A Bayesian hierarchical model of categorical data rating and classification

Bob Carpenter   Center for Computational Mathematics, Flatiron Institute

Abstract

We introduce a Bayesian model of categorical data rating and classification. Rater effects capture raters' accuracy and bias and item-level effects capture the bias introduced by the items being classified such as difficulty. We show that item-level effects are crucial for ensuring calibrated predictions. We use multivariate priors to capture mean task accuracy, bias, and correlation among responses, which allows sharper and better calibrated predictions for new data raters as might be found in an ongoing data rating task with crowdsourcing. Item-level predictors (aka features) can be used to jointly train a classifier, where no predictors results in a prevalence-only model. We show that training a classifier with a probabilistic data set regularizes estimates and improves the calibration of probabilistic classification.

Keywords: data rating, Bayesian modeling, multivariate priors, classification, item difficulty

Contents

1 Introduction

Supervised training of classifiers is based on labeled data sets. Labeled data typically arises from humans or systems rating the data (also known as "coding" or "annotating" in different literatures). Because human and machine raters are never 100% accurate, the problem arises as to how to deal with disagreements in ratings. For example, given a radiology image, one human rater might say it shows a stage 1 cancer tumor and another may say it is nothing or given a social media post, one rater might say it is positive toward a product and another might say it is neutral. How do we adjudicate these disagreements among raters and get on with building classifiers? One traditional approach is to vote—use multiple raters and take a majority vote. Often this is done in stages, where if two raters disagree, a third is brought in to settle the dispute. This can be problematic when both raters make the same error or when there is disparity in accuracy or correlated bias among the raters. And in the end, we have no measure of certainty. Another traditional approach is to censor data where there is disagreement—that is, use only items in the training data for which the raters agreed. This approach may lead to clean data, but it will not be representative of the "wild type" data from which it was selected.

In order to make the best use of our data, we need to appropriately model it to correct for the bias and accuracy of the raters as well as the difficulty and biases introduced by the items being annotated. In the end, we will be left with a "soft" data set, with probabilities assigned to outcomes for each item. A traditional data set uses a one-hot encoding (where one category has probabity 1 and the others probability 0) and is often derived by assigning the most probable category. We will show in this paper that taking the best category is inferior to selecting a category at random based on the probability distribution, which in turn is inferior to training directly with the probabilistic weights.

Among the contributions of this paper are a new crowdsourcing and classifier training model that introduces (1) item-level effects for difficulty and item bias, (2) multivariate priors on item-level and rater-level effects, and (3) joint classifier and data set training with full Bayesian inference. The model strictly generalizes the model of (**dawid-skene?**). We show how the item-level effects are necessary to achieve calibrated prediction on new data, and how training

a model jointly with full Bayesian inference is preferable to factoring the problem.

2   Data format

2.1   Rating data

We assume there are $K \in \mathbb{N}$ categories into which items are classified, $I \in \mathbb{N}$ items being rated, and $J \in \mathbb{N}$ raters. We assume there are $N \in \mathbb{N}$ ratings, with $y_n \in 1:K$ being the rating given by rater $jj[n]$ for item $ii[n]$. The result is a long-form table of N rows; the first few rows of an example are shown in Table 1.

Table 1: Long-form data format for ratings. Annotation n is for item $ii[n]$ by annotator $jj[n]$, who supplied rating $y[n]$. For example, rating $n = 3$ was made for item $ii[3] = 1$ by rater $jj[3] = 6$, who provided label $y[3] = 6$. Three raters, with ids 1, 2, and 6, rated item $i = 1$ providing labels 4, 4, and 3 respectively.

n	ii	jj	y
1	1	1	4
2	1	2	4
3	1	6	3
4	2	3	1
5	2	4	1
6	3	2	6
⋮	⋮	⋮	⋮

This data format is flexible enough to allow each item to be rated by a zero or more raters. While it is possible to represent a single rater rating the same item multiple times, our models will treat the ratings as independent.

2.2   Item-level predictors

In addition to the ratings, we will assume there are L predictors (features, etc.) for each item. We let $x \in \mathbb{R}^{I \times L}$ be the data matrix, with rows $x_i \in \mathbb{R}^L$ being the L-vector of predictors for item $i \in 1:I$. We model intercepts separately and thus do not assume that there is a column of 1s in the matrix x. Although it would be possible to have rater-level predictors, such as the geographical location or age or sex of the rater, we do not consider that extension in this paper.

