ИНФОРМАЦИОННЫЕ КРИТЕРИИ

Сергей Николенко

СПбГУ — Санкт-Петербург

01 ноября 2022 г.

Random facts:

- 1 ноября в Мексике и других странах — День мёртвых (El Día de Muertos); ещё индейцы майя и ацтеки приносили дары богине Миктлансиуатль и сооружали стены с изображением черепов — цомпантли; а в России 1 ноября — День судебного пристава
- 1 ноября 1478 г. Сикст IV направил Изабелле Кастильской и Фердинанду Арагонскому буллу, разрешившую учреждение Испанской инквизиции, а 1 ноября 1512 г. был впервые показан публике расписанный Микеланджело потолок Сикстинской капеллы
- 1 ноября 1800 г. Джон Адамс первым из президентов переехал в Исполнительный особняк (позднее переименованный в Белый дом)
- 1 ноября 1851 г. Огюст Мариэтт обнаружил в Саккаре под развалинами Серапеума вырубленные в скале грандиозные погребальные катакомбы священных быков — Аписов, вестников бога Птаха
- 1 ноября 1938 г. Seabiscuit победил War Admiral в «матче века»; по ходу гонки он отставал, но на финише опередил на четыре корпуса

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 $p\left(\mathcal{M}_k | D\right)$:

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

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

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

- 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

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

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

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

- Итого получается, что

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

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

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

- 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 $\mathcal{M}$.

Информационный критерий Акаике

Идея и вывод AIC

- Пусть данные $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ были получены из истинного распределения $p_{\text{data}}(\mathbf{x})$, а мы пытаемся приблизить их некоторой параметрической моделью $p\left(\mathbf{x}|\theta\right)$, $\theta \in \mathbb{R}^d$.
- Предположим, что мы обучили модель методом максимального правдоподобия, получив $p\left(\mathbf{x}|\theta_{\text{ML}}\right)$.
- Давайте попробуем оценить, насколько модель $p\left(\mathbf{x}|\theta_{\text{ML}}\right)$ отличается от неизвестного истинного распределения $p_{\text{data}}(\mathbf{x})$:

$$\begin{aligned}\mathrm{KL}\left(p_{\text{data}} \| p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right) &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log \frac{p_{\text{data}}(\mathbf{x})}{p\left(\mathbf{x}|\theta_{\text{ML}}\right)}\right] = \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p_{\text{data}}(\mathbf{x})\right] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right].\end{aligned}$$

Идея и вывод AIC

- Модель будет тем лучше, чем больше будет $\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\mathrm{ML}}\right)\right]$, и для оценки расхождения между $p\left(\mathbf{x}|\theta_{\mathrm{ML}}\right)$ и $p_{\mathrm{data}}(\mathbf{x})$ нужно получить оценку ожидаемого логарифма правдоподобия.
- Во всех критериях важен логарифм правдоподобия в точке его максимума, ведь это как раз выборочная оценка ожидания:

$$\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\mathrm{ML}}\right)\right] = \int p_{\mathrm{data}}(\mathbf{x}) \log p\left(\mathbf{x}|\theta_{\mathrm{ML}}\right) \mathrm{d}\mathbf{x} \approx$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \log p\left(\mathbf{x}_n|\theta_{\mathrm{ML}}\right).$$

- Но это смещённая оценка как минимум потому, что мы обучаем параметры максимального правдоподобия θ_{ML} на том же датасете $\mathbf{X}$, который используется в этой оценке.

Идея и вывод AIC

- Если истинная модель p_{data} тоже из семейства $p\left(\mathbf{x}|\theta\right)$ с некоторым истинным параметром θ_0, то

$$\theta_0 = \arg\max_\theta \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta\right)\right],$$

  но это ожидание берётся по всему распределению.
- θ_0 — это «истинная» гипотеза максимального правдоподобия; при некоторых условиях регулярности можно доказать, что:
    - $\theta_{\mathrm{ML}}(\mathbf{X}) \to \theta_0$ при $N \to \infty$;
    - для $\theta_{\mathrm{ML}}(\mathbf{X})$ верна асимптотическая нормальность, т.е. распределение величины $\sqrt{N}\left(\theta_{\mathrm{ML}} - \theta_0\right)$ сходится по вероятности к распределению $N\left(0, I(\theta_0)^{-1}\right)$, где $I(\theta)$ — это матрица информации Фишера

$$I(\theta) = \int p\left(\mathbf{x}|\theta\right) \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta} \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta^\top} \mathrm{d}\mathbf{x}.$$

