ИНФОРМАЦИОННЫЙ КРИТЕРИЙ АКАИКЕ

Сергей Николенко

СПбГУ — Санкт-Петербург

07 ноября 2024 г.

Random facts:

- 7 ноября 1631 г. Пьер Гассенди впервые наблюдал прохождение Меркурия по диску Солнца, предсказанное Кеплером
- 7 ноября 1902 г. в Туле, по инициативе местного врача Фёдора Архангельского, открылся первый в России вытрезвитель под названием «Приют для опьяневших»
- 7 ноября 1917 г. в России произошла Октябрьская революция, вооружённое восстание в Петрограде; именно 7 ноября Владимир Ленин сказал: «Товарищи! Рабочая и крестьянская революция, о необходимости которой всё время говорили большевики, свершилась!»; а 7 ноября 1987 г. в Тунисе произошла Жасминовая революция — бескровная смена власти, приведшая к отмене пожизненного президентства
- 7 ноября 1944 г. в Токио был казнён Рихард Зорге, а в США Франклин Рузвельт избран президентом в четвёртый раз подряд
- 7 ноября 1967 г. Карл Стоукс стал первым темнокожим мэром крупного города США (Кливленда), 7 ноября 1989 г. Дуглас Уайлдер и Дэвид Динкинс стали первыми темнокожими губернатором штата (Виргинии!) и мэром Нью-Йорка; а 7 ноября 1990 г. Мэри Робинсон стала первой женщиной-президентом Ирландии

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 $p_{\text{data}}(\mathbf{x})$:

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

Идея и вывод AIC

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

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

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

- 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 $p(\mathbf{x}|\theta)$ с некоторым истинным параметром θ_0, то

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

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

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

Идея и вывод AIC

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

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

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 $\sqrt{N}\,(\theta_{\mathrm{ML}} - \theta_0)$ сходится по вероятности к нормальному распределению

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

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

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

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

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

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 $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, 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 $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$ 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 $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right]$

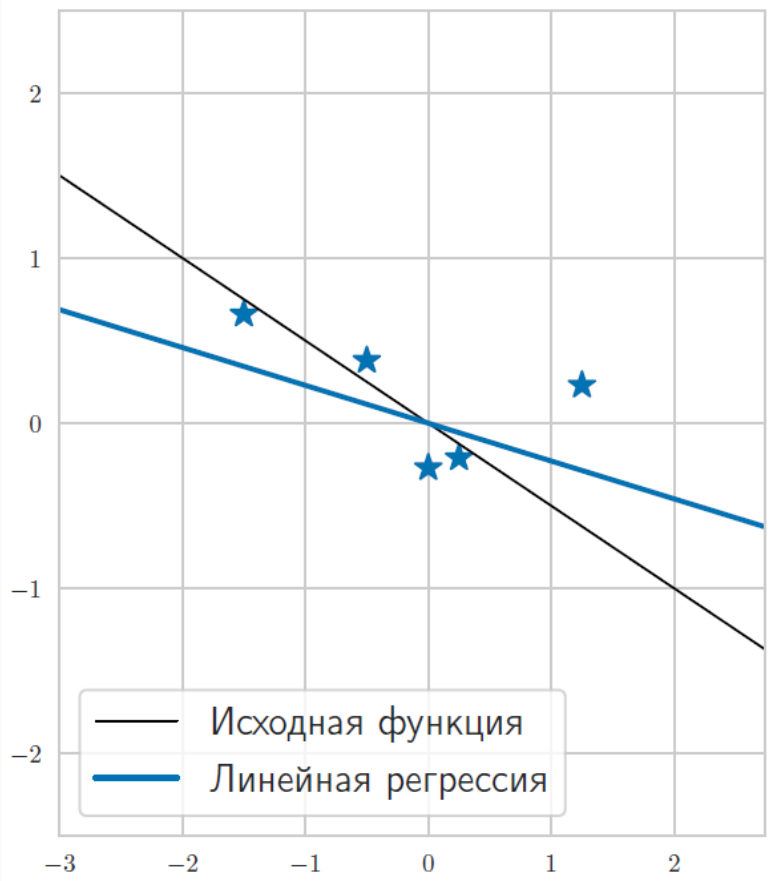
$$b(p_{\text{data}}) = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right],$$

  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:

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

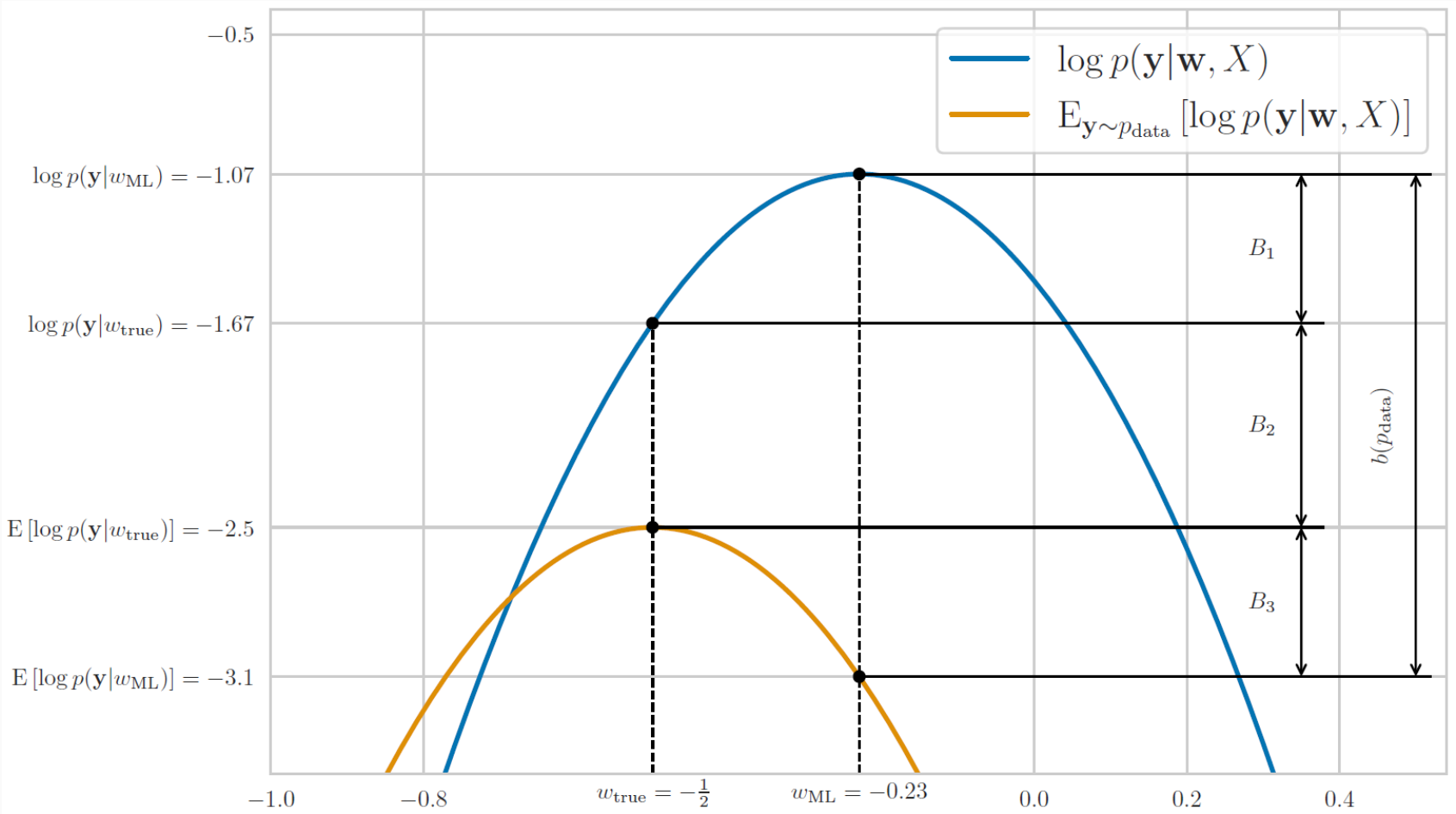
Идея и вывод AIC

- Пример — давайте рассмотрим линейную регрессию с одним параметром w:



Идея и вывод AIC

- И нарисуем графики $\log p\left(\mathbf{y}|w\right)$ и $\mathbb{E}_{\mathbf{y}\sim p_{\text{data}}}\left[\log p\left(\mathbf{y}|w\right)\right]$:



Идея и вывод AIC

- Давайте попробуем оценить $b(p_{\text{data}})$, разложив как на картинке на три слагаемых (но теперь в ожидании):

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

- Будем оценивать слагаемые по отдельности.