3   Data-generating process

We formulate our statistical model generatively in the sense that it is able to generate a complete data set given the items and predictors. We will start with the model for the items then consider a model for the ratings.

3.1 Prevalence and classification

We will assume that each item $i \in 1:I$ has a true category $z_i \in 1:K$. The $z[i]$ are not observed and may be considered missing data and represented by means of a discrete parameter in the model. We model the category based on item-level predictors using a logistic regression, where we assume $\beta \in \mathbb{R}^{L \times K}$ is our matrix of regression coefficients and $\alpha \in \mathbb{R}^K$ is an intercept.

$$z_i \sim \text{categorical}\big(\text{softmax}(\alpha + \beta \cdot x_i^\top)\big),$$

where x_i is the i-th row of the matrix x and $\text{softmax}(u) = \exp(u)/\text{sum}(\exp(u)) \in \Delta^{K-1}$, with $\exp()$ applied elementwise. We will be able to use the fitted model to make predictions for new items not in the training set assuming we have their predictor vectors. That is, the result will be a classifier for new items.

In the case where we have no item-level predictors (i.e., $L = 0$), our model reduces to an intercept-only model where $\text{softmax}(\alpha) \in \Delta^{K-1}$ represents the simple prevalence of the categorical outcomes.

3.2 Generative model for ratings

We will model the sampling distribution for rating using a logistic regression with effects for the item being rated and the rater performing the rating. This section describes the four types of effects we assume on that rating. All of the effects are vector parameters in $\mathbb{R}^K$ (i.e., the size of the number of categories).

3.2.1 Intercept for global bias

In order to model potential biases in ratings that are independent of the category of the item being rated, we will assume there is an intercept term in our logistic regression, $\xi \in \mathbb{R}^K$. A high value for ξ_k means there is an overall bias toward category k whereas a low value represents an overall bias away from category k.

3.2.2 True category effect

The category assigned to item i by a rater is strongly influenced by the true category z_i of that item. We thus assume there is an effect ψ_k based on the true category k of the item being rated. This will contribute a term $\psi_{z_{ii[n]}}$ for the n-th rating, which has category $ii[n]$ and true category $z_{ii[n]}$. Another way to consider the true category effect is as the prior location for the item effects for an item of category k and as a prior location for rater responses to items of true category k.

3.2.3 Rater effects

In order to allow raters to vary in their accuracies and biases, we will model each rater to have their own probabilistic response to items of a given true category. Specifically, we assume that each rater $j \in 1:J$ has a response simplex $\theta_{j,k} \in \Delta^{K-1}$ which says how they respond to items of category k, all else being equal. A perfect rater has $\theta_{j,k,k'}$ equal to 1 if $k = k'$ and 0 otherwise.

That is, $\theta_{j,k,k}$ represents rater j's accuracy on items of category k and the off-diagonal elements of θ_j represent the biases.

3.2.4 Item effects

We will further assume that each item $i \in 1:I$ has a vector of effects $\varphi_i \in \mathbb{R}^K$. If $\varphi_i = 0$, the item has no effect on ratings and raters will just return results according to $\theta_{j,k}$ for items of category k. If $\varphi_{i,z[i]}$ is high, the item is relatively easy to rate, whereas if it's low, the item is difficult to rate, with the other terms determining the response bias.

3.2.5 Sampling distribution for ratings

Given the rating and item-level effects, the generative model for ratings is a multi-logit regression,

$$y_n \sim \text{categorical}\left(\text{softmax}\left(\xi + \psi_{z_{ii[n]}} + \varphi_{ii[n]} + \theta_{jj[n],\, z_{ii[n]}}\right)\right).$$

Breaking this down, item $ii[n] \in 1:I$ is being given a rating of $y_n \in 1:K$ by rater $jj[n] \in 1:J$. The true rating for item $ii[n]$ is $z_{ii[n]} \in 1:K$. The effects being added are vectors in $\mathbb{R}^K$. The softmax() function transforms the unconstrained vector to a simplex, which means the vector components are on the log probability scale. The first effect is the intercept ξ, which accounts for overall bias in response by the raters. The second term $\psi_{z_{ii[n]}}$ is the effect of the true category $zii[n]$. The third term $\varphi_{ii[n]}$ is the effect of the item being rated. The final term $\theta_{jj[n],\, z_{ii[n]}}$ is the effect of rater $jj[n]$ responding to an item whose true category is $z_{ii[n]}$.

