

EM cont'd: Semi-supervised learning

Semi-supervised learning is learning when you have both labeled and unlabeled data. The Nigam-McCallum-Mitchell approach is standard for text classification in a naive Bayes model.

Bag of words for document classification: multinomial. Vocabulary of K words, let x_k be the count of how many times word k is in the document. Then using parameters $\theta_{1:K} \geq 0$, $\sum_k \theta_k = 1$, we have:

$$p(x | \theta) \propto \prod_k \theta_k^{x_k}$$

1. For each data point $x_{1:N}$,
 - (a) Choose a class y by π_y
 - (b) Generate x_n by a multinomial with parameter θ_y .

This gives data probability:

$$p(x_{1:N}, y_{1:N} | \theta_{1:L}, \pi) \propto \prod_n \prod_l \left(\pi_l \prod_k \theta_{lk}^{x_{nk}} \right)^{y_{nl}}$$

Log likelihood is:

$$\ell(x_{1:N}, y_{1:N} | \theta_{1:L}, \pi) = \sum_n \sum_l y_{nl} \left(\log \pi_l + \sum_k x_{nk} \log \theta_{lk} \right)$$

We can obtain an analytical solution for θ, π by:

$$\begin{aligned} \hat{\pi}_l &= \frac{\# \text{ examples with label } l}{\text{total } \# \text{ examples}} \\ &= \frac{\sum_n y_{nl}}{N} \\ \hat{\theta}_{lk} &= \frac{\# \text{ times word } k \text{ occurs in examples with label } l}{\# \text{ examples with label } l} \\ &= \frac{\sum_n y_{nl} x_{nk}}{\sum_n y_{nl}} \end{aligned}$$

Now, the semi-supervised case. We have N labeled points, $x_{1:N}, y_{1:N}$ and U unlabeled points, $x_{N+1:N+U}$. We can write down the data probability exactly as before, except we have to introduce *hidden variables* corresponding to the unknown labels. For consistency with previous lectures (but not the paper), we'll call these $z_{N+1:N+U}$ (the paper just calls them y). We write down the data probability as:

$$p(x_{1:N+U}, y_{1:N} \mid \theta_{1:L}, \pi) \propto \left[\prod_{n=1}^N \prod_l \left(\pi_l \prod_k \theta_{lk}^{x_{nk}} \right)^{y_{nl}} \right] \left[\prod_{n=N+1}^{N+U} \sum_{z_n} \prod_l \left(\pi_l \prod_k \theta_{lk}^{x_{nk}} \right)^{z_{nl}} \right]$$

Taking logs, we get:

$$\sum_{n=1}^N \sum_l y_{nl} \left(\log \pi_l + \sum_k x_{nk} \log \theta_{lk} \right) + \sum_{n=N+1}^{N+U} \log \sum_{z_n} \prod_l \left(\pi_l \prod_k \theta_{lk}^{x_{nk}} \right)^{z_{nl}}$$

And now we're stuck with a $\log \sum$ again, so we do EM! Here, $\pi_l^{z_{nl}}$ is playing the role of λ in Jensen, and the multinomial term is playing the role of f . Writing h_{nl} for the expectation of z_{nl} , we get the lower bound of the form:

$$\begin{aligned} & \sum_{n=1}^N \sum_l y_{nl} \left(\log \pi_l + \sum_k x_{nk} \log \theta_{lk} \right) + \sum_{n=N+1}^{N+U} \sum_l h_{nl} \left(\log \pi_l + \sum_k x_{nk} \log \theta_{lk} \right) \\ &= \sum_{n=1}^{N+U} \sum_l (y_{nl} + h_{nl}) \left(\log \pi_l + \sum_k x_{nk} \log \theta_{lk} \right) \end{aligned}$$

Here, I implicitly assume that for $n \leq N$ we set $h_n = 0$ and for all $n > N$, we set $y_n = 0$.

So the two things we need to do are:

1. Compute h_{nl} as the posterior probability that example $n > N$ is in class l .
2. Maximize the above with respect to π, θ .

The maximization term is trivial: it's exactly the same as the supervised case, except we replace some of the y s with h s. This gives:

$$\begin{aligned} \hat{\pi}_l &= \frac{\sum_n (y_{nl} + h_{nl})}{N + U} \\ \hat{\theta}_{lk} &= \frac{\sum_n (y_{nl} + h_{nl}) x_{nk}}{\sum_n (y_{nl} + h_{nl})} \end{aligned}$$

The expectation step for example n is as follows:

$$\begin{aligned} h_{nl} &= p(y_n = l \mid x_n, \theta, \pi) \\ &= \frac{1}{Z} p(y_n = l, x_n \mid \theta, \pi) \\ &= \frac{1}{Z} \pi_l \prod_k \theta_{lk}^{x_{nk}} \end{aligned}$$

Where Z is obtained by summing h_n and normalizing.

Making semi-supervised learning work is sometimes hard. The NMM paper describes several ideas, the most important of which is *annealing*. Essentially, we incrementally let the unlabeled data have more and more influence, but we don't want it to "swamp" the labeled data.

Standard learning curves show that the semi-supervised performance begins about the same as fully-supervised, then increases faster, and then rejoins the curve. So there's usually a "sweet spot" in terms of data (both labeled and unlabeled).