

# CS838-1 Advanced NLP: The EM Algorithm

Xiaojin Zhu

2007

Send comments to jerryzhu@cs.wisc.edu

“Nice intuitions have nice mathematical explanations.”

“Not every iterative procedure can be called EM.”

## 1 Naive Bayes Revisited

Recall in Naive Bayes models, we are given a training set  $\{(x, y)_{1:n}\}$ , and our goal is to train a classifier that classifies any new document  $x$  into one of  $K$  classes. In the case of MLE, this is achieved by estimating parameters  $\Theta = \{\pi, \theta_1, \dots, \theta_K\}$ , where  $p(y = k) = \pi_k$ , and  $p(x|y = k) = \theta_k$ , to maximize the joint log likelihood of the training data:

$$\Theta = \arg \max_{\pi, \theta_1, \dots, \theta_K} \log p(\{(x, y)_{1:n}\} | \pi, \theta_1, \dots, \theta_K). \quad (1)$$

The solution is

$$\begin{aligned} \pi_k &= \frac{\sum_{i=1}^n [y_i = k]}{n}, \quad k = 1, \dots, K \\ \theta_{kw} &= \frac{\sum_{i: y_i = k} x_{iw}}{\sum_{i: y_i = k} \sum_{u=1}^V x_{iu}}, \quad k = 1, \dots, K. \end{aligned} \quad (2)$$

Note  $\pi$  is a probability vector of length  $K$  over classes, and each  $\theta_k$  is a probability vector of length  $V$  over the vocabulary.

Classification is done by computing  $p(y|x)$ :

$$\hat{y} = \arg \max_k p(y = k|x) \quad (3)$$

$$= \arg \max_k p(y = k)p(x|y = k) \quad ; \text{ Bayes rule, ignore constant denominator} \quad (4)$$

$$= \arg \max_k \pi_k \prod_{w=1}^V \theta_{kw}^{x_w} \quad (5)$$

$$= \arg \max_k \log \pi_k + \sum_{w=1}^V x_w \log \theta_{kw} \quad ; \text{ log is monotonic.} \quad (6)$$

## 2 K-means Clustering

So far so good. What if the training data is *mostly unlabeled*? For example, let's assume we have only one labeled example per class:

$$\{(x_1, y_1 = 1), (x_2, y_2 = 2), \dots, (x_K, y_K = K), x_{K+1}, \dots, x_n\}.$$

We can simply ignore unlabeled data  $\{x_{K+1}, \dots, x_n\}$  and train our Naive Bayes classifier on labeled data. Can we do better?

Here is an iterative procedure, known as *K-means clustering*, which we apply to our classification task:

1. Estimate  $\Theta^{(t=0)}$  from labeled examples only.
2. Repeat until things no longer change:
  - (a) Classify all examples (labeled and unlabeled)  $x$  by its most likely class under the current model:  $\hat{y} = \arg \max_k p(y = k|x, \Theta^{(t)})$ .
  - (b) Now we have a fully labeled dataset  $\{(x, \hat{y})_{1:n}\}$ . Retrain  $\Theta^{(t+1)}$  on it. Let  $t = t + 1$ .

There are a couple details:

- We re-classify all the labeled points. This is mostly for notational convenience. It is certainly fine (and probably more desirable) to fix their labels during the iterations. The derivation follows similarly, with separate terms for the labeled data.
- We use the K-means clustering algorithm for classification. But it was originally designed for clustering. For example, one can randomly pick  $\theta_{1:K}$  to start with, and the algorithm will converge to  $K$  final clusters. In that case, we do not have the correspondence between clusters and classes.
- K-means clustering is usually presented on mixture of Gaussian distributions. Here we instead have mixture of multinomial distributions. In either case,  $\theta_k$  is the ‘centroid’ of cluster  $k$ . The distance is measured differently though [1].

