

1 Информационные критерии (черновик)

1.1 Акаике информационный критерий (AIC)

Предложен Хиротугу Акаике в 1971 году, описан и исследован им же в 1973, 1974, 1983 годах. Первоначально аббревиатура AIC, предложенная автором, расшифровывалась как «"an information criterion"» («некий информационный критерий»), однако последующие авторы называли его "Akaike information criterion".

Критерий AIC основан на информационной теории. Предположим, что данные подчиняются некоей неизвестной модели f . Мы рассматриваем две модели g_1 и g_2 в качестве кандидатов для описания данных. Если бы мы знали f , то могли бы выбрать модель, у которой меньше $D_{\text{KL}}(f \parallel g_i)$ (интерпретируется как модель с меньшей потерей информации); называется расхождением/дивергенцией Кульбака-Лейблера $D_{\text{KL}}(P \parallel Q) = \int_X p \log \frac{p}{q} d\mu$, равно перекрестной энтропии минус просто энтропия. Но мы не знаем f . Akaike (1974) показал, что сравнивать модели-кандидаты можно и не зная f . Правда, это сравнение будет эквивалентно сравнению дивергенций только асимптотически по размеру выборки. Исходная расчетная формула критерия имеет вид:

$$AIC = 2k - 2\ln L,$$

где L — максимальное значение функции правдоподобия построенной модели, k — количество параметров в модели (оцененных по методу максимального правдоподобия). Но сами оценки максимального правдоподобия для сравнения моделей, вообще говоря, не нужны, а нужны только максимальные значения функции правдоподобия.

Заметим, что критерий AIC сравнивает критерии, но не оценивает качество моделей, т.е. обе модели могут плохо соответствовать данным, но одна из них 'менее плохо'.

Также заметим, что для применения AIC модели не обязаны быть вложенными.

1.2 Байесовский информационный критерий (BIC) или критерий Шварца (SC)

"Байесовский информационный критерий" (Bayesian information criterion — BIC) предложен Шварцем в 1978 году, поэтому иногда он называется также критерием Шварца (Schwarz criterion — SIC). Он разработан исходя из байесовского подхода и является наиболее часто используемой модификацией AIC:

$$BIC = SIC = k \ln n - 2 \ln L.$$

Как видно из формулы, данный критерий налагает больший штраф на увеличение количества параметров по сравнению с AIC, так как $\ln n$ больше 2 уже при количестве наблюдений 8.

Если коротко описать первоначальную идею метода, то в нем выбирается модель с большей апостериорной вероятностью (поэтому он и называется байесовским), далее используется разложение в значении второго порядка через информационную матрицу Фишера, и потом оценка интеграла от экспоненты методом Лапласа; все эти действия приводят к асимптотическому по размеру данных результату. Надо различать априорное распределение параметров модели (это непонятно, что такое) и априорные вероятности (предпочтения) для моделей.

Рассмотрим сначала использование априорных вероятностей для параметров. Тогда апостериорная вероятность

$$p(x | M) = \int p(x | \theta, M) \pi(\theta | M) d\theta,$$

где $\pi(\theta | M)$ — априорное распределение параметра θ в случае справедливости модели M .

Konishi and Kitigawa (2008, p. 217) вывели BIC с использованием аппроксимации Лапласа интеграла от экспоненты. Для этого сначала $\log$ -likelihood $\ln(p(x|\theta, M))$ раскладывается в ряд Тейлора до второго порядка в точке максимального правдоподобия $\hat{\theta}$ (дважды дифференцируемость предполагается):

$$\ln(p(x | \theta, M)) = \ln(\hat{L}) - 0.5(\theta - \hat{\theta})^T n \mathcal{J}(\theta)(\theta - \hat{\theta}) + R(x, \theta),$$

где $\mathcal{J}(\theta)$ — малое информационное количество Фишера (т.е. информационное количество Фишера, деленное на число индивидов n). Отбрасывая остаток $R(x, \theta)$ и используя линейное приближение $\pi(\theta | M)$ в окрестности $\hat{\theta}$, получаем приближение интеграла по θ аппроксимацией Лапласа:

$$p(x | M) \approx \hat{L}(2\pi/n)^{k/2} |\mathcal{J}(\hat{\theta})|^{-1/2} \pi(\hat{\theta}).$$

При больших n можно отбросить $|\mathcal{J}(\hat{\theta})|$ и $\pi(\hat{\theta})$, так как они являются $O(1)$. Поэтому получаем, что асимптотический результат не зависит от априорного распределения параметра

$$p(x | M) = \exp\{\ln \hat{L} - (k/2) \ln(n) + O(1)\} = \exp(-\text{BIC}/2 + O(1)),$$

где BIC определено выше. Здесь $\hat{L}$ либо (a) the Bayesian posterior mode (т.е. максимум апостериорной функции правдоподобия), либо (b) максимум функции правдоподобия при условии, что априорное распределение $\pi(\theta | M)$ имеет ненулевую производную в точке MLE. Таким образом, если $p(M)$ одинаковые, то сравнение апостериорных вероятностей сводится к сравнению BIC. Иначе, идет сравнение апостериорных вероятностей в виде

$$p(M | x) \propto p(x | M) p(M) \approx \exp(-\text{BIC}/2) p(M)$$

Есть рассуждение о том, что сравнение по BIC совпадает с применением метода MDL Minimum Description Length (из теории информации). Таким образом, исходно байесовский подход приводит к результату, не зависящему от априорного распределения параметра, и приводится к виду, имеющему отношение к теории информации. Так что название байесовский здесь, по-видимому, не имеет отношения к делу.

