Approximate Bayesian Binary and Ordinal Regression with Structured Uncertainty in the Inputs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a novel approach to binary and ordinal prediction with structured uncertainty in the input variables. It is based on efficiently approximating the prediction model conditional on the inputs and then marginalizing the conditional model over the input space using Monte Carlo approximation. For efficiency, the well-known Laplace approximation is used for the binary case and we derive a similar approximation for the ordinal case. Empirical evaluation on sports data shows that the proposed approach substantially improves forecasting accuracy and highlights the severity of the problem of uncertainty in the input variables in sports.

1 Introduction

This work is motivated by a problem in sports forecasting [6] – predicting future outcomes from past performances. To illustrate the problem, imagine we are interested in predicting the outcome of a basketball game between two teams. Typically, we would first compile a set of past games with outcomes and relevant performance-related variables, which are usually count variables, such as shots made/missed, rebounds, etc. Then we aggregate these variables across past performances and possibly transform them into variables that are known to be good descriptors of team quality and predictors of future performance, e.g. shooting percentage. Finally, we would use some binary regression model to model the relationship between the inputs and the outcomes. What has so far not been taken into account in related work is that past performance variables are noisy and that the noise is transferred to any input variables we derive from them. Fitting a model and failing to account for this uncertainty will result in being overconfident in the model parameters. Thus, forecasting accuracy could be improved by first modeling the uncertainty in the inputs and second taking this uncertainty into account in the model, and we focus on developing a Bayesian approach for the second task, while addressing the first in a limited case with prior information.

2 Methods

Let $X \in \mathbb{R}^{n \times m}$, be our training data and $y \in \mathbb{R}^n$ the target variable. We train a model that can produce predictions y^* for a given new x^* by using the training data X. We adopt the Bayesian approach $p(\beta|X, y) \propto p(X, y, |\beta)p(\beta)$, where β are the parameters of the model. For the prediction setting, we obtain the probability by marginalizing over the model parameters: $p(y^*|x^*, X, y) = \int p(y^*|\beta, x^*)p(\beta|X, y)d\beta$, where the data are treated as constant. A more general problem arises in our setting, where we instead treat the inputs X and x^* as random variables, with densities $p(X|w)$ and $p(x^*|w^*)$, where w and w^* are known. In this case, to train a model and obtain its posterior

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

distribution $p(\beta|w,y)$ we must marginalize over the input space of possible training data sets X:

$$p(\beta|w,y) = \int p(\beta, X|w,y)dX = \int p(\beta|X,y)p(X|w)dX.$$

Similarly, to obtain a prediction for y^* we must not only marginalize over the parameters β, but also over the distribution of the test sample x^* as follows:

$$p(y^*|w^*,w,y) = \int_{x^*}\int_{\beta} p(y^*|x^*,\beta)p(x^*|w^*)p(\beta|w,y)dx^*d\beta. \quad (1)$$

This approach is general, since it can be applied with any model that produces a posterior probability distribution over its parameters $p(\beta|X)$ and a distribution over its predictions $p(y^*|x^*,X)$ for a given test sample x^*. What remains for the model to be fully specified is defining the distributions that generate the training and test data sets. The integral in Eq. 1 will generally be intractable even for the simplest of Bayesian models and is typically approximated using Monte Carlo: $E[y^*|w^*,w,y] \approx \frac{1}{N}\sum_{i=1}^{N} y^*_{(i)}$, where $y^*_{(i)}$ is a random sample from the posterior predictive distribution in Eq. 1 and can be obtained by sequentially sampling $X_{(i)}$ from $p(X|w)$, $\beta_{(i)}$ from $p(\beta|X = X_{(i)},y)$, $x^*_{(i)}$ from $p(x^*|w^*)$, and finally, $y^*_{(i)}$ from $p(y^*|x^* = x^*_{(i)},\beta = \beta_{(i)})$. The densities $p(X|w)$, $p(x^*|w^*)$ represent our structural measurement error model and are in most practical cases easy to sample from efficiently. The densities $p(\beta|X = X_{(i)},y)$ and $p(y^*|x^* = x^*_{(i)},\beta = \beta_{(i)})$ are the posterior and posterior predictive for the selected prediction model, conditional on the inputs being fixed, and we have to be able to efficiently sample from them. This implies that we need an efficient prediction model or, in the case of Bayesian models, which are typically computationally intensive, an efficient structural approximation to $p(\beta|X = X_{(i)},y)$.