Идея и вывод AIC

- Более того, те формулы предполагали, что $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, но аналогичные результаты можно получить и если $p_{\text{data}}(\mathbf{x})$ не принадлежит параметрическому семейству $p\left(\mathbf{x}|\theta\right)$.
- Пусть θ_0 — максимум ожидания логарифма правдоподобия по p_{data}, то есть решение системы

$$\int p_{\text{data}}(\mathbf{x}) \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta} \mathrm{d}\mathbf{x} = 0.$$

Идея и вывод AIC

- Тогда при тех же условиях можно доказать, что:
  - $\theta_{\mathrm{ML}}(\mathbf{X}) \to \theta_0$ при $N \to \infty$;
  - распределение величины $\sqrt{N}\left(\theta_{\mathrm{ML}} - \theta_0\right)$ сходится по вероятности к нормальному распределению

$$\sqrt{N}\left(\theta_{\mathrm{ML}} - \theta_0\right) \to_{N\to\infty} N\left(0, J^{-1}(\theta_0) I(\theta_0) J^{-1}(\theta_0)\right),$$

где $I(\theta)$ — это та же матрица информации Фишера, только по распределению p_{data}:

$$I(\theta) = \int p_{\mathrm{data}}(\mathbf{x}) \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta} \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta^\top} \mathrm{d}\mathbf{x};$$

а $J(\theta)$ — это ожидание матрицы вторых производных

$$J(\theta) = -\int p_{\mathrm{data}}(\mathbf{x}) \frac{\partial^2 \log p\left(\mathbf{x}|\theta\right)}{\partial \theta \partial \theta^\top} \mathrm{d}\mathbf{x}.$$

Идея и вывод AIC

- Иначе говоря, на позиции (i, j) у матрицы $I(\theta)$ стоит ожидание произведения $\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i}$ и $\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_j}$, а у матрицы $J(\theta)$ — ожидание второй производной $\frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial \theta_j}$.
- И если всё-таки $p_{\mathrm{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, то

$$\frac{\partial^2 \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\left(\frac{1}{p\left(x|\theta\right)} \frac{\partial p\left(\mathbf{x}|\theta\right)}{\partial \theta_j}\right) = \\ = \frac{1}{p\left(x|\theta\right)} \frac{\partial^2 p\left(\mathbf{x}|\theta\right)}{\partial \theta_i \partial \theta_j} - \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_i} \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_j},$$

а в ожидании по $p_{\mathrm{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$ при подстановке $\theta = \theta_0$

$$\mathbb{E}_{p(\mathbf{x}|\theta_0)}\left[\frac{1}{p\left(\mathbf{x}|\theta_0\right)} \frac{\partial^2 p\left(\mathbf{x}|\theta_0\right)}{\partial \theta_i \partial \theta_j}\right] = \int \frac{p\left(\mathbf{x}|\theta_0\right)}{p\left(\mathbf{x}|\theta_0\right)} \frac{\partial^2 p\left(\mathbf{x}|\theta_0\right)}{\partial \theta_i \partial \theta_j} \mathrm{d}\mathbf{x} = \\ = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int p\left(\mathbf{x}|\theta_0\right) = 0, \quad \text{то есть здесь} \quad I(\theta_0) = J(\theta_0).$$

Идея и вывод AIC

- Различные информационные критерии для сравнения моделей оценивают смещение выборочной оценки для величины $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right]$

$$b(p_{\text{data}}) = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right],$$

  где мы взяли ожидание по датасетам $\mathbf{X} \sim p_{\text{data}}$.
- Если мы сможем оценить смещение $b(p_{\text{data}})$, то информационный критерий можно будет построить, умножив на -2 аналогично BIC:

$$\begin{aligned} &\mathrm{IC}\left(\mathbf{X},\theta\right) = \\ &= -2\left(\text{логарифм правдоподобия } \mathbf{X} \text{ в } \theta_{\text{ML}} - \text{оценка смещения}\right) = \\ &= -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right) + 2\left(\text{оценка } b(p_{\text{data}})\right). \end{aligned}$$

Идея и вывод AIC

- Давайте попробуем оценить $b(p_{\text{data}})$:

$$\begin{aligned} b(p_{\text{data}}) &= \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right] = \\ &= \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - \log p\left(\mathbf{X}|\theta_0\right)\right] + \\ &+ \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_0\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right]\right] + \\ &+ \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right] = \\ =& B_1 + B_2 + B_3. \end{aligned}$$

- Будем оценивать слагаемые по отдельности.