4 Priors

4.1 Prior for prevalence regression coefficients

For the prevalence regression, we provide weakly informative priors for the components of the intercept $\alpha \in \mathbb{R}^K$,

$$\alpha_k \sim \text{normal}(0, 3),$$

and the components of the slopes $\beta \in \mathbb{R}^{L \times K}$,

$$\beta_{l,k} \sim \text{normal}(0, 3).$$

4.2 Sum-to-zero constraint for unconstrained simplex parameters

The underlying dimensionality of a simplex is one less than the number of categories it ranges over. In order to match our unconstrained parameterization on the log odds scale to the dimensionality of a simplex, we will constrain it to sum to zero. This removes what would otherwise be an additive non-identifiability in the regression that would allow us to add a constant c to every dimension any of the effects without changing the sampling distribution.

4.3 Prior for global effect

We assign the global intercept a weakly informative prior,

$$\xi \sim \text{normal}(0, 3 \cdot I),$$

where I is the identity matrix.

4.4 Prior for category-level effects

Without any prior knowledge of which categories are likely to be correlated with category k, we will assume a weakly inforamtive prior on the category-level effects,

$$\psi_k \sim \text{normal}(0, 3 \cdot I),$$

for $k \in 1 : K$.

4.5 Prior for item-level effects

The item-level effects $\varphi_i \in \mathbb{R}^K$ are assigned multivariate normal priors centered at zero,

$$\varphi_i \sim \text{normal}(0, \Sigma^\varphi),$$

for $i \in 1 : I$, where Σ^φ is a symmetric, positive-definite covariance matrix parameteter.

4.6 Prior for rater-level effects

The rater-level effects $\theta_{j,k} \in \mathbb{R}^K$ are assigned to a prior conditioned on the true category k,

$$\theta_{j,k} \sim \text{normal}(0, \Sigma^\theta_k),$$

where Σ^θ_k is a positive definite covariance matrix for $k \in 1 : K$.

4.7 Hyperpriors for the covariance of varying effects

We have symmetric, positive-definite covariance parameters Σ^φ and Σ^θ_k for $k \in 1 : K$. We factor covariance matrices into a vector of scales and a correlation matrix,

$$\Sigma^\varphi xtrmdiag(\sigma^\varphi) \cdot \Omega^\varphi \cdot \text{diag}(\sigma^\varphi),$$

for strictly positive $\sigma^\varphi \in \mathbb{R}^K_+$ and a correlation matrix Ω^φ (i.e., a symmetric, positive definite matrix with unit diagonal). We factor the rater-level covariances by category k, taking

$$\Sigma^\theta_k = \text{diag}(\sigma^\theta_k) \cdot \Omega^\theta_k \cdot \text{diag}(\sigma^\theta_k)$$

for $k \in 1 : K$, where $\sigma^\theta_k \in \mathbb{R}^K_+$ is a vector of strictly positive scales and Ω^θ_k is a correlation matrix.

The components of the scale parameters are assigned weakly informative half-normal priors independently by component, taking

$$\sigma^{\theta}_{k,k'}, \sigma^{\varphi}_{k} \sim \mathrm{normal}_{+}(0, 3),$$

for $k, k' \in 1 : K$. We assign Lewandowski-Kurowicka-Joe (LKJ) priors to the correlation matrix,

$$\Omega^{\theta}_{k}, \Omega^{\varphi}_{k} \sim \mathrm{LKJ}(5),$$

for $k \in 1 : K$, where the LKJ density is defined for a symmetric positive-definite, unit-diagonal correlation matrix Ω and shape $\eta > 0$ by

$$\mathrm{LKJ}(\Omega \mid \eta) \propto \det(\Omega)^{\eta - 1}.$$

For $\eta = 1$, the LKJ distribution is uniform over correlation matrices Ω. For $\eta > 1$, it concentrates mass around the unit correlation matrix (with $\eta < 0$ it concentrates toward the boundaries). Thus when used as a prior on a correlation matrix parameter, it has the effect of shrinking the correlation estimates (i.e., the off-diagonal elements of an estimated Ω).

5 Full data and marginal likelihood

Session information

{r session-info} sessionInfo()