2.1 Approximate proportional-odds model

For a binary outcome, Bayesian logistic regression is commonly used, typically with the well-known Laplace approximation (see [1], pages 213-215) for the posterior. We now derive a Laplace approximation to the proportional-odds model (ordinal logistic regression), which is the most commonly used model for the ordinal setting [3]. Let n and m again be the be the number of samples, and input variables, respectively and k the number of (ordered) categories. The model is based on the assumption that the odds of all binary decisions between categories are proportional to each other or, equivalently, that the $k-1$ logit surfaces are parallel:

$$logit(P(Y \leq j|x)) = log(\frac{P(Y \leq j|x)}{P(Y > j|x)} = \beta x + \alpha_j,$$

for $j \in \{1,..,k-1\}$, where β and α_j are parameters. For convenience, let $\alpha_0 = -\infty$ and $\alpha_k = +\infty$. The outcome probabilities can then be written as:

$$P(Y = j,x) = P(Y \leq j|x) - P(Y \leq j-1|x) = \sigma(\beta x + \alpha_j) - \sigma(\beta x + \alpha_{j-1}), \quad (2)$$

for $j \in \{1,..,k\}$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

The proportional odds model and its generalizations (see [4]) have received very little attention in the Bayesian setting. We now derive the Laplace approximation to the posterior of this model. First, we place priors on the parameters. To ensure in-order intercepts, we introduce parameters d_j, $j \in \{1,..,k-1\}$ and a stick-breaking reparameterization of the $k-1$ parameters a_j with $a_j = \Sigma_{i=1}^{j} = d_i$. We place flat priors on the parameters $p(d_i) \propto 1$ and all d_i are restricted to be positive, except for d_1. As in the binary case, we place normal priors on the coefficients $\beta_1,...,\beta_m \sim \mathcal{N}(0,\sigma_\beta)$. The model's likelihood is:

$$p(\beta,d,\mathcal{D}) \propto \prod_{i=1}^{n}\prod_{j=1}^{k}(R_{i,j} - R_{i,j-1})^{y_i=j}\prod_{i=1}^{m} exp(\frac{\beta_i^2}{2\sigma_\beta^2}),$$

where $R_{i,j} = \sigma(\beta x_i + \alpha_j)$, and the log-likelihood is:

$$L = C - \frac{1}{2\sigma_\beta^2}(\beta \circ \beta) + \sum_{i=1}^{n}\sum_{j=1}^{k} I(y_i = j)log(R_{i,j} - R_{i,j-1}),$$

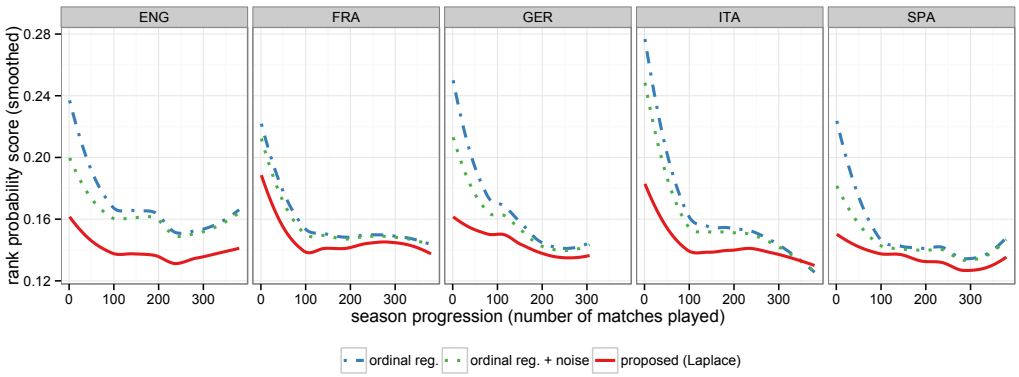

Figure 1: Prediction errors over the course of a season, averaged across all seasons and for each football competition separately.

