Crowd Labelling EM Derivation

Matthew Alger
The Australian National University

August 2, 2016

In this document, I elaborate on the derivation of the expectation-maximisation formulae from Yan et al. [2010]. Notation etc. is taken from Yan et al. [2010]. I only consider the case of binary labels.

1 Formulation

We have N data points $\{\vec{x}_1, \dots, \vec{x}_N\}$, where $\vec{x}_i \in \mathbb{R}^D$. We also have a set of labels $\{y_1^{(1)}, \dots, y_n^{(1)}, \dots, y_1^{(T)}, \dots, y_N^{(T)}\}$, where $y_i^{(t)} \in \mathbb{Z}_2$ is the (potentially incorrect) binary label assigned to $\vec{x}_i$ by annotator t. We want to train a classifier to predict labels of new data points, we want to estimate the groundtruth labels $\{z_1, \dots, z_N\}$ where $z_i \in \mathbb{Z}_2$, and we want to model the quality of each annotator's labels. Let $\vec{x}$, y, and z be random variables representing data points, labels, and groundtruths, respectively. The classification task is then to model $p(z \mid \vec{x})$. Define matrices to represent the data:

$$
\begin{aligned}
X &= [\vec{x}_1^T; \dots; \vec{x}_N^T] \in \mathbb{R}^{N \times D} \\
Y &= [y_1^{(1)}, \dots, y_1^{(T)}; \dots; y_N^{(1)}, \dots, y_N^{(T)}] \in \mathbb{Z}_2^{N \times T} \\
Z &= (z_1, \dots, z_N) \in \mathbb{Z}_2^N
\end{aligned}
$$

We assume that annotator labels depend on both the data point and the groundtruth, that annotator labels have annotator-dependent noise, and that annotator labels follow a Bernoulli distribution:

$$
\begin{aligned}
p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) &= (1 - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t))^{|y_i^{(t)} - z_i|} \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t)^{1 - |y_i^{(t)} - z_i|} \\
\eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) &= \sigma(\vec{w}_t^T \vec{x}_i - \gamma_t)
\end{aligned}
$$

We use logistic regression to model the posterior distribution:

$$
p(z_i = 1 \mid \vec{x}_i, \alpha, \beta) = \sigma(\vec{\alpha}^T \vec{x}_i + \beta)
$$

The parameters of the model are $\vec{\theta} = \{\vec{\alpha}, \beta, \vec{w}_1, \dots, \vec{w}_T, \gamma_1, \dots, \gamma_T\}$.

2 Expectation-Maximisation

We can find optimum values of the parameters by maximising the log-likelihood $p(Y \mid X, \vec{\theta})$, i.e.

$$
\begin{aligned}
\vec{\theta}^\star &= \underset{\vec{\theta}}{\operatorname{argmax}} \sum_{i=1}^N \sum_{t=1}^T \log p(y_i^{(t)} \mid \vec{x}_i, \vec{\theta}) \\
&= \underset{\vec{\theta}}{\operatorname{argmax}} \sum_{i=1}^N \sum_{t=1}^T \log \sum_{z_i=0}^1 p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta})
\end{aligned}
$$

noting that we assume independence between labels of different data points and labels from different annotators. Since z_i are latent variables, we must use expectation-maximisation.

For the expectation step, we want to evaluate $p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta})$ for $i = 1, \ldots, N$. We can write this in terms of the parameters:

$$\begin{aligned}
p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta}) &= \frac{1}{A_i} p(z_i, y_i^{(1)}, \ldots, y_i^{(T)} \mid \vec{x}_i, \vec{\theta}) \\
&= \frac{1}{A_i} \prod_{t=1}^{T} p(z_i, y_i^{(t)} \mid \vec{x}_i, \vec{\theta}) \\
&= \frac{1}{A_i} \prod_{t=1}^{T} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{\theta}) p(z_i \mid \vec{x}_i, \vec{\theta}) \\
&= \frac{1}{A_i} \prod_{t=1}^{T} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)
\end{aligned}$$

A_i is a normalisation term given by

$$A_i = \sum_{z_i=0}^{1} p(z_i, y_i^{(1)}, \ldots, y_i^{(T)} \mid \vec{x}_i, \vec{\theta}).$$

To simplify notation, let $\tilde{p}(z_i) = p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta})$.