- Проще всего оценить B_2, потому что в нём нет θ_{ML}:

$$B_2 = \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[\log p\left(\mathbf{X}|\theta_0\right) - N\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right]\right] =$$
$$= \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_0\right)\right] - N\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] = 0.$$

- Это не значит, что B_2 всегда равно нулю; для конкретного датасета $\mathbf{X}$ значение B_2 будет ненулевым, но в ожидании получится ноль.

Идея и вывод AIC

- Чтобы оценить B_3, рассмотрим функцию $\eta(\theta_{\mathrm{ML}}) = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\mathrm{ML}}\right)\right]$ и разложим её по формуле Тейлора в окрестности точки θ_0 (её максимума):

$$\eta(\theta_{\mathrm{ML}}) \approx \eta(\theta_0) - \frac{1}{2}\left(\theta_{\mathrm{ML}} - \theta_0\right)^\top J(\theta_0)\left(\theta_{\mathrm{ML}} - \theta_0\right),$$

где

$$J(\theta_0) = -\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\left.\frac{\partial^2 \log p\left(\mathbf{z}|\theta\right)}{\partial\theta\partial\theta^\top}\right|_{\theta_0}\right] = \\ = -\int p_{\mathrm{data}}(\mathbf{z})\left.\frac{\partial^2 \log p\left(\mathbf{z}|\theta\right)}{\partial\theta\partial\theta^\top}\right|_{\theta_0} \mathrm{d}\mathbf{z}.$$

Идея и вывод AIC

- А B_3 — это ожидание $\eta(\theta_0) - \eta(\theta_{\mathrm{ML}})$ по распределению $p_{\mathrm{data}}(\mathbf{X})$:

$$
\begin{aligned}
B_3 = \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[N\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] - N\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\mathrm{ML}}\right)\right]\right] = \\
= \frac{N}{2}\mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[\left(\theta_{\mathrm{ML}} - \theta_0\right)^\top J(\theta_0)\left(\theta_{\mathrm{ML}} - \theta_0\right)\right] = \\
= \frac{N}{2}\mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[\mathrm{Tr}\left(J(\theta_0)\left(\theta_{\mathrm{ML}} - \theta_0\right)\left(\theta_{\mathrm{ML}} - \theta_0\right)^\top\right)\right] = \\
= \frac{N}{2}\mathrm{Tr}\left(J(\theta_0)\mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}}\left[\left(\theta_{\mathrm{ML}} - \theta_0\right)\left(\theta_{\mathrm{ML}} - \theta_0\right)^\top\right]\right).
\end{aligned}
$$

- Теперь можно вместо ожидания матрицы ковариаций по датасету $\mathbf{X}$ подставить асимптотический результат:

$$
B_3 = \frac{N}{2}\mathrm{Tr}\left(J(\theta_0)\frac{1}{N}J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}\right) = \frac{1}{2}\mathrm{Tr}\left(I(\theta_0)J(\theta_0)^{-1}\right).
$$

Идея и вывод AIC

- Для оценки B_1 нужно провернуть аналогичный трюк с $\ell(\theta) = \log p\left(X|\theta\right)$, разложив его вокруг своего максимума θ_{ML}:

$$\ell(\theta) = \ell(\theta_{\mathrm{ML}}) + \frac{1}{2}\left(\theta - \theta_{\mathrm{ML}}\right)^{\top} \left.\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_{\mathrm{ML}}} \left(\theta - \theta_{\mathrm{ML}}\right).$$

- Мы знаем, что $\theta_{\mathrm{ML}} \to \theta_0$ при $N \to \infty$; а по закону больших чисел можно получить, что

$$-\frac{1}{N} \left.\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\top}}\right|_{\theta} = -\frac{1}{N}\sum_{n=1}^{N} \left.\frac{\partial^2 \log p\left(\mathbf{x}_n|\theta\right)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_0} \to J(\theta_0).$$

Идея и вывод AIC

- Следовательно, в нашу оценку можно подставить

$$\ell(\theta_{\mathrm{ML}}) - \ell(\theta_0) \approx -\frac{N}{2} \left(\theta - \theta_{\mathrm{ML}}\right)^\top J(\theta_0) \left(\theta - \theta_{\mathrm{ML}}\right).$$

