Bayesianische Regression

lineare und logistische Modelle

Lona Koers

LMU

25. Juli 2025

Motivation und Intuition

TODO: gutes Beispiel

- Generalisierte Lineare Modelle (GLMs)
- Punktvorhersage vs. Verteilung vorhersagen
- warum reicht uns ein CI / PI

Outline

1. Bayesianische **lineare** Modelle

2. Bayesianische **generalisierte** lineare Modelle (GLMs)

3. Posterior Inference

Frequentistisches $\to$ bayesianisches lineares Modell

Annahmen:

1. i.i.d. Daten $\boldsymbol{D} = (\boldsymbol{y}, \boldsymbol{X})$
2. Kondition auf $\boldsymbol{X}$ (implizit)

Frequentistisches lineares Modell: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$

3. Gewichtsparameter $\boldsymbol{\theta}$ als Zufallsvariable interpretieren

Bayesianisches lineares Modell:

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$$

Modelldefinition (Prior-Verteilungen)

Prior-Annahme für $\boldsymbol{\theta}$ (und evtl. σ^2) notwendig $\rightarrow$ sehr vielseitige Modell-Anpassung möglich

1. Normal-Invers-Gamma Prior:

$$\begin{aligned}
\boldsymbol{\theta} \mid \sigma^2 &\sim \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \breve{\Sigma}) \\
\sigma^2 &\sim \mathsf{IG}(\breve{a}, \breve{b}) \\
\boldsymbol{\theta}, \sigma^2 &\sim \mathsf{NIG}(\breve{\boldsymbol{\mu}}, \sigma^2 \breve{\Sigma}, \breve{a}, \breve{b})
\end{aligned}$$

mit Prior Parametern: $\breve{\boldsymbol{\mu}}, \breve{\Sigma}, \breve{a}$ und $\breve{b}$

Vorteil: NIG-Prior ist mit Normalverteilungs-Likelihood konjugiert $\rightarrow$ exakte Inferenz möglich (mehr dazu später)

TODO: Bild

Uninformative Prior als Spezialfall der NIG-Prior

2. Uninformative Prior

z.B. mit NIG-Prior mit Prior Parametern

$$\breve{\boldsymbol{\mu}} = \mathbf{0}, \quad \breve{\Sigma}^{-1} = \mathbf{0} \text{ i.e., } \breve{\Sigma} \to \infty$$
$$\breve{a} = -\frac{p}{2}, \quad \breve{b} = 0$$

$\implies$ flache (und damit uninformative) Prior und maximaler Einfluss der Daten auf die Posterior:

$$\boldsymbol{\theta} \mid \sigma^2 \overset{a}{\sim} \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \infty) \implies p(\boldsymbol{\theta} \mid \sigma^2) \propto 1$$

TODO: Bild

Regularisierung: frequentistisch vs. bayesianisch

Erinnerung: *frequentistische* Regularisierung durch Minimierung von

$$\mathsf{PLS}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda \ \mathsf{pen}(\boldsymbol{\theta})$$

mit Regularisierungs-Parameter $\lambda > 0$.

Bayesianische Regularisierung durch Wahl der Prior-Verteilung für $\boldsymbol{\theta}$

Regularisierung durch Prior Wahl

3. Ridge Regularisierung

Frequentistisch [8, 9]: $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$
Bayesianisch [10]: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I})$ mit $\tau^2 \propto \frac{1}{\lambda}$

4. Lasso Regularisierung

Frequentistisch [14]: $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$
Bayesianisch [13]:

$$\boldsymbol{\theta} \mid \boldsymbol{\tau}^2 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\tau}^2 \boldsymbol{I})$$

$$\tau_j^2 \overset{\text{i.i.d.}}{\sim} \text{Exp}(0.5\lambda^2), \quad j = 1, \ldots, p$$

Problem: keine Variablenselektion (im Gegensatz zu frequentistischem Lasso)

→ Alternative Priors für Variablenselektion: Spike and Slab [11], Horseshoe [1], u.v.m.

