

Hierarchical Bayesian Models of Categorical Data Annotation

Bob Carpenter (carp@alias-i.com) *Alias-i, Inc., Brooklyn, NY*

Consider building an automated system for diagnosing caries (a pre-cavity) in dental patients on the basis of X-rays. Standard practice is to build an annotated corpus based on the judgements of one or more domain experts, which are combined to form a deterministic “gold standard” to estimate and evaluate image classifiers. Here’s a contingency table from a study [3] of five dentists:

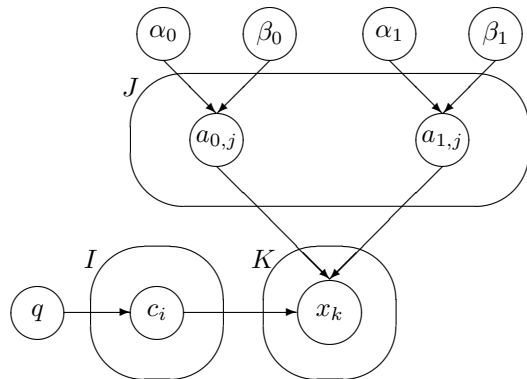
Anno	#	Anno	#	Anno	#	Anno	#	Anno	#	Anno	#	Anno	#
00000	1880	00001	789	00010	43	00011	75	00100	23	00101	63	00110	8
01000	188	01001	191	01010	17	01011	67	01100	15	01101	85	01110	8
10000	22	10001	26	10010	6	10011	14	10100	1	10101	20	10110	2
11000	2	11001	20	11010	6	11011	27	11100	3	11101	72	11110	1
												11111	100

Similar contingency table counts recur in other image-recognition tasks, other diagnostic situations, and in almost all natural-language classification tasks. Coding standards may be vague or ambiguous, coders may misunderstand the standards, and coders are both inconsistent and error prone.

We modestly extend existing epidemiological models [1] to Bayesian reasoning over hierarchical parameters and missing and replicated data. We suggest three applications: One, use category uncertainty to probabilistically supervise [5] and evaluate automatic classifiers; two, evaluate task difficulty and annotator accuracy and agreement with tools other than κ statistics; and three, use posterior category uncertainty to schedule reviews and assign annotator effort for particularly difficult items to ensure reliable gold standards, clear coding standards, and accurate annotators [2,4].

Graphical Model

Suppose our observed annotation data involves J annotators, I items, and K binary (dichotomous) annotations $x_k \in \{0, 1\}$, where annotation k is of item i_k by annotator j_k . This representation explicitly allows missing or replicated data, as would occur by assigning annotation tasks to annotators at random with replacement. Here’s the joint model of the full data annotation process:



$$\begin{aligned}
 c_i &\sim \text{Bernoulli}(q) \\
 a_{0,j} &\sim \text{Beta}(\alpha_0, \beta_0) \\
 a_{1,j} &\sim \text{Beta}(\alpha_1, \beta_1) \\
 x_k &\sim \text{Bernoulli}(c_{i_k} a_{1,j_k} + (1 - c_{i_k})(1 - a_{0,j_k})) \\
 q &\sim \text{Beta}(1, 1) \\
 \alpha_0 / (\alpha_0 + \beta_0) &\sim \text{Beta}(1, 1) \\
 \alpha_0 + \beta_0 &\sim \text{Polynomial}(-5/2) \\
 \alpha_1 / (\alpha_1 + \beta_1) &\sim \text{Beta}(1, 1) \\
 \alpha_1 + \beta_1 &\sim \text{Polynomial}(-5/2)
 \end{aligned}$$

The true categories c_i are distributed as a binomial with prevalence parameter $q \in [0, 1]$ representing the probability of an item having category 1. Prevalence q has a noninformative uniform prior, expressed as a conjugate beta to indicate the ease of incorporating prior knowledge.

