Master projects

Aliaksandr Hubin

Associate Professor, BIAS
NMBU

aliaksandr.hubin@nmbu.no

06.04.2022

Bayesian Generalized Nonlinear Model (JAIR 2021) [3]

Sample of observations $i = 1, \ldots, n$

- Y_i . . . response data;
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ . . . p-dimensional vector of input covariates.

Specification of the model

From input variables a huge (but finite) number of features can be generated: $F_j(\boldsymbol{x}_i)$, $j = 1, ..., q$ (consider ordering w.r.t. complexity)
The **model** is then specified as GLM:

$$Y_i|\mu_i, \phi \sim \mathfrak{f}(y|\mu_i, \phi)\ , i = 1, ..., n; \tag{1}$$

$$\mathrm{h}(\mu_i) = \beta_0 + \sum_{j=1}^{q} \gamma_j \beta_j F_j(\boldsymbol{x}_i, \boldsymbol{\alpha}_j)\ . \tag{2}$$

- $\mathfrak{f}(\cdot|\mu, \phi)$ - density from exponential family with mean μ_i and dispersion parameter ϕ;
- h - link function;
- $\beta_j \in \mathbb{R}$ - regression coefficients of j-th feature;
- $\gamma_j \in \{0, 1\}$ - indicator variable for j-th feature.

Hierarchy of the features

A feature $F_j(\boldsymbol{x}, \boldsymbol{\alpha}_j), j \in \{p+1, ..., q\}$ can be constructed recursively through:

$$F_j(\boldsymbol{x}, \boldsymbol{\alpha}_j) = v(\boldsymbol{\alpha}_j^T \mathrm{F}_{1:j-1}(\boldsymbol{x}))$$

$v \in \mathcal{G}$ is one of the allowed basic function from set $\mathcal{G}$.

Types and meaning of functions in $\mathcal{G}$

- Neural Networks: $\mathrm{logit}(x)$, $\tanh(x)$, $\mathrm{erf}(x)$, $\mathrm{ReLU}(x)$;
- Polynomials: $F_k(\boldsymbol{x}) * F_l(\boldsymbol{x}) = \exp\left(\log(F_k(\boldsymbol{x})) + \log(F_l(\boldsymbol{x}))\right)$;
- CART: $\mathrm{I}(x \geq 1)$;
- MARS: $\max\{0, x - t\}$ and $\max\{0, t - x\}$;
- Fractional polynomials: $x^{\frac{1}{a}} = \exp\left(b \log(x)\right), b = \frac{1}{a}$;
- **Logical *AND*, *OR* and *NOT*:** $L_k \wedge L_l = L_k * L_l$, $L_k \vee L_l = L_k + L_l - L_k * L_l$, **and** $\overline{L}_k = 1 - L_k$.

Examples on Inference [3]

Dataset: $n = 223$ exoplanets. We want to recover 2 basic physical laws. **Input variables** include: *TypeFlag, RadiusJpt, PeriodDays, PlanetaryMassJpt, Eccentricity, HostStarMassSlrMass, HostStarRadiusSlrRad, HostStarMetallicity, HostStarTempK, PlanetaryDensJpt* denoted as x_1-x_{10}.

Planetary mass

$$m_p \approx K_1 R_p^3 \times \rho_p.$$

Planetary mass m_p is proportional to cube of radius R_p times the density of the planet ρ_p

Kepler's third law

The square of the orbital period P of a planet is directly proportional to the cube of the semi-major axis a of its orbit:

$$a \approx K_2 \left(P^2 M_h\right)^{\frac{1}{3}}.$$

Examples on Prediction [3]

Three binary classification tasks

- Asteroids data: Mean anomaly, Inclination, Argument of perihelion, Longitude of the ascending node, Rms residual, Semi major axis, Eccentricity, Mean motion, Absolute magnitude;
- Breast cancer data: Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension;
- Spam emails: 58 characteristics, including 57 continuous and 1 nominal variable, where most of these are concerned with the frequency of particular words or characters. 3 provide different measurements on the sequence length of consecutive capital letters.

Regression task

- Abalone age prediction: sex, length, diameter, height, whole weight, shucked weight, viscera weight and shell weight.

Open questions for master theses (BGNLM)

Scalability within the populations:

- Try variational inference within the populations;
- Further develop subsampling in MCMC.

Study in detail particular model sub-classes:

- Study fractional polynomials as a special case;
- Study logical causal inference through logic regression.