1.3 Частный случай оценок по МНК в нормальной модели.

Часто, когда есть модель и измерения с нормально распределенными ошибками, AIC и BIC можно записать в другом виде, откинув то, от чего сравнение не зависит, и используя оценку величины остатков в модели. Традиционно, используются те же обозначения, так как результат сравнения одинаков, хотя абсолютные значения меняются.

В частном случае классической нормальной линейной регрессии логарифмическая функция правдоподобия равна

$$\ln L = -n/2(1 + \ln 2\pi + \ln \hat{\sigma}^2),$$

где $\hat{\sigma}^2$ — состоятельная оценка (метода максимального правдоподобия) дисперсии случайной ошибки модели, равная отношению суммы квадратов остатков к объему выборки — SSE/n или, в других обозначениях, RSS/n (SSE - sum of squares, errors; RSS — residual sum of squares).

Подставив значение логарифмической функции правдоподобия в формулу AIC (с делением на объем выборки), а также не учитывая постоянные слагаемые 1 и $\ln 2\pi$ (так как при сравнении моделей они не влияют на результат) получим следующую формулу:

$$AIC = 2k/n + \ln \hat{\sigma}^2.$$

Аналогично,

$$BIC = k \ln n/n + \ln \hat{\sigma}^2.$$

Через RSS можно записать так (опять же, сами значения другие): $AIC = 2k + n \ln RSS$, $AIC = k \ln n + n \ln RSS$.

Заметим, что k — это полное число параметров в модели, включая, в случае линейной регрессии, свободный член и σ^2 .

Исходно, все равно, максимизировать функцию правдоподобия или минимизировать AIC/BIC по параметрам (в каждой модели), чтобы получить значение критерия. Однако, после преобразования критериев, это становится не так (ОМП могут не давать минимум AIC/BIC, когда мы их преобразуем).

1.4 Вспомогательная информация

1.4.1 Расстояние Кульбака-Лейблера

Для дискретных вероятностных распределений P и Q с числом элементарных событий n расхождение Кульбака-Лейблера распределения Q относительно распределения P (или «расстояние от P до Q ») определяется как:

$$D_{KL}(P \parallel Q) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

Другими словами, это математическое ожидание логарифмической разности между вероятностями p и q , где математическое ожидание берётся по распределению P . Расстояние Кульбака-Лейблера определено, только если $q_i = 0 \Rightarrow p_i = 0$, для всех $i = 1, \dots, n$ (абсолютная непрерывность). Всякий раз, когда $p_i = 0$, вклад i -го члена интерпретируется как ноль, потому что $\lim_{x \rightarrow 0} x \ln(x) = 0$.

Для k -мерных абсолютно непрерывных распределений P и Q расстояние Кульбака-Лейблера задаётся выражением

$$D_{KL}(P \parallel Q) = \int_X p(x) \ln \frac{p(x)}{q(x)} dx = H(P, Q) - H(P),$$

где $p(x)$ и $q(x)$ — функции плотности распределений P и Q соответственно, определённые на интервале $X \subseteq \mathbb{R}^k$.

Оценка по максимуму правдоподобия асимптотически эквивалентна оценке, полученной по минимуму дивергенции KL между распределением с оценкой параметра и истинным распределением или, эквивалентно, минимуму кросс-энтропии:

$$\hat{\theta} = \arg \max_{\theta} L_{P_{\theta}}(\mathbf{y}) = \arg \min_{\theta} D_{KL}(P_{\theta_0} \parallel P_{\theta}) = \arg \min_{\theta} H(P_{\theta_0}, P_{\theta}).$$

1.4.2 Метод Лапласа вычисления интеграла от экспоненты

Аппроксимация Лапласа выглядит так:

$$\int e^{Mf(\mathbf{x})} d\mathbf{x} \approx \left(\frac{2\pi}{M}\right)^{d/2} \frac{e^{Mf(\mathbf{x}_0)}}{|-H(f)(\mathbf{x}_0)|^{1/2}} \text{ при } M \rightarrow \infty \quad (1)$$

где $H(f)(\mathbf{x}_0)$ — гессиан f , вычисленный в точке $\mathbf{x}_0$, и $|\cdot|$ обозначает определитель матрицы. Предполагается, что гессиан отрицательно определен (функция выпукла вверх).