text{true}})$...

Правильный ответ лучше

- ...то получится

$$\mathrm{E}\left[\ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})}\right] = \int p(D \mid \mathcal{M}_{\mathrm{true}}) \ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})} dD.$$

- Это называется *расстоянием Кульбака-Лейблера* (Kullback-Leibler divergence) между распределениями $p(D \mid \mathcal{M}_{\mathrm{true}})$ и $p(D \mid \mathcal{M})$.
- Но это только самый грубый результат о сравнении моделей...

Байесовский информационный критерий

BIC в руки

- Мы хотим сравнить несколько моделей $\mathcal{M}_1, \ldots, \mathcal{M}_K$ с наборами параметров $\theta_1, \ldots, \theta_K$ на наборе данных D, т.е. сравнить между собой $p(\mathcal{M}_k|D)$:

$$p(\mathcal{M}_k|D) \propto p(\mathcal{M}_k)\, p(D|\mathcal{M}_k)\,.$$

- Будем полагать $p(\mathcal{M}_k)$ равномерными. А $p(D|\mathcal{M}_k)$ — это как раз знаменатель теоремы Байеса:

$$p(\theta_k|D, \mathcal{M}_k) = \frac{p(\theta_k|\mathcal{M}_k)\, p(D|\theta_k, \mathcal{M}_k)}{p(D|\mathcal{M}_k)}.$$

BIC в руки

- Нам нужно оценить интеграл

$$p\left(D\right) = \int p\left(\theta\right) p\left(\theta|D\right) \mathrm{d}\theta = \int p\left(\theta\right) e^{\ell(\theta)} \mathrm{d}\theta,$$

где $\ell(\theta) = \log p\left(\theta|D\right)$.

- Применим лапласовскую аппроксимацию в окрестности точки максимума правдоподобия θ_{ML}:

$$\ell(\theta) \approx \ell(\theta_{\mathrm{ML}}) - \frac{N}{2} \left(\theta - \theta_{\mathrm{ML}}\right)^{\top} J(\theta_{\mathrm{ML}}) \left(\theta - \theta_{\mathrm{ML}}\right),$$

где

$$J\left(\theta_{\mathrm{ML}}\right) = -\frac{1}{N} \left. \frac{\partial^2 \log p\left(\theta|D\right)}{\partial\theta\partial\theta^{\top}} \right|_{\theta_{\mathrm{ML}}}.$$

BIC в руки

- Аналогично можно разложить априорное распределение окрестности θ_{ML}, но там не пропадёт член первого порядка, поэтому давайте им и ограничимся:

$$p(\theta) \approx p(\theta_{\mathrm{ML}}) + (\theta - \theta_{\mathrm{ML}})^{\top} \nabla_{\theta} p(\theta)|_{\theta_{\mathrm{ML}}}.$$

- Итого получается, что

$$p(D) \approx \int \left(p(\theta_{\mathrm{ML}}) + (\theta - \theta_{\mathrm{ML}})^{\top} \nabla_{\theta} p(\theta)|_{\theta_{\mathrm{ML}}} \right) \times \\ \times e^{\ell(\theta_{\mathrm{ML}}) - \frac{N}{2}(\theta - \theta_{\mathrm{ML}})^{\top} J(\theta_{\mathrm{ML}})(\theta - \theta_{\mathrm{ML}})} \mathrm{d}\theta.$$

BIC в руки

- Но теперь можно заметить, что

$$\int (\theta - \theta_{\mathrm{ML}})\, e^{-\frac{N}{2}(\theta-\theta_{\mathrm{ML}})^{\top} J(\theta_{\mathrm{ML}})(\theta-\theta_{\mathrm{ML}})} \mathrm{d}\theta = 0,$$

  потому что это величина, пропорциональная матожиданию $(\theta - \theta_{\mathrm{ML}})$ по гауссиану со средним $(\theta - \theta_{\mathrm{ML}})$ и матрицей ковариаций $J(\theta_{\mathrm{ML}})^{-1}$.

- А значит, наша аппроксимация превращается в

$$p\left(D\right) \approx e^{\ell(\theta_{\mathrm{ML}})} p\left(\theta_{\mathrm{ML}}\right) \int e^{-\frac{N}{2}(\theta-\theta_{\mathrm{ML}})^{\top} J(\theta_{\mathrm{ML}})(\theta-\theta_{\mathrm{ML}})} \mathrm{d}\theta.$$

BIC в руки

- Интеграл теперь можно взять — из него получится нормировочная константа для того же самого гауссиана:

$$p\left(D\right) \approx e^{\ell(\theta_{\mathrm{ML}})} p\left(\theta_{\mathrm{ML}}\right) \left(2\pi\right)^{\frac{d}{2}} N^{-\frac{d}{2}} \left(\det J(\theta_{\mathrm{ML}})\right)^{-\frac{1}{2}}, \quad \text{или}$$

$$\log p\left(D\right) \approx \ell(\theta_{\mathrm{ML}}) - \frac{d}{2}\log N + \log p\left(\theta_{\mathrm{ML}}\right) - \frac{1}{2}\log\det J(\theta_{\mathrm{ML}}) + \frac{d}{2}\log(2\pi).$$

- Выбросим всё, что не растёт с N, и умножим на -2; получится *байесовский информационный критерий* (Bayesian information criterion, BIC), он же *критерий Шварца* (Schwartz criterion):

$$\mathrm{BIC}\left(\mathcal{M}\right) = -2\log p\left(D|\theta_{\mathrm{ML}}, \mathcal{M}\right) + d\log N,$$

где d — это размерность вектора θ, или число *свободных параметров* в модели $\mathcal{M}$.

Спасибо!

Спасибо за внимание!