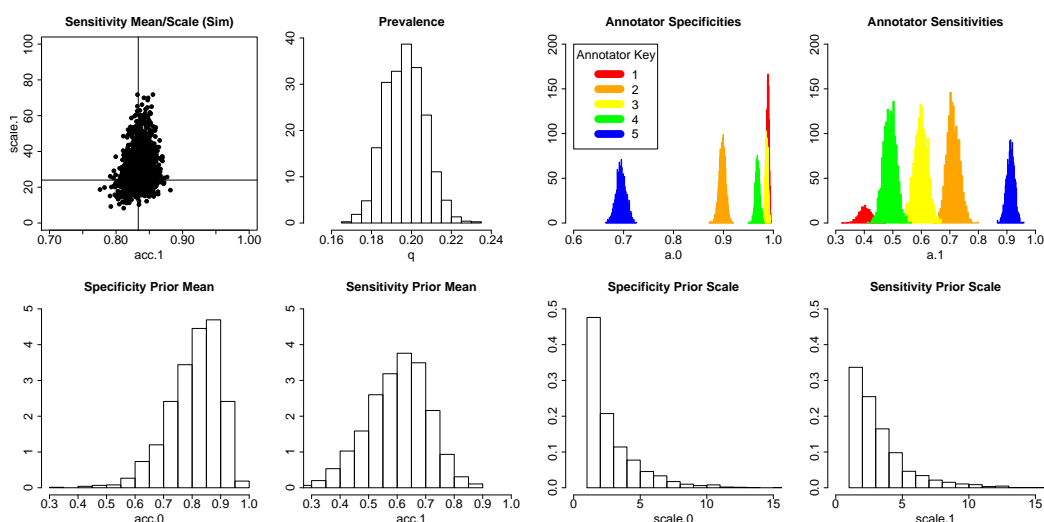
Each annotator has a specificity (accuracy on true category 0 items) of $a_{0,j} \in [0, 1]$ and a sensitivity (accuracy on true category 1 items) of $a_{1,j} \in [0, 1]$. These are generated from independent beta priors with parameters α_0, β_0 and α_1, β_1 , which represent the difficulty of the task. The mean specificity of annotators is $\frac{\alpha_0}{\alpha_0 + \beta_0}$, and the scale $\alpha_0 + \beta_0$ accounts for variance among annotator specificities. The beta parameters have convenient noninformative priors which are uniform on the means $\frac{\alpha}{\alpha + \beta}$ and inverse root scales $(\alpha + \beta)^{-\frac{1}{2}}$. The beta priors on annotator accuracy could also easily be made informative.

An annotation x_k is generated based on the annotator’s specificity or sensitivity given the true category c_{i_k} . If $c_{i_k} = 1$, $x_k \sim \text{Bernoulli}(a_{1,j_k})$, and if $c_{i_k} = 0$, $x_k \sim \text{Bernoulli}(1 - a_{0,j_k})$.

Bayesian Inference

Simulated Data: We first simulated data in R similar to the dentistry data to verify that BUGS could identify the model parameters. With diffuse random initial values for all unknown variables, different chains mixed rapidly (\hat{R} approaching 1 for all variables within 50 samples). Many item-level category assignments were quite uncertain, even with five annotators. The 95% interval for prevalence is (.07, .10) with a simulated prevalence parameter of .075 and sample prevalence of .081. The first two annotator simulated specificities were .74, and .81, with posterior 95% estimates being (.72, .76) and (.72, .81). Due to low prevalence, intervals are wider for sensitivity, with the first two annotators having .68 and .70 simulated sensitivities with 95% intervals of (.60, .73) and (.63, .78).

The main problem is that five annotators are not enough to accurately estimate the group-level beta parameters. The 95% interval for mean specificity is (.56, .85) and for mean sensitivity is (.54, .87), both for simulated values of .83. The 95% intervals on the scales for specificity and sensitivity are (1.4, 65) and (1.1, 20) from simulated values of 48 and 24. With 30 annotators, 1000 items, and prevalence .2, beta parameter intervals are quite tight, as shown in the following scatterplot of sensitivity means and scales, with horizontal and vertical lines indicating the simulated values being estimated:



Dentistry Data: Posterior estimates for all model parameters other than categories are shown above as histograms. Our 95% interval for caries prevalence is (.18, .22). In contrast, a consensus vote estimates prevalence at .026, a majority at .13. Our estimate differs from the voting scheme's because prevalence and annotator accuracy as well as their uncertainty are considered. Annotator 1 (red) and annotator 5 (blue) show opposite biases with median specificity/sensitivity estimates of .99/.40 and .69/.91.

A posterior check shows the model estimates the number of 00000 annotations reasonably with a 95% interval of (1765, 1902) compared to an actual count of 1880. It underestimates 11111 cases, with a 95% interval of (48, 72) compared to an actual 100. Other counts are estimated fairly accurately, which does not occur for models in which all annotators are assumed to have equal accuracies, or where sensitivity is not separated from specificity. The poster will include evaluations of a mixture model that treats some positive items as “easy” [1, 4] analogously to a zero-inflated Poisson.

References

1. Albert, P. S., L. M. McShane, J. H. Shih, and the NCI Bladder Tumor Marker Network. 2001. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biom.* **57**(2):610–619.
2. Bruce, R. F. and J. M. Wiebe. 1999. Recognizing subjectivity: a case study of manual tagging. *NLE* **1**(1):1–16.
3. Espeland, M. A. and S. L. Handelman. 1989. Using latent class models to characterize and assess relative-error in discrete measurements. *Biom.* **45**:587–599.
4. Klebanov, B. B., E. Beigman, and D. Diermeier. 2008. Analyzing disagreements. *COLING 2008*.
5. Smith, P. 1995. Learning with probabilistic supervision. In T. Petsche, ed., *Computational Learning Theory and Natural Learning Systems Volume III*. MIT Press.