Other topics:

- Model selection in Bayesian neural networks and VAE;
- Developing neural networks with latent Gaussian fields;
- Weak supervision in NLP (SKWEAK);
- Fraud detection in Bank transactions (with DNB and NR);
- Python library for BGNLM (R library with C++ blocks today).

Inference and predictions with Fractional Polynomials

Fractional polynomials is a popular class of statistical models that creates an alternative to regular polynomials and allow flexible parameterization of continuous variables. There exist several implementation of fractional polynomials in the literature. The aim of this project would be to prepare and carry out a simulation study, where the ability of the approaches to recover the correct polynomials will be studied. The focus will be on BGNLM and EMJMCMC or GMJMCMC packages here and fractional polynomials will be addressed as a specific case. Also, comparisons of predictive performance of the studied approaches on real data sets will be performed. The project might also require careful specification of the priors for the problem at hand.

Logical causalities by Bayesian logic regression

Michael Baumgartner and Christoph Falk from the University of Bergen recently published a paper entitled Configurational causal modeling and logic regression [5]. In this paper, they show that logic regression model can be used to recover causal structures featuring conjunctural causation and equifinality.The aim of this master project, thus, would be to check if a Bayesian version of logic regression can perform the task better than the frequentist version that is used in [5]. The project will involve carefully reading through the paper by Baumgartner and Falk, understanding the problem and repeating their simulation study (their code and data are available) with a version of Bayesian logic regression implemented in EMJMCMC or GMJMCMC packages or in LogicReg package. Also real data analysis will be provided. The project might also require careful specification of the priors for the problem at hand.

Deep learning with latent Gaussian spatio-temporal structures

Random effects combined with more modern machine learning approaches [3], in which case more flexible functions functions are allowed, has been less explored. In this project, we will develop a class of neural networks with latent Gaussian fields models and perform experiments with such models on a relative large data set of lamb weights, a total of 1358139 observations from 182 municipalities in Norway, with some covariates that are on individual level (e.g. age, age of mother) while others are on regional level (e.g. elevation). Other relevant data sets with spatio-temporal structure like a data set of forest fires in the Nordics may be considered instead or in addition.

Weak supervision in NLP (SKWEAK)

In [6,7] we developed a SKWEAK library for weak supervision in NLP. Currently, it is based on a hidden Markov model with the latent states corresponding to the true NER classes, whilst the observed processes are vectors of multinomials giving classes produced by the pre-trained (on other domains) labelling functions. We then run EM inference and use the sequence of most probable latent states as the "true" underlying classes for our NER on a new domain. In the project, we can implement and test adding other sources of signal to the model (raw emdeddings) and/or perform model selection on the components of the embeddings and labelling functions.

Fraud detection in Bank transactions (with NR and DNB)

Every time period we observe 0 or more transactions between all pairs of customers in a bank. In this thesis, one would apply variational inference to learn the commuting parties across all periods and conditional on that distributions of the transactions sizes between the commuting parties in each period. Then, blocks of parties with significantly different patterns of transactions in different periods will be identified using posterior probabilities of the sizes of transactions with commuting counter-parties. The implementation will be in either torch library in R or in Pytorch.

Literature

1. Hubin, A., Storvik G. (2018). **Mode jumping MCMC for Bayesian variable selection in GLMM.** *Journal of Computational Statistics and Data Analysis; 2018 November; 127:281-297.*
2. Hubin, A., Storvik G., Frommlet F. (2018). **A novel algorithmic approach to Bayesian Logic Regression (with Discussion)**. *Bayesian analysis; 2020.*
3. Hubin, A., Storvik G., Frommlet F. (2018). **Flexible Bayesian Nonlinear Model Configuration.** *Journal of Artificial Intelligence Research (2021)*.
4. Hubin, A., Storvik G. (2022). **Variational Bayes for Inference on Bayesian Neural Networks Under Model and Parameter Uncertainty.** *in revisoin for the second review in Bayesian analysis*;
5. Baumgartner, M., Falk Christoph. (2021). **Configurational Causal Modeling and Logic Regression.** *Multivariate Behavioral Research (2021)*;
6. Lison, P., Barnes, J., Hubin, A., Touileb, S. (2020). **Named Entity Recognition without Labelled Data: A Weak Supervision Approach.** In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1518-1533).
7. Lison, P., Barnes, J., Hubin, A. (2021). **skweak: Weak Supervision Made Easy for NLP.** In Proceedings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (System Demonstrations Track).