- А затем и оценить B_1 так же, как оценивали B_3:

$$\begin{aligned} B_1 = \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\log p\left(\mathbf{X}|\theta_{\mathrm{ML}}\right) - \log p\left(\mathbf{X}|\theta_0\right)\right] = \\ = \frac{N}{2}\mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\left(\theta - \theta_{\mathrm{ML}}\right)^\top J(\theta_0)\left(\theta - \theta_{\mathrm{ML}}\right)\right] = \\ = \frac{N}{2}\mathrm{Tr}\left(J(\theta_0)\mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\left(\theta - \theta_{\mathrm{ML}}\right)^\top \left(\theta - \theta_{\mathrm{ML}}\right)\right]\right), \end{aligned}$$

то есть

$$B_3 = \frac{1}{2}\mathrm{Tr}\left(I(\theta_0)J(\theta_0)^{-1}\right). \qquad (1)$$

- Осталось только объединить три оценки:

$$b(p_{\text{data}}) = B_1 + B_2 + B_3 = \text{Tr}\left(I(\theta_0)J(\theta_0)^{-1}\right).$$

- $I(\theta_0)$ и $J(\theta_0)$ нам неизвестны, т.к. зависят от p_{data}; если взять оценки $\hat{I}$ и $\hat{J}$, это приведёт нас к *информационному критерию Такеучи* (Takeuchi information criterion, TIC):

$$\text{TIC} = -2\sum_{n=1}^{N} \log p\left(\mathbf{x}_n | \theta_{\text{ML}}\right) + 2\text{Tr}\left(\hat{I}\hat{J}^{-1}\right).$$

- В качестве $\hat{I}$ и $\hat{J}$ можно подставить просто усреднённые значения по датасету в точке максимума правдоподобия:

$$\hat{I}_{i,j} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial \log p\left(\mathbf{x}_n|\theta\right)}{\partial \theta_i} \frac{\partial \log p\left(\mathbf{x}_n|\theta\right)}{\partial \theta_j}\bigg|_{\theta_{\text{ML}}}, \quad \hat{J}_{i,j} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial^2 \log p\left(\mathbf{x}_n|\theta\right)}{\partial \theta_i \partial \theta_j}\bigg|_{\theta_{\text{ML}}}.$$

Идея и вывод AIC

- А если можно всё-таки предполагать, что истинное распределение данных p_{data} лежит в параметрическом семействе $p\left(\mathbf{x}|\theta\right)$, то есть $p_{\mathrm{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, то, как мы обсуждали выше, $I(\theta_0) = J(\theta_0)$, и информационный критерий Такеучи превращается в *информационный критерий Акаике* (Akaike information criterion, AIC):

$$\begin{aligned}\mathrm{AIC} = -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\mathrm{ML}}\right) + 2\mathrm{Tr}\left(\mathbf{I}_d\right) = \\ = -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\mathrm{ML}}\right) + 2d.\end{aligned}$$

- Очень простая формула!

Идея и вывод AIC

- Пример: вернёмся к полиномиальной регрессии с логарифмом правдоподобия

$$\ell(\mathbf{w}) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top\mathbf{x}_n\right)^2.$$

- Давайте в этом разделе для разнообразия будем дисперсию тоже обучать:

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}, \quad {\sigma_{\mathrm{ML}}}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}_{\mathrm{ML}}^\top\mathbf{x}_n\right)^2.$$

- Тогда при подстановке гипотезы максимального правдоподобия получится

$$\ell(\mathbf{w}_{\mathrm{ML}}) = -\frac{N}{2}\log\left(2\pi{\sigma_{\mathrm{ML}}}^2\right) - \frac{N}{2}.$$

Идея и вывод AIC

- AIC и BIC в таком примере будут, скорее всего, выбирать примерно одну и ту же модель, хотя разница между ними всё-таки есть:

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).

- 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

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

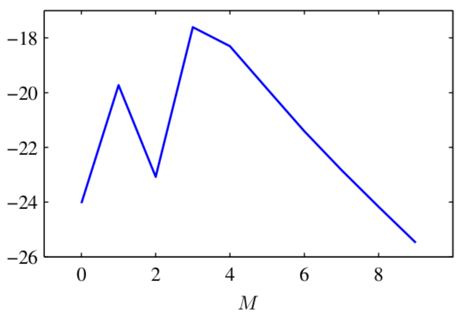
- Получается:

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

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

Эмпирический байес

- 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

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

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

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

- 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!