

ЭМПИРИЧЕСКИЙ БАЙЕС И AIC

Сергей Николенко

СПбГУ — Санкт-Петербург

6 ноября 2025 г.

Random facts:

- 6 ноября 447 г. мощное землетрясение разрушило большие участки стен Константинополя, в том числе 57 башен
- 6 ноября 1736 г. Михаил Ломоносов был зачислен в Марбургский университет в Германии
- 6 ноября 1928 г. американский полковник Джейкоб Шик запатентовал электробритву
- 6 ноября 1947 г. на NBC впервые вышла *Meet the Press*, самая долгоживущая телепрограмма в истории (уже 77 сезонов и почти 5000 выпусков)
- 6 ноября 1957 г. на Марсовом поле в Ленинграде зажгли первый в стране Вечный огонь, а 6 ноября 1968 г. на Кутузовском проспекте у Поклонной горы была открыта Триумфальная арка
- 6 ноября 1975 г. группа «Sex Pistols» дала свой первый концерт в художественной школе Святого Мартина, а 6 ноября 1995 г. был выпущен последний альбом группы «Queen» с участием Фредди Меркьюри



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:

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

- 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

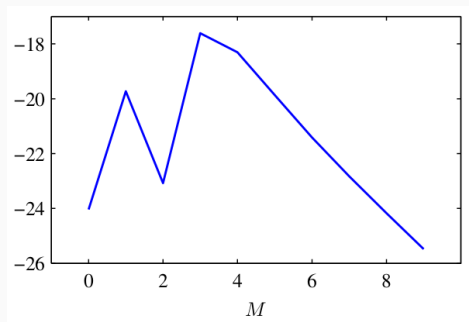
$$\int e^{-\frac{1}{2}(\mathbf{w}-\mu_N)^\top A(\mathbf{w}-\mu_N)} d\mathbf{w} = (2\pi)^{\frac{d}{2}} \sqrt{\det A^{-1}}.$$

- 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 $\beta \mathbf{X}^\top \mathbf{X}$:

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

для некоторых собственных векторов \mathbf{u}_i .

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

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

- Теперь будем оптимизировать логарифм маргинального правдоподобия $\log p(\mathbf{y}|\mathbf{X}, \alpha, \beta)$, 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

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

- 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

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

- Аналогично по β : можно заметить, что

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

- Значит,

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

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

- Итого по β получаем

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

ИНФОРМАЦИОННЫЙ КРИТЕРИЙ АКАИКЕ

- 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 $p(\mathbf{x}|\theta)$, $\theta \in \mathbb{R}^d$.
- Предположим, что мы обучили модель методом максимального правдоподобия, получив $p(\mathbf{x}|\theta_{\text{ML}})$.
- Давайте попробуем оценить, насколько модель $p(\mathbf{x}|\theta_{\text{ML}})$ отличается от неизвестного истинного распределения $p_{\text{data}}(\mathbf{x})$:

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

- Модель будет тем лучше, чем больше будет $\mathbb{E}_{p_{\text{data}}(\mathbf{x})} [\log p(\mathbf{x}|\theta_{\text{ML}})]$, и для оценки расхождения между $p(\mathbf{x}|\theta_{\text{ML}})$ 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:

$$\begin{aligned}\mathbb{E}_{p_{\text{data}}(\mathbf{x})} [\log p(\mathbf{x}|\theta_{\text{ML}})] &= \int p_{\text{data}}(\mathbf{x}) \log p(\mathbf{x}|\theta_{\text{ML}}) d\mathbf{x} \approx \\ &\approx \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}_n|\theta_{\text{ML}}) .\end{aligned}$$

- 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.

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

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

но это ожидание берётся по всему распределению.

- θ_0 — это «истинная» гипотеза максимального правдоподобия; при некоторых условиях регулярности можно доказать, что:
 - $\theta_{\text{ML}}(\mathbf{X}) \rightarrow \theta_0$ при $N \rightarrow \infty$;
 - 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

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

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

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

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

$$\sqrt{N}(\theta_{\text{ML}} - \theta_0) \rightarrow_{N \rightarrow \infty} N(0, J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)),$$

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

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

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

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

- 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 $p_{\text{data}}(\mathbf{x}) = p(\mathbf{x}|\theta_0)$, то

