Kontekstinis duomenų, aprašomų erdviniais apibendrintais tiesiniais modeliais, klasifikavimas

Doktorantės Eglės Zikarienės ataskaita už 2019/2020 mokslo metus

Vadovas: prof. dr. Kęstutis Dučinskas

Konsultantas: prof. dr. Julius Žilinskas

Doktorantūros pradžios ir pabaigos metai: 2013 – 2021

2020-10-22

Ataskaitos turinys

1. Tyrimo objektas ir tikslai.

2. Planuojami doktorantūros rezultatai.

3. 2019/2020 m. m. darbo planas

- Dalyvavimas konferencijoje
- Publikacija
- Disertacijos rengimas
    - Teorinis tyrimas
    - Empirinis tyrimas

Tyrimo objektas ir tikslai

Tyrimo objektas:

- Erdvinių duomenų modeliai,
- Klasifikavimo procedūra, paremta Bajeso diskriminantine funkcija (BDF)

Tyrimo tikslas:

Sudaryti ir ištirti BDF išraiškas eksponentinės šeimos skirstiniams ir eliptinių šeimos *t* skirstiniui, siekiant praplėsti klasifikavimo procedūros pritaikymo galimybes.

Planuojami doktorantūros rezultatai

Planuojami rezultatai:

- BDF išraiškos Bernulio, Binominis, Puasono, Gamma, Beta, *t* skirstiniams;
- Tikslios tikrųjų klasifikavimo klaidų formulės;
- Vidutinių klasifikavimo klaidų tikimybių asimptotinės aproksimacijos;
- Pasiūlytos procedūros teorinis pagrindas;
- Pasiūlytos procedūros algoritminė realizacija;

2019/2020 m. m. darbo planas

Dalyvavimas konferencijoje:

Zikarienė E., Dučinskas K. Implementation of generalized additive models for spatial beta regression. Computer data analysis and modeling: stochastics and data science. Minskas, Baltarusija 2019 m. rugsėjo 18 - 22 d.

2019/2020 m. m. darbo planas

Mokslinė publikacija:

Zikarienė E., Dučinskas K.
Implementation of generalized additive models for spatial beta regression. Proceedings of the XII International Conference. Computer data analysis and modeling: stochastics and data science. p. 341-343.

Dučinskas K., Zikarienė E. 2015.
Actual error rates in classification of the T-distributed random field observation based on plug in linear discriminant function.
Informatica. vol.26, p. 557-568

2019/2020 m. m. darbo planas

Disertacijos rengimas:

- Teorinis tyrimas. Erdvinių duomenų, aprašomų apibendrintais tiesiniais modeliais, kontekstinio klasifikavimo klaidos tikimybių analitinių išraiškų ir aproksimacijų išvedimas naudojant klasikinius ir Bajeso parametrų įvertinimus diskrečių požymių atveju.

- Empirinis tyrimas. Siūlomų klasifikavimo procedūrų empirinis tyrimas, palyginimas ir optimizavimas naudojant generuotus ir realius duomenis.

Kontekstinis klasifikavimas

Tikslas:

$$D_t^*(\bullet) = \min_{D_t}\left(1 - \sum_{\vec{l}=1}^{m} \pi_l P_l\left(D_t(\bullet) = l\right)\right)$$

čia $D_t(\bullet)$ - klasifikavimo taisyklė, P_l - teisingo klasifikavimo tikimybė, π_i - apriorinės klasių tikimybės

Bajeso taisyklė:

$$D_t^B(Z_0) = \arg\max_{\{k=1,\ldots m\}}\left\{\sum_{l=1}^{m} \pi_l p_l\left(Z_0 \middle| \mathbf{t}, \boldsymbol{\Psi}\right)\right\}$$

π_i - apriorinės klasių tikimybės, $p_l\left(Z_0 \middle| \mathbf{t}, \boldsymbol{\Psi}\right)$ - klasės sąlyginio tankio funkcija, $\mathbf{t}$ - mokymo aibė,

$\boldsymbol{\Psi}$ - nežinomų parametrų vektorius

Bajeso diskriminantinė funkcija

Bajeso diskriminantinė funkcija,

$$W_{lk}^{B}\left(Z_0,\boldsymbol{\Psi}\right)=\ln\left(\frac{p_l\left(Z_0\left|\mathbf{t},\boldsymbol{\Psi}\right.\right)}{p_k\left(Z_0\left|\mathbf{t},\boldsymbol{\Psi}\right.\right)}\right)+\gamma_{lk}$$

$p_l\left(Z_0\left|\mathbf{t},\boldsymbol{\Psi}\right.\right)$ - klasės sąlyginio tankio funkcija, γ_{kl} - apriorinių tikimybių logaritmų santykis