where I is the indicator function and C is a constant. The gradient of the log-likelihood cannot be written as succinctly as is the case with logistic regression. Instead, we start with the derivative for some parameter θ:

$$\frac{\partial}{\partial\theta}L = -\frac{1}{2\sigma_\beta^2}\frac{\partial}{\partial\theta}(\beta\circ\beta) + \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\partial}{\partial\theta}L_{i,j},$$

where $\frac{\partial}{\partial\theta}L_{i,j} = 0$ if $y_i \neq j$ and

$$\begin{aligned}\frac{\partial}{\partial\theta}L_{i,j} &= \frac{1}{R_{i,j}-R_{i,j-1}}\frac{\partial}{\partial\theta}(R_{i,j}-R_{i,j-1}) = \\ &= \frac{1}{R_{i,j}-R_{i,j-1}}\left[R_{i,j}(1-R_{i,j})\frac{\partial}{\partial\theta}(\beta x_i+\alpha_j)R_{i,j-1}(1-R_{i,j-1})\frac{\partial}{\partial\theta}(\beta x_i+\alpha_{j-1})\right]\end{aligned}$$

otherwise. Due to negligible effect on accuracy and running time, we omit the derivation of the Hessian and use a numerical approximation for our empirical evaluation.

3 Results

We test our binary approach on basketball game outcomes (which always have a winner) and ordinal approach on football match outcomes (where draws are possible). In both cases, the count data are first preprocessed to model the uncertainty in the input variables. As a baseline, binary logistic regression and ordered logistic regression are included, using mean counts as well as using noisy test cases. This is obtained by treating the test input variables as random, using the same structural model for the inputs as for the proposed model, and approximating the expected prediction of the baseline models with Monte Carlo sampling. For the binary case we also include the proposed model without marginalization - jointly modeling β and X, including the uncertainty in the form of an informative prior on X. We implement this model in the probabilistic programming language and tool for Bayesian inference Stan [5]. Our empirical evaluation procedure is a straightforward measurement of out-of-sample forecasting accuracy, while respecting the time line. We use train-test season pairs, training on one season and forecasting on the next. Only data available prior to a match are used in forecasting it. We measure forecasting accuracy with mean squared error (MSE) in the binary case and the rank probability score (RPS) in the ordinal case [2].

3.1 Data sets and preprocessing

The basketball data used in our experiments consist of all the regular season and play-off games in the past 13 seasons (12 train-test season pairs) of the National Basketball Association (NBA) from 2001/02 to 2013/14. [1] The count variables included in the data are counts of two-point and

[1] The data were obtained from http://www.basketball-reference.com/.

