Akaike information criterion und Bayesian information criterion

Machine Learning

Vanessa Dietze

Universität zu Köln

6. Mai 2019

Gliederung

Motivation

- Repräsentation eines Prozesses durch statistisches Modell so gut wie nie exakt
- Finden eines „besten“ Modells
- Kriterien zur Entscheidungsfindung $\Rightarrow$ entstanden aus der Informationstheorie
- Reduzierung des Informationsverlustes
- Je weniger Informationen verloren gehen, desto besser ist das Modell
- Kriterien, die auch für „nonnested models“ gelten

Grundlagen

- <u>relative</u> Bewertung von statistischen Modellen anhand gegebener Daten
- Schätzung der verloren gegangenen Daten
- „goodness of fit“
- Strafterme für hohe Parameterwahl
- asymptotische Approximationen
- anwendbar für viele verschiedene Modellarten

Herleitung AIC

Definition (Kullback-Leibler Divergenz)

Sei $y := (y_1, \dots, y_n)^T$ *Vektor mit* n *unabhängigen zufälligen Beobachtungen mit der selben Verteilungsfunktion* $F(y)$*. Sei* $G(y)$ *eine Verteilungsfunktion eines weiteren Modells. Seien* $f(y)$ *und* $g(y)$ *die jeweiligen Dichtefunktionen. Der Zusammenhang der beiden Funktionen* $G(y)$ *und* $F(y)$ *wird dargestellt durch:*

$$I(f,g) := \mathbb{E}_{F(y)}\left[\log\left(\frac{f(Y)}{g(Y)}\right)\right] = \int \log\left(\frac{f(y)}{g(y)}\right) \mathrm{d}F(y)$$

Hierbei sei Y *die Zufallsvariable zu den genutzten Beobachtungen.*

Die Kullback-Leibler Information gibt an, wie viel Information verloren geht, versucht man $F(y)$ durch $G(y)$ zu approximieren.
Ziel: Minimierung des Informationsverlustes durch:

$$\min_{g(y)} I(f,g)$$

Die Kullback-Leibler Divergenz lässt sich auch schreiben als:

$$I(f,g) := \mathbb{E}_{F(y)}\left[\log\left(\frac{f(Y)}{g(Y|\theta)}\right)\right] = \mathbb{E}_{F(y)}\left[log(f(Y))\right] - \mathbb{E}_{F(y)}\left[log(g(Y|\theta))\right].$$

Sei nun $I_{\text{rel}}(f,g) = \mathbb{E}_{F(y)}\left[log(g(Y|\theta))\right]$.

Somit ist die relative Kullback-Leibler Divergenz die erwartete log-likelihood Funktion.

$$\begin{aligned}
\mathbb{E}_{\hat{F}(y)}\left[log\left(g\left(Y|\hat{\theta}_{\text{MLE}}(y)\right)\right)\right] &= \int \log\left(g\left(y|\hat{\theta}_{\text{MLE}}(y)\right)\right) \mathrm{d}\hat{F}(y) \\
&= \frac{1}{n}\sum_{i=1}^{n} log\left(g\left(y_i|\hat{\theta}_{\text{MLE}}(y)\right)\right) \\
&= \frac{1}{n} log(\hat{L})
\end{aligned}$$

Der Bias lässt sich berechnen durch

$$b(F) = \mathbb{E}_{F(y)}\left[\frac{1}{n}\log(\hat{L}) - \mathbb{E}_{F(y)}\left[\log\left(g\left(Y|\hat{\theta}_{\text{MLE}}(Y)\right)\right)\right]\right]$$

Somit wird die relative Kullback-Leibler Divergenz

$$\hat{I}_{\mathrm{rel}}(f,g) = \frac{1}{n} log(\hat{L}) - \hat{b}(F)$$

Das AIC ist nun definiert als das zweifach-negative von $\hat{I}_{rel}(f,g)$:

$$\mathrm{AIC} = -2\hat{I}_{\mathrm{rel}}(f,g) = -2\log \hat{L} + 2\hat{b}(F)$$

Durch Berechnung des Bias lässt sich dann die folgende Identität für das AIC aufschreiben:

$$\mathrm{AIC} = -2\log(\hat{L}) + 2k,$$

wobei k die Anzahl der Parameter oder Terme ist.

Relative Likelihood des Modells kann angegeben werden durch:

$$\exp\left(\frac{\mathrm{AIC}_{\mathrm{min}} - \mathrm{AIC}_i}{2}\right)$$

Falls die Datenmenge zu gering ist, dann wird eine Korrektur des AIC vorgenommen:

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

Wird die Datenmenge groß, dann fällt der hintere Term weg und es gilt $\mathrm{AIC}_c \rightarrow \mathrm{AIC}$

- Strafterm um vor "Overfitting" zu schützen
- Term oder Modell mit geringstem AIC wird ausgewählt

Allgemeines

Das BIC ist definiert über:

$$\text{BIC} = k\log(n) - 2\log(\hat{L}),$$

wobei $\hat{L}$ die Maximum-Likelihoodfunktion ist.

- Hierbei ist zu beachten, dass n deutlich größer sein sollte als k
- Größerer Strafterm als bei AIC
- für höhere Dimensionen muss eine Korrektur durchgeführt werden

Das BIC lässt sich auch über die Varianz des Fehlers definieren und lautet dann:

$$\text{BIC} = n\log(\hat{\sigma}^2) + k\log(n),$$

wobei $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$.

Vergleich

- BIC “bestraft„ härter als AIC (ab einer Datenmenge von n=8)
- AIC ist die Schätzung einer Konstanten plus der relativen Distanz zwischen der unbekannten Likelihood-Funktion der Daten und der gefitteten Likelihood-Funktion des Modells
- BIC ist die Schätzung einer Funktion der Posterior-Wahrscheinlichkeiten, dass das Modell das Richtige ist
- im Falle von einfachen Modellen werden die Kriterien äquivalent zum Likelihood-Quotienten Test für bestimmte alpha

- AIC hat höhere Wahrscheinlichkeit ein zu großes Modell zu wählen, BIC ein zu geringes Modell
- AIC besser für Modell, bei dem falsch-negatives Ergebnis schlechter ist als falsch-positives
- BIC besser für Modell, bei dem falsch-positives Ergebnis genauso schlecht oder noch schlechter ist als falsch-negatives Ergebnis
- Meist besser beide Kriterien kombiniert zu verwenden

Literatur

Verwendete Literatur

[HTF] Trevor Hastie Robert Tibshirani Jerome Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer Series in Statistics, Seiten 230-235, 2008

[BA] Kenneth P. Burnham David R. Anderson, *Multimodel Inference - Understanding AIC and BIC in Model Selection*, SOCIOLOGICAL METHODS and RESEARCH, Vol. 33, No. 2, November 2004, Sage Publications

[FFRA] Frank J. Fabozzi Sergio M. Focardi Svetlozar T. Rachev Bala G. Arshanapalli, *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*, Published 2014 by John Wiley and Sons, Inc.

Danke für eure Aufmerksamkeit.