## 3 The EM Algorithm

In K-means we made a *hard* classification: each  $x$  is assigned a unique label  $\hat{y}$ . A *soft* version would be to use the posterior  $p(y|x)$ . Intuitively, each  $x_i$  is split into  $K$  copies, but copy  $k$  has weight  $\gamma_{ik} = p(y_i = k|x_i)$  instead of one. We now have a dataset with fractional counts, but that poses no difficulty in retraining the parameters. One can show that the MLE is the weighted version of (2)

$$\begin{aligned} \pi_k &= \frac{\sum_{i=1}^n \gamma_{ik}}{n}, \quad k = 1, \dots, K \\ \theta_{kw} &= \frac{\sum_{i=1}^n \gamma_{ik} x_{iw}}{\sum_{i=1}^n \sum_{u=1}^V \gamma_{ik} x_{iu}}, \quad k = 1, \dots, K. \end{aligned} \tag{7}$$

The change from hard to soft leads us to the EM (Expectation Maximization) algorithm:

1. Estimate  $\Theta^{(t=0)}$  from labeled examples only.
2. Repeat until convergence:
  - (a) **E-step:** For  $i = 1 \dots n, k = 1 \dots K$ , compute  $\gamma_{ik} = p(y_i = k | x_i, \Theta^{(t)})$ .
  - (b) **M-step:** Compute  $\Theta^{(t+1)}$  from (7). Let  $t = t + 1$ .

## 4 Analysis of EM

EM might look like a heuristic method. However, as we show now, it has a rigorous foundation: EM is guaranteed to find a local optimum of data log likelihood.

Recall if we have complete data  $\{(x, y)_{1:n}\}$  (as in standard Naive Bayes), the joint log likelihood under parameters  $\Theta$  is

$$\log p((x, y)_{1:n} | \Theta) = \sum_{i=1}^n \log p(y_i | \Theta) p(x_i | y_i, \Theta). \quad (8)$$

However, now  $y_{1:n}$  are hidden variables. We instead maximize the *marginal* log likelihood

$$\ell(\Theta) = \log p(x_{1:n} | \Theta) \quad (9)$$

$$= \sum_{i=1}^n \log p(x_i | \Theta) \quad (10)$$

$$= \sum_{i=1}^n \log \sum_{y=1}^K p(x_i, y | \Theta) \quad (11)$$

$$= \sum_{i=1}^n \log \sum_{y=1}^K p(y | \Theta) p(x_i | y, \Theta). \quad (12)$$

We note that there is a summation *inside* the log. This couples the  $\Theta$  parameters. If we try to maximize the marginal log likelihood by setting the gradient to zero, we will find that there is no longer a nice closed form solution, unlike the joint log likelihood with complete data. The reader is encouraged to attempt this to see the difference.

EM is an iterative procedure to maximize the marginal log likelihood  $\ell(\Theta)$ . It constructs a concave, easy-to-optimize lower bound  $Q(\Theta, \Theta^{(t)})$ , where  $\Theta$  is the variable and  $\Theta^{(t)}$  is the previous, fixed, parameter. The lower bound has an interesting property  $Q(\Theta^{(t)}, \Theta^{(t)}) = \ell(\Theta^{(t)})$ . Therefore the new parameter  $\Theta^{(t+1)}$  that maximizes  $Q$  is guaranteed to have  $Q \geq \ell(\Theta^{(t)})$ . Since  $Q$  lower bounds  $\ell$ , we have  $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$ .

The lower bound is obtained via *Jensen's inequality*

$$\log \sum_i p_i f_i \geq \sum_i p_i \log f_i, \quad (13)$$

which holds if the  $p_i$ 's form a probability distribution (i.e., non-negative and sum to 1). This follows from the concavity of  $\log$ .

$$\ell(\Theta) = \sum_{i=1}^n \log \sum_{y=1}^K p(x_i, y|\Theta) \quad (14)$$

$$= \sum_{i=1}^n \log \sum_{y=1}^K p(y|x_i, \Theta^{(t)}) \frac{p(x_i, y|\Theta)}{p(y|x_i, \Theta^{(t)})} \quad (15)$$

$$\geq \sum_{i=1}^n \sum_{y=1}^K p(y|x_i, \Theta^{(t)}) \log \frac{p(x_i, y|\Theta)}{p(y|x_i, \Theta^{(t)})} \quad (16)$$

$$\equiv Q(\Theta, \Theta^{(t)}). \quad (17)$$

Note we introduced a probability distribution  $p(y|x_i, \Theta^{(t)}) \equiv \gamma_{iy}$  separately for each example  $x_i$ . This is what E-step is computing.

The M-step maximizes the lower bound  $Q(\Theta, \Theta^{(t)})$ . It is worth noting that now we can set the gradient of  $Q$  to zero and obtain a closed form solution. In fact the solution is simply (7), and we call it  $\Theta^{(t+1)}$ .