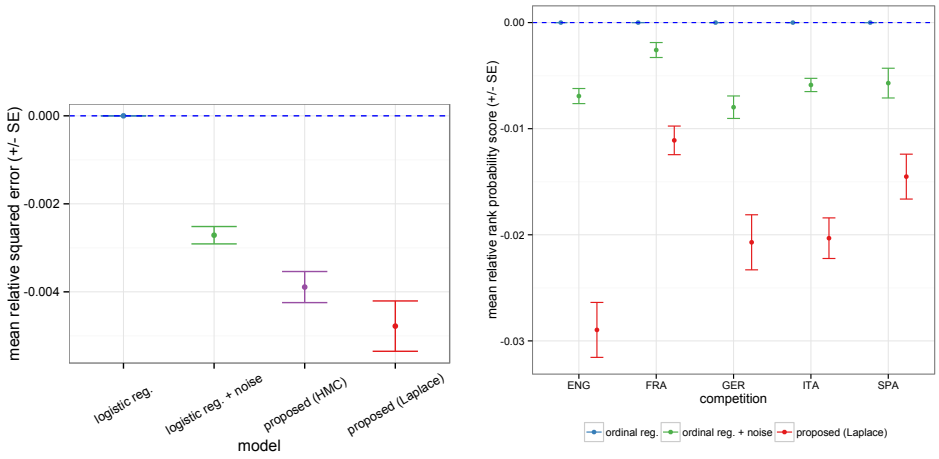

Figure 2: Estimated prediction errors for the basketball and football model, respectively across all train-test season pairs. All errors are relative to the baseline for comparison (logistic and ordinal regression, respectively).

three-point shots made and missed, turnovers, offensive and defensive rebounds. We use these counts indirectly by transforming them into 8 ratios, described in [7], which are known to be good predictors. Our football data set consists of 5 complete seasons: 2010/11-2014/15 (4 train-test season pairs) for 5 football leagues. [2] In addition to the outcome, we include, for each match and each of the two teams that participated in the match, the number of goals scored, shots and shots on target, corners, fouls committed, and yellow and red cards received.

Input variables that are used as predictors in sports are typically count variables or ratios of count variables, in particular ratios of the form $\frac{A}{A+B}$, where A and B are sums of count variables. We assume that the count variables follow time-homogenous independent Poisson distributions. A natural choice for the prior distribution of the rate parameter λ is the Gamma distribution $\lambda \sim Gamma(a_0, b_0)$, which is conjugate. Therefore, for each count variable with mean $\bar{\lambda}_i$ over n_i games, the posterior is again Gamma$(a_0 + \lambda_i n_i, b_0 + n_i)$, where we select weakly informative priors $a_0 = b_0 = 0.001$. A sum of Gamma distributed random variables with the same scale is again Gamma distributed with same scale. Furthermore, if A and B are Gamma distributed random variables with the same scale $A \sim Gamma(\alpha, \theta)$ and $B \sim Gamma(\beta, \theta)$, then $X = \frac{A}{A+B}$ is distributed $X \sim Beta(\alpha, \beta)$. Therefore, a ratio variable derived from Poisson posterior rates of the form $R = \frac{\sum_i \lambda_{Ai}}{\sum_i \lambda_{Ai} + \sum_i \lambda_{Bi}}$ is Beta distributed: $R \sim Beta(\sum_i \lambda_{Ai}, \sum_i \lambda_{Bi})$.

4 Discussion

As anticipated, the proposed model excels at the beginning of each season and the differences between models' prediction errors decrease as the season progresses and input variables become more certain as can be seen in Fig 1. Similar results were observed for basketball, but are omitted for brevity. The HMC variant of the proposed ordinal model was not included in the comparison on football data, because the computation times make it infeasible for practical use. Although the HMC-based approximation yields relatively good predictions, it is discernibly worse than the structural approximation. This can, at least in part, be explained by the inferior accuracy of the HMC-based approximation due to slow mixing (effective sample sizes for β were, on average, $\approx 20\%$ of the total number of iterations and at 1000 iterations, the estimated MCMC sampling error was, on average, approximately $\approx 10\%$ of posterior standard deviation). The structural approximation variant of the proposed model outperforms all other models, as can be seen in Fig. 2, even more convincingly across the football data sets than basketball. Compared to NBA basketball, football seasons are much shorter (in terms of matches per team) and there is more uncertainty in the inputs derived from match statistics.

[2] The data were obtained from http://football-data.co.uk/data.php.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[3] P. McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.

[4] B. Peterson and F. E. Harrell Jr. Partial proportional odds models for ordinal response variables. *Applied statistics*, pages 205–217, 1990.

[5] Stan developement team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*, 2015.

[6] H. O. Stekler, D. Sendor, and R. Verlander. Issues in sports forecasting. *International Journal of Forecasting*, 26(3):606–621, 2010.

[7] E. Štrumbelj and P. Vračar. Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, 28(2):532–542, 2012.