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

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

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

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

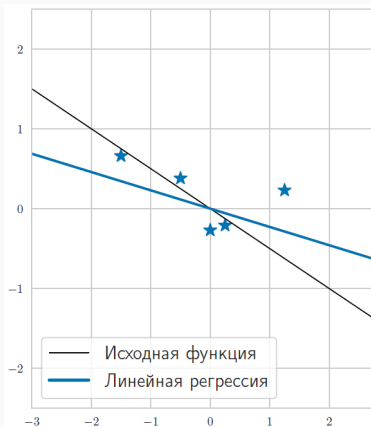
$$b(p_{\text{data}}) = \mathbb{E}_{\mathbf{X} \sim p_{\text{data}}} \left[\log p(\mathbf{X}|\theta_{\text{ML}}) - N \mathbb{E}_{p_{\text{data}}(\mathbf{z})} [\log p(\mathbf{z}|\theta_{\text{ML}})] \right],$$

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:

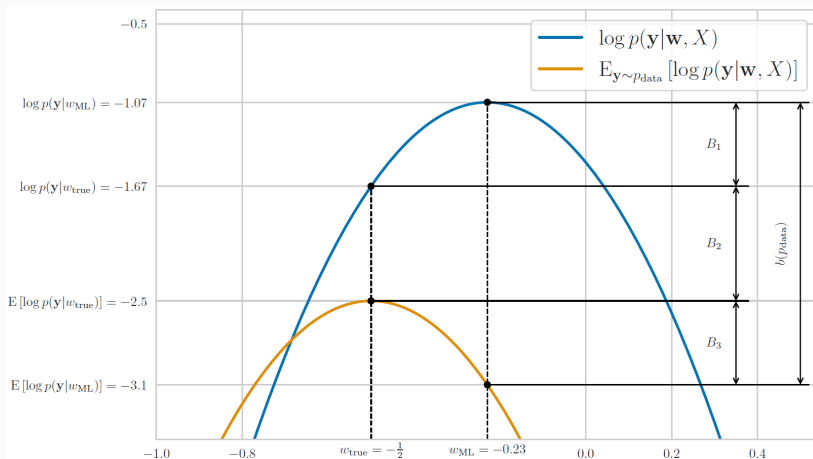
$$\begin{aligned} \text{IC}(\mathbf{X}, \theta) &= \\ &= -2 (\text{логарифм правдоподобия } \mathbf{X} \text{ в } \theta_{\text{ML}} - \text{оценка смещения}) = \\ &= -2 \sum_{n=1}^N \log p(\mathbf{x}_n|\theta_{\text{ML}}) + 2 (\text{оценка } b(p_{\text{data}})). \end{aligned}$$

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



Идея и вывод AIC

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



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

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

- Будем оценивать слагаемые по отдельности.

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

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

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

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

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

где

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

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

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

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

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

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

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

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

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

- Следовательно, в нашу оценку можно подставить

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

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

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

то есть

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

- Осталось только объединить три оценки:

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

- $I(\theta_0)$ 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):

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

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

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

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

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

- Очень простая формула!

- Пример: вернёмся к полиномиальной регрессии с логарифмом правдоподобия

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

- Давайте теперь для разнообразия будем дисперсию тоже обучать:

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

- Тогда при подстановке гипотезы максимального правдоподобия получится

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

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

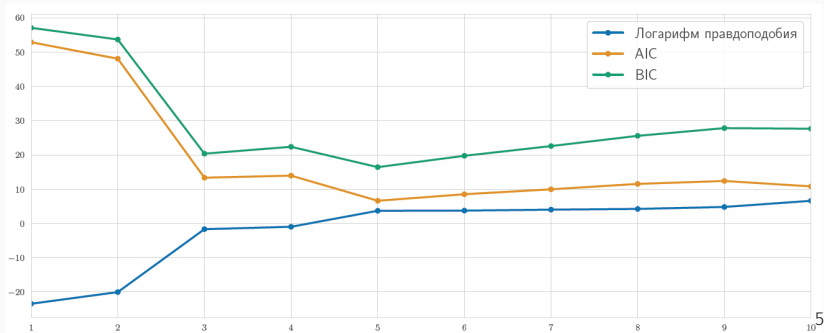
Идея и вывод AIC

- Тогда AIC и BIC получаются такими:

$$AIC(d) = N (\log (2\pi) + \log \sigma_{ML}^2 + 1) + 2(d + 2),$$

$$BIC(d) = N (\log (2\pi) + \log \sigma_{ML}^2 + 1) + (d + 2) \log N.$$

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



СПАСИБО!

Спасибо за внимание!