Идея и вывод AIC

- 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 $\eta(\theta_0) - \eta(\theta_{\text{ML}})$ по распределению $p_{\text{data}}(\mathbf{X})$:

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

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

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

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

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

Идея и вывод AIC

- Осталось только объединить три оценки:

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

- $I(\theta_0)$ 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):

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

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

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

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 $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, то, как мы обсуждали выше, $I(\theta_0) = J(\theta_0)$, и информационный критерий Такеучи превращается в *информационный критерий Акаике* (Akaike information criterion, AIC):

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

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

- 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 теперь 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

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

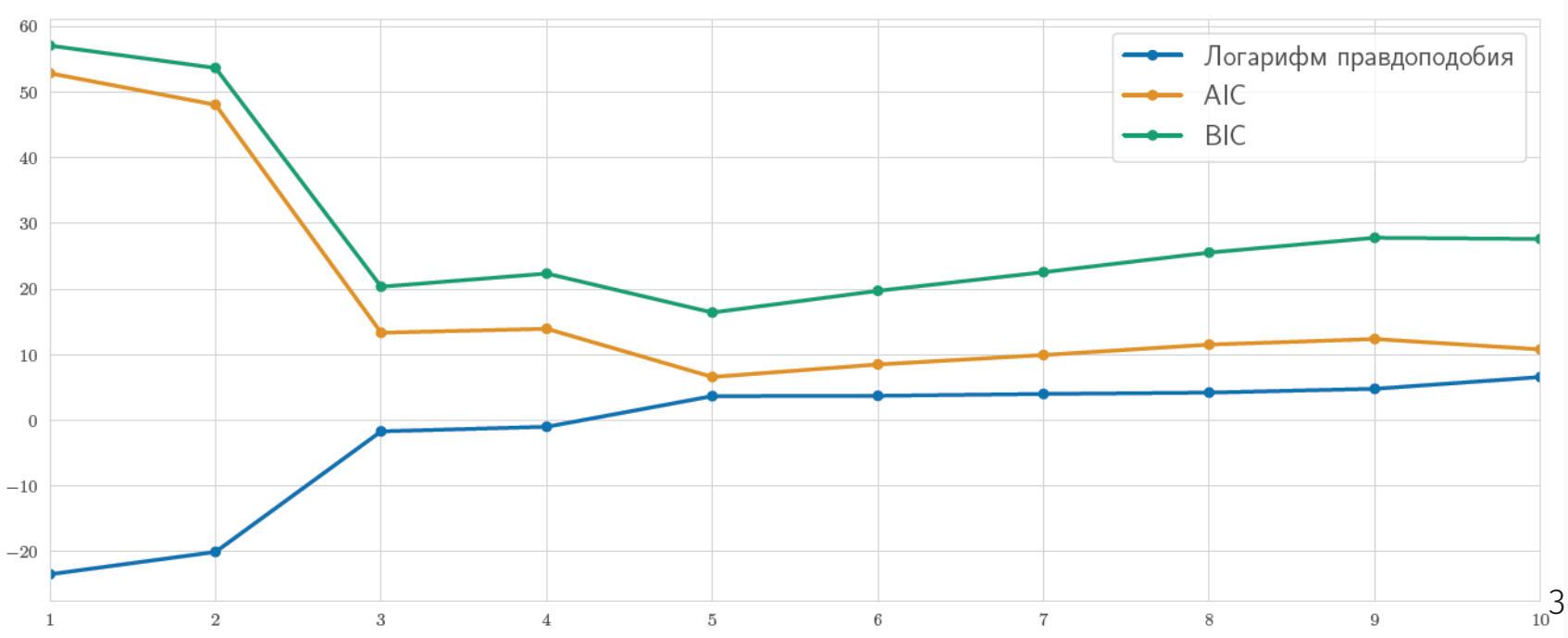
  а параметров в модели будет $d + 2$, где d — степень многочлена.

Идея и вывод AIC

- Тогда AIC и BIC получаются такими:

$$\mathrm{AIC}(d) = N\left(\log(2\pi) + \log {\sigma_{\mathrm{ML}}}^2 + 1\right) + 2(d+2),$$
$$\mathrm{BIC}(d) = N\left(\log(2\pi) + \log {\sigma_{\mathrm{ML}}}^2 + 1\right) + (d+2)\log N.$$

- AIC и BIC в этом примере будут, скорее всего, выбирать примерно одно и то же, хотя разница всё-таки есть:



Спасибо за внимание!