For the maximisation step, we want to set

$$\vec{\theta}^{\text{new}} = \underset{\vec{\theta}^{\text{new}}}{\text{argmax}} \sum_{i=1}^{N} Q_i(\vec{\theta}^{\text{new}}, \vec{\theta})$$

where

$$Q_i(\vec{\theta}^{\text{new}}, \vec{\theta}) = \sum_{z_i=0}^{1} p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta}) \log p(\vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, z_i \mid \vec{\theta}^{\text{new}}).$$

Once again, we need to write this in terms of the parameters.

$$\begin{aligned}
Q_i(\vec{\theta}^{\text{new}}, \vec{\theta}) &= \sum_{z_i=0}^{1} \tilde{p}(z_i) \log p(\vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, z_i \mid \vec{\theta}^{\text{new}}) \\
&= \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log p(\vec{x}_i, y_i^{(t)}, z_i \mid \vec{\theta}^{\text{new}}) \\
&= \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log(p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}}) p(\vec{x}_i \mid \vec{\theta}^{\text{new}})) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}}) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \\
&\qquad \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) (\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{\theta}^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}})) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \\
&\qquad \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) (\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t^{\text{new}}, \gamma_t^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}^{\text{new}}, \beta^{\text{new}}))
\end{aligned}$$

Then the maximisation step is

$$\vec{\theta}^{\text{new}} = \underset{\vec{\theta}^{\text{new}}}{\text{argmax}} \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) (\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t^{\text{new}}, \gamma_t^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}^{\text{new}}, \beta^{\text{new}}))$$

noting that $T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}})$ is the same for all $\vec{\theta}^{\text{new}}$ as x_i is observed. To simplify notation, let

$$f(\vec{\theta}) = \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i)(\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)).$$

where $\tilde{p}(z_i)$ is evaluated using the old value of $\vec{\theta}$.

3 Gradients of the Optimisation Target

In this section, I derive the gradients of f with respect to the parameters $\vec{\theta}$.

3.1 $\nabla_{\vec{\alpha}} f(\vec{\theta})$

$$\begin{aligned}
\nabla_{\vec{\alpha}} f(\vec{\theta}) &= T \sum_{i=1}^{N} \sum_{z_i=0}^{1} \nabla_{\vec{\alpha}} (\tilde{p}(z_i) \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)) \\
&= T \sum_{i=1}^{N} \sum_{z_i=0}^{1} \tilde{p}(z_i) \nabla_{\vec{\alpha}} \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta) \\
&= T \sum_{i=1}^{N} (\tilde{p}(z_i = 1) - \tilde{p}(z_i = 0)) \nabla_{\vec{\alpha}} \log \sigma(\vec{\alpha}^T \vec{x}_i + \beta) \\
&= T \sum_{i=1}^{N} (\tilde{p}(z_i = 1) - \tilde{p}(z_i = 0)) \frac{\nabla_{\vec{\alpha}} \sigma(\vec{\alpha}^T \vec{x}_i + \beta)}{\sigma(\vec{\alpha}^T \vec{x}_i + \beta)} \\
&= T \sum_{i=1}^{N} (\tilde{p}(z_i = 1) - \tilde{p}(z_i = 0)) \frac{\sigma(\vec{\alpha}^T \vec{x}_i + \beta)(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta))}{\sigma(\vec{\alpha}^T \vec{x}_i + \beta)} \vec{x} \\
&= T \sum_{i=1}^{N} (\tilde{p}(z_i = 1) - \tilde{p}(z_i = 0))(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)) \vec{x}
\end{aligned}$$

3.2 $\frac{\partial f}{\partial \beta}(\vec{\theta})$

The derivative is mostly identical to the derivative for $\nabla_{\vec{\alpha}} f(\vec{\theta})$, but with $\frac{\partial}{\partial \beta}(\vec{\alpha}^T \vec{x}_i + \beta) = 1$.

$$\frac{\partial f}{\partial \beta}(\vec{\theta}) = T \sum_{i=1}^{N} (\tilde{p}(z_i = 1) - \tilde{p}(z_i = 0))(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta))$$

References

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.