Tikroji klaidos tikimybė:

$$P_0^B\left(\hat{\boldsymbol{\Psi}}\right)=1-\sum_{l=1}^{m}\pi_l P\hat{C}_l$$

π_l - apriorinės klasių tikimybės, $P\hat{C}_l$ - teisingo klasifikavimo tikimybė

Sąlyginiai eksponentiniai skirstiniai

Tankio funkcija:

$$\mathrm{P}\left(z(s_i)\middle|\left\{z(s_j): j\neq i\right\}\right)=\exp\Big[A_i\left(\left\{z(s_j): j\neq i\right\}\right)B_i\left(z(s_i)\right)+C_i\left(z(s_i)\right)$$
$$+D_i\left(\left\{z(s_j): j\neq i\right\}\right)\Big],\qquad i=1,\ldots,n,$$

čia $\{B_i(\cdot)\}$ pakankamos statistikos, $\{D(\cdot)\}$ - parametrų funkcija, priklausanti nuo pasirinktos kaimynų schemos, $\{C_i(\cdot)\}$ - funkcija taške z_i, nepriklausanti nuo parametrų.

Natūralus parametras:

$$A_i\left(\left\{z(s_j): j\neq i\right\}\right)=\alpha_i+\sum_{j=1}^{n}\eta_{ij}B_j\left(z(s_j)\right),\qquad i=1,\ldots,n,$$

α_i - trendo parametras, η_{ij} - erdvinio ryšio parametras.

Klasifikavimas naudojant BDF

BDF Algorithm:

Inputs: Data set $\left\{Z(s): s \in D \subset R^p\right\}$, Model M for population Ω_l, parameter estimation function *f*, prior probability function *g*, neighbourhood system $\partial s \subset D$, $Z(s_0)$ – classification observation.

Algorithm: Fit the model M to the data $\left\{Z(s): s \in D \subset R^p\right\}$

Estimate unknown parameters: $\hat{\boldsymbol{\Psi}} = f\left(Z(s), \Psi\right)$

Evaluate the prior probability function for class *l*: $\hat{\pi}_l = g\left(\left\{Z(s)\middle|\mathbf{t}, y = l\right\}, \partial s\right)$

Evaluate the BDF function: $W\left(Z(s_0), \hat{\boldsymbol{\Psi}}\right)$

Make decision for $Z(s_0)$: $\hat{W}_{lk}^B\left(Z(s_0), \hat{\boldsymbol{\Psi}}\right) \geq 0, l = 1,..,m, k \neq l$

Evaluate probability of missclasification (actual error rate): $P_0^B\left(\hat{\boldsymbol{\Psi}}\right)$

Outputs: Class label for $Z(s_0)$

Probability of missclassification: $P_0^B\left(\hat{\boldsymbol{\Psi}}\right)$

Beta skirstinys

Sąlyginio tankio funkcija:

$$p\left(Z_0^l = z_0 \middle| T_{-i} = t_{-i}, y_i = l; \mu_i^l \phi_i^l\right) = \exp\left\{\left(\mu_i^l \phi_i^l\right)\ln z_0 + \left(\left(1-\mu_i^l\right)\phi_i^l - 1\right)\ln\left(1-z_0\right) - \ln\left(B\left(\mu_i^l \phi_i^l; \left(1-\mu_i^l\right)\phi_i^l\right)\right)\right\}$$

$$\mu_i^l = E\left(Z_i \middle| T_{-i} = t_{-i}, y_i = l\right) = \frac{1 + A_{i1}^l}{2 + A_{i1}^l + A_{i2}^l} \text{ ir } \phi_i^l = 2 + A_{i1}^l + A_{i2}^l$$

čia μ_i^l - sąlyginis vidurkis, ϕ_i^l - tikslumo parametras.

Natūralus parametras: $A_{i1}^l = \alpha_{i1}^l - \sum_{i \neq j} \eta_{ij} \ln\left(1 - z_j\right)$ $A_{i2}^l = \alpha_{i2}^l - \sum_{i \neq j} \eta_{ij} \ln\left(z_j\right)$,

čia $\alpha_{ik}^l = x_i' \beta_k^l$, kai x_i' - vektorius aiškinamųjų kintamųjų erdvės taške s_i β_k^l - nežinomi regresijos parametrai, $k, l = 1, 2$.