Regularisierung in Anwendung

Vorteile von bayesianischer Regularisierung sind u.a.:

- Probabilistisches Modell trotz Regularisierung
- Regularisierung-Parameter muss nicht als Hyperparameter optimiert werden (z.B. durch Prior auf τ^2)
- Mehr Anpassungsmöglichkeiten durch Prior-Spezifikation

TODO: Bild regularization Priors + update

Outline

Bayesianisches LM $\rightarrow$ **GLM**

$$\text{LM: } \boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}) \quad \rightarrow \quad \text{GLM: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim F(g^{-1}(\boldsymbol{X}\boldsymbol{\theta}))$$

- Verteilungsannahme von $\boldsymbol{y}$ wird (äquivalent zum frequentistischen GLM) auf alle Verteilungen F der Exponentatialfamilie ausgeweitet
- Skala des linearen Prädiktors $\boldsymbol{X}\boldsymbol{\theta}$ wird mit der Link-Funktion g^{-1} angepasst

GLM $\rightarrow$ **logistisches** Modell

Bayesianisches logistisches Modell

$$\boldsymbol{y}_i \mid \boldsymbol{\theta} \sim \mathsf{Bin}(1, g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta})), \quad i = 1, \ldots, n$$
$$g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta}) = \sigma(\boldsymbol{x}_i\boldsymbol{\theta})$$

Für Beobachtungen $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ und Sigmoid-Link $\sigma(y) = \frac{\exp(y)}{1+\exp(y)}$

Prior Wahl

- Im Allgemeinen äquivalent zum LM möglich, z.B. Normalverteilung-Prior
- Verteilungen mit schweren Rändern (z.B. t-Verteilung, Cauchy Verteilung) verringern Separation und fördern Shrinkage [5, 7]
- Für Regularisierung können dieselben Priors verwendet werden [12, 4, 3]

Outline

Posterior Inference

Erinnerung: Bayes-Regel

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

wobei $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ die Modell-Likelihood ist.

bayesianisches LM: exakte Inferenz mit konjugierten Prioris

bayesianisches GLM: approximative Inferenz

Metropolis Hastings (MCMC)

Idee: Approximation der Posterior mit Markov chain Monte Carlo - was muss man für Regression anpassen?

Laplace Approximation (LA)

Idee: Approximation der Posterior mit einer Normalverteilung

- Mittelwert und Varianz werden mit IWLS berechnet

PPD

Literatur Empfehlungen

- Bayesianische Regression (v.a. für praktische Anwendung): Gelman et al. [6]
- Prior Verteilungen (v.a. Shrinkage): Erp, Oberski, and Mulder [3] und Celeux et al. [2]
- Software: z.B. `brms` in `R`, `PyMC` in `python`

Referenzen I

[1] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "The horseshoe estimator for sparse signals". English. In: *BIOMETRIKA* 97.2 (June 2010), pp. 465–480. ISSN: 0006-3444. DOI: 10.1093/biomet/asq017.

[2] Gilles Celeux et al. "Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation". In: *Bayesian Analysis* 7.2 (June 2012), pp. 477–502. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/12-BA716.

[3] Sara van Erp, Daniel L. Oberski, and Joris Mulder. "Shrinkage priors for Bayesian penalized regression". In: *Journal of Mathematical Psychology* 89 (Apr. 2019), pp. 31–50. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2018.12.004.

[4] Ludwig Fahrmeir, Thomas Kneib, and Susanne Konrath. "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection". In: *Statistics and Computing* 20.2 (2010), pp. 203–219. ISSN: 1573-1375. DOI: 10.1007/s11222-009-9158-3.

Referenzen II

[5] Andrew Gelman et al. “A weakly informative default prior distribution for logistic and other regression models”. In: *The Annals of Applied Statistics* 2.4 (Dec. 2008), pp. 1360–1383. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/08-AOAS191.

[6] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC, 2013. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

[7] Joyee Ghosh, Yingbo Li, and Robin Mitra. *On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression*. Feb. 2017. DOI: 10.48550/arXiv.1507.07170.

[8] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Applications to Nonorthogonal Problems”. In: *Technometrics* 12.1 (1970), pp. 69–82. ISSN: 0040-1706. DOI: 10.2307/1267352.

[9] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 12.1 (1970), pp. 55–67. ISSN: 0040-1706. DOI: 10.2307/1267351.

[10] David J. C. MacKay. “Bayesian Interpolation”. In: *Neural Computation* 4.3 (May 1992), pp. 415–447. ISSN: 0899-7667. DOI: 10.1162/neco.1992.4.3.415.

Referenzen III

[11] Tj Mitchell and Jj Beauchamp. “Bayesian Variable Selection in Linear-Regression”. English. In: *Journal of the American Statistical Association* 83.404 (Dec. 1988), pp. 1023–1032. ISSN: 0162-1459. DOI: 10.2307/2290129.

[12] R. B. O’Hara and M. J. Sillanpää. “A review of Bayesian variable selection methods: what, how and which”. In: *Bayesian Analysis* 4.1 (Mar. 2009), pp. 85–117. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/09-BA403.

[13] Trevor Park and George Casella. “The Bayesian Lasso”. English. In: *Journal of the American Statistical Association* 103.482 (June 2008), pp. 681–686. ISSN: 0162-1459. DOI: 10.1198/016214508000000337.

[14] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 0035-9246.