Logistic regression with a latent binary variable and noisy labels

September 6, 2018

1 Model

Following [1], consider a set of N training data points $D = \{(\mathbf{x}_1, c_1) \dots (\mathbf{x}_N, c_N)\}$ where $\mathbf{x}_n \in \mathbb{R}^M$ denote M-dimensional real-valued explanatory data (e.g. gene expression levels) and $c_n \in \{0, 1 \dots K-1\}$ denotes a categorical label with K possible values (e.g. clinician-assigned label incorporating some degree of uncertainty).

We aim to fit this with a two-stage model, first regressing the explanatory data $\mathbf{x}_n$ to a latent binary variable representing ground truth $b_n \in \{0,1\}$ (e.g. disease state), then modeling clinical labeling as a categorical variable $c_n|b_n$ that is conditionally-independent of the explanatory data given the ground truth

$$\begin{aligned} P(b=1|\mathbf{x},\mathbf{w}) &= \sigma\left(\mathbf{w}^T\mathbf{x}\right) \\ P(c=k|b=j,\mathbf{z}) &= z_{j,k} \end{aligned}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function, $\mathbf{w} \in \mathbb{R}^M$ are weight parameters for the logistic regression model, and $\mathbf{z}$ are probability parameters for the label observation model.

We put a Laplace double-exponential (Lasso) prior on $\mathbf{w}$, and a uniform[1] Dirichlet prior on each row of $\mathbf{z}$

$$\begin{aligned} P(\mathbf{w}) &\propto \prod_{m=1}^{M} \exp(-|x^{(m)}|) \\ P(\mathbf{z}) &\propto \prod_{j\in\{0,1\}} \delta\left(1 - \sum_{k=0}^{K-1} z_{j,k}\right) \end{aligned}$$

This is equivalent to Section 2.2 of [1], with the sum over j in equation (8) of that paper constrained to $j \in \{0,1\}$ instead of $j \in \{0, 1 \dots K-1\}$. The paper derives a conjugate gradient optimization algorithm, and proves its convergence.

[1]For identifiability of b, we need to break the symmetry of the Dirichlet prior slightly; e.g. by adding a pseudocount of 1 for all $b \to c$ mappings that "agree".

1.1 Quartile approach

An alternate model is to use the interpretation of logistic regression where a latent *continuous-valued* random variable (obtained by adding logistically-distributed noise to $\mathbf{w}^T\mathbf{x}$) is used to obtain the labels (b, c), e.g. with c corresponding to the quartiles.

I haven't pursued this model, as the assumption that c corresponds to quartiles of the latent variable underlying logistic regression seems like a possible misfit to the situation of arbitrarily designated clinical labels (although, conceivably, my assumption that c is independent of $\mathbf{x}$ given b is just as bad, or worse).

2 EM algorithm

How to use the training data D to fit the weights $\mathbf{w}$ and probabilities $\mathbf{z}$? One approach is to use the EM (Expectation Maximization) algorithm [2], treating the binary-valued latent variables $B = \{b_n\}$ as *missing data*, the dataset $D = (X, C)$ as *observed data* (with inputs $X = \{\mathbf{x}_n\}$ and observed labels $C = \{c_n\}$), and the weights and probabilities $\theta = (\mathbf{w}, \mathbf{z})$ as the *parameters* to be fit by the algorithm.

The conjugate gradient parameter optimization approach derived by [1] may well be superior to the EM method. However I've outlined the EM approach here for reference.

The likelihood to be maximized is

$$P(B, C, \theta|X) = P(\mathbf{w})P(\mathbf{z})P(B|\mathbf{w}, X)P(C|\mathbf{z}, B)$$

At the i'th iteration, the parameters found by the EM algorithm are given by maximizing the expected log-likelihood

$$\begin{aligned}
\theta^{(i)} &= \mathrm{argmax}_\theta \; \mathcal{E}\left(\theta || \theta^{(i-1)}\right) \\
\mathcal{E}\left(\theta || \theta^{(i-1)}\right) &= \sum_B P(B|\theta^{(i-1)}, X, C) \log P(B, C, \theta | X) \\
&= \log P(\mathbf{w}) + \log P(\mathbf{z}) + \sum_B P(B|\theta^{(i-1)}, X, C) \left[\log P(B|\mathbf{w}, X) + \log P(C|\mathbf{z}, B)\right] \\
&= \log P(\mathbf{w}) + \log P(\mathbf{z}) \\
&\quad + \sum_n \sum_{j \in \{0,1\}} P(b_n = j|\theta^{(i-1)}, \mathbf{x}_n, c_n) \left[\log P(b_n = j|\mathbf{w}, \mathbf{x}_n) + \log P(c_n|\mathbf{z}, b_n = j)\right] \\
&= \mathcal{E}_{\mathbf{w}} \; + \; \mathcal{E}_{\mathbf{z}} \\
\mathcal{E}_{\mathbf{w}} &= \log P(\mathbf{w}) + \sum_n \left[(1 - \beta_n^{(i-1)}) \log(1 - \sigma\left(\mathbf{w}^T \mathbf{x}_n\right)) + \beta_n^{(i-1)} \log \sigma\left(\mathbf{w}^T \mathbf{x}_n\right)\right] \\
\mathcal{E}_{\mathbf{z}} &= \log P(\mathbf{z}) + \sum_n \left[(1 - \beta_n^{(i-1)}) \log z_{0,c_n} + \beta_n^{(i-1)} \log z_{1,c_n}\right] \\
\beta_n^{(i)} &= P(b_n = 1|\theta^{(i)}, \mathbf{x}_n, c_n) \\
P(b_n = 1|\theta, \mathbf{x}_n, c_n) &= \frac{1}{1 + \frac{P(c_n, b_n = 0|\theta, \mathbf{x}_n)}{P(c_n, b_n = 1|\theta, \mathbf{x}_n)}} \\
&= \frac{1}{1 + \frac{(1 - \sigma(\mathbf{w}^T \mathbf{x}_n)) z_{0,c_n}}{\sigma(\mathbf{w}^T \mathbf{x}_n) z_{1,c_n}}}
\end{aligned}$$

The maximization of $\mathcal{E}_{\mathbf{w}}$ w.r.t. $\mathbf{w}$ is a weighted, Lasso-penalized logistic regression (the weights being the $\beta_n^{(i)}$), for which closed formulae may exist (if not, it may be better to use the conjugate gradient derivations of [1] rather than derive gradients for this more indirect EM approach).

The maximization of $\mathcal{E}_{\mathbf{z}}$ w.r.t. $\mathbf{z}$ should be solvable exactly.

References

[1] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECML PKDD'12, pages 143–158, Berlin, Heidelberg, 2012. Springer-Verlag.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.