Pakankamos statistikos: $B\left(z_i\right) = \left[\log\left(z_i\right), \log\left(1 - z_i\right)\right]^T$

Parametrų vertinimo funkcija :

$$L\left(\Psi\right) = \prod_{i=1}^{n_1} p_{i1} \prod_{i=n_1+1}^{n} p_{i2} \quad \hat{\Psi} = \arg\max_{\Psi}\left(\log\left(L\left(\Psi\right)\right)\right)$$

Empirinis tyrimas

Duomenys:

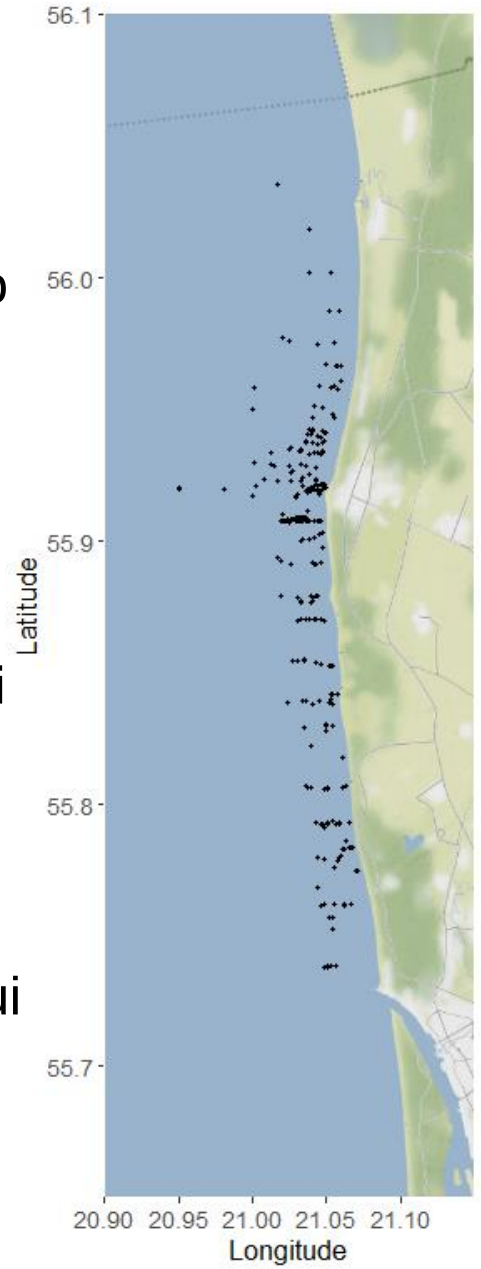
- Kiekybinis atsako kintamasis $\{Z(s): s \in D \subset R^p\}$ - Šakotojo banguolio padengimas gardelėje $1m \times 1m$.
- Kokybinis kintamasis (klasės žymė) – dugno paviršius (smėlis, ne smėlis).

Uždavinys:

- Identifikuoti dugno paviršių, kurį vizualiai sunku įvertinti dėl ten augančių augalų.

Siūlomas sprendimas:

- Atlikti klasifikavimą taikant BDF, kuri vertinama taikant sąlyginius tankius stebimam dumblių padengimo kiekiui
- Pilnai sąlyginio Beta tankio modelis.



Literatūra

Cepeda-Cuervo E., Gamerman D. (2005). Bayesian methodology for Modelling Parameters in the two Parameter Exponential Family. Revista Estadística. vol. 57, pp. 93-105.

Ferrari S.L.P., Cribari-Neto F. (2004). Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics. vol. 31, pp. 799-815.

Kalhori Nadrabadi L., Mohhamadzadeh M. (2018). Bayesian Inference for Spatial beta Generalized Linear Mixed Models. Journal of Sciences, Islamic Republic of Iran. vol. 29(2), pp. 173-185.

Martins TG, Simpson D, Lindgren F, Rue H (2013). Bayesian Computing with INLA: New Features. Computational Statistics and Data Analysis. vol. 67, pp. 68-83.

Rue H, Martino S, Chopin N. (2009). Approximate Bayesian Inference for Latent Gaussian Models using Integrated Nested Laplace Approximations (with discussion). Journal of the Royal Statistical Society B. vol. 71, pp. 319-392.

Simas A. B., Barreto-Souza W., Rocha A. V. (2010). Improved Estimators for a General Class of Beta Regression Models. Computational Statistics and Data Analysis. vol. 54, pp. 348-366

Wood S.N. (2017). Generalized Additive Models: An Introduction with R (2nd Edition). Chapman and Hall/CRC, Boca Raton.

Zhang Y. (2002). On Estimation and Prediction for Spatial Generalized Linear Mixed Models. Biometrics. Vol. 58, pp. 129-136.

Zuur A. F., Ieno E. L., 2018. *Beginner's Guide to Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA. Volume II: GAM and zero – inflated models*, Highland Statistics Ltd. United Kingdom.

Ačiū