It is easy to see that

$$Q(\Theta^{(t)}, \Theta^{(t)}) = \sum_{i=1}^n \log p(x_i|\Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (18)$$

Since  $\Theta^{(t+1)}$  maximizes  $Q$ , we have

$$Q(\Theta^{(t+1)}, \Theta^{(t)}) \geq Q(\Theta^{(t)}, \Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (19)$$

On the other hand,  $Q$  lower bounds  $\ell$ . Therefore

$$\ell(\Theta^{(t+1)}) \geq Q(\Theta^{(t+1)}, \Theta^{(t)}) \geq Q(\Theta^{(t)}, \Theta^{(t)}) = \ell(\Theta^{(t)}). \quad (20)$$

This shows that  $\Theta^{(t+1)}$  is indeed a better (or no worse) parameter than  $\Theta^{(t)}$  in terms of the marginal log likelihood  $\ell$ . By iterating, we arrive at a local maximum of  $\ell$ .

## 5 Deeper Analysis of EM

You might have noticed that we never referred to the concrete model  $p(x|y), p(y)$  (Naive Bayes) in the above analysis, except when we say the solution is simply (7). Does this suggest that EM is more general than Naive Bayes? Besides, where did the particular probability distribution  $p(y|x_i, \Theta^{(t)})$  come from in (15)?

The answer to the first question is yes. EM applies to joint probability models where some random variables are missing. It is advantageous when the marginal is hard to optimize, but the joint is. To be general, consider a joint distribution  $p(X, Z|\Theta)$ , where  $X$  is the collection of observed variables, and  $Z$  unobserved variables. The quantity we want to maximize is the marginal log likelihood

$$\ell(\Theta) \equiv \log p(X|\Theta) = \log \sum_Z p(X, Z|\Theta), \quad (21)$$

which we assume to be difficult. One can introduce an arbitrary distribution over hidden variables  $q(Z)$ ,

$$\ell(\Theta) = \sum_Z q(Z) \log P(X|\Theta) \quad (22)$$

$$= \sum_Z q(Z) \log \frac{P(X|\Theta)q(Z)P(X, Z|\Theta)}{P(X, Z|\Theta)q(Z)} \quad (23)$$

$$= \sum_Z q(Z) \log \frac{P(X, Z|\Theta)}{q(Z)} + \sum_Z q(Z) \log \frac{P(X|\Theta)q(Z)}{P(X, Z|\Theta)} \quad (24)$$

$$= \sum_Z q(Z) \log \frac{P(X, Z|\Theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{P(Z|X, \Theta)} \quad (25)$$

$$= F(\Theta, q) + KL(q(Z)||p(Z|X, \Theta)). \quad (26)$$

Note  $F(\Theta, q)$  is the RHS of Jensen's inequality. Since  $KL \geq 0$ ,  $F(\Theta, q)$  is a lower bound of  $\ell(\Theta)$ .

First consider the maximization of  $F$  on  $q$  with  $\Theta^{(t)}$  fixed.  $F(\Theta^{(t)}, q)$  is maximized by  $q(Z) = p(Z|X, \Theta^{(t)})$  since  $\ell(\Theta)$  is fixed and  $KL$  attains its minimum zero. This is why we picked the particular distribution  $p(Z|X, \Theta^{(t)})$ . This is the E-step.

Next consider the maximization of  $F$  on  $\Theta$  with  $q$  fixed as above. Note in this case  $F(\Theta, q) = Q(\Theta, \Theta^{(t)})$ . This is the M-step.

Therefore the EM algorithm can be viewed as coordinate ascent on  $q$  and  $\Theta$  to maximize  $F$ , a lower bound of  $\ell$ .

Viewed this way, EM is a particular optimization method. There are several variations of EM:

- Generalized EM (GEM) finds  $\Theta$  that improves, but not necessarily maximizes,  $F(\Theta, q) = Q(\Theta, \Theta^{(t)})$  in the M-step. This is useful when the exact M-step is difficult to carry out. Since this is still coordinate ascent, GEM can find a local optimum.
- Stochastic EM: The E-step is computed with Monte Carlo sampling. This introduces randomness into the optimization, but asymptotically it will converge to a local optimum.
- Variational EM:  $q(Z)$  is restricted to some easy-to-compute subset of distributions, for example the fully factorized distributions  $q(Z) = \prod_i q(z_i)$ .

In general  $p(Z|X, \Theta^{(t)})$ , which might be intractable to compute, will not be in this subset. There is no longer guarantee that variational EM will find a local optimum.

## References

- [1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.