Active Learning with Crowdsourced Labels

Matthew Alger
The Australian National University

April 1, 2016

Crowdsourcing provides an active learning domain where many standard active learning assumptions are broken: There is no longer just one labeller, labellers may be non-expert, labellers may not be independent, labellers' accuracy may differ depending on the examples presented, and different labellers may have different accuracies.

Yan et al.[1] introduce a probabilistic model of the crowdsourced active learning problem. Denote examples as $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with $\mathbf{x}_i \in \mathbb{R}^D$, true (unknown) labels as $z_1, \ldots, z_N$, and labels given by the labeller t as $y_1^t, \ldots, y_N^t$. Not all y_i^t are observed and generally no z_i are observed. Denote the $N \times D$ matrix of all examples as X, the $N \times 1$ matrix of all true labels as Z, and the $N \times T$ matrix of all labeller-generated labels as Y (where T is the number of labellers). Then

$$p(Y, Z \mid X) = \prod_i p(z_i \mid \mathbf{x}_i) \prod_{t=1}^{T} p(y_i^t \mid \mathbf{x}_i, z_i).$$

This model makes the label y_i^t dependent on not only the true label z_i but also the specific example $\mathbf{x}_i$. As such, it addresses the problem of labellers' accuracy differing depending on the examples presented as well as differing from each other in general. $p(z_i \mid \mathbf{x}_i)$ models the likelihood (e.g. logistic regression) and $p(y_i^t \mid \mathbf{x}_i, z_i)$ models the labeller. For binary classification, Yan et al. use a Bernoulli model with

$$p(y_i^t \mid \mathbf{x}_i, z_i) = (1 - \eta_t(\mathbf{x}_i))^{|y_i^t - z_i|} \eta_t(\mathbf{x}_i)^{1 - |y_i^t - z_i|}$$

References

[1] Yan Yan, Rómer Rosales, Glenn Fung, Mark W Schmidt, Gerardo H Valadez, Luca Bogoni, Linda Moy, and Jennifer G Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International conference on artificial intelligence and statistics*, pages 932–939, 2010.