CS769 Spring 2010 Advanced Natural Language Processing

The EM Algorithm

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

Given labeled examples $(x_1, y_1), \ldots, (x_l, y_l)$, one can build a classifier. If in addition one has many unlabeled examples $x_{l+1}, \ldots, x_n$, can one use both labeled and unlabeled examples to build a better classifier? This setting is known as *semi-supervised learning*. There are many algorithms for semi-supervised learning, the EM algorithm is one of them. It applies when you have a generative model, but part of the data is missing.

1 Fully Labeled Data: Naive Bayes Revisited

Recall in Naive Bayes models, we are given a training set $(x_1, y_1), \ldots, (x_n, y_n)$, 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 $\Theta = \{\pi, \theta_1, \ldots, \theta_K\}$, where $p(y = k) = \pi_k$, and $p(x|y = k) = \theta_k$, to maximize the joint log likelihood of the training data:

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

The solution is

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

Note π is a probability vector of length K over classes, and each θ_k is a probability vector of length V over the vocabulary.

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

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

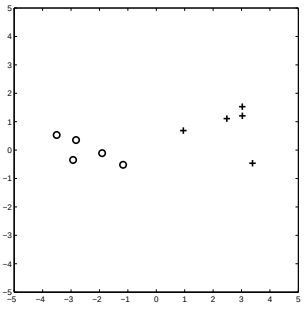
$$= \arg \max_k p(y = k)p(x|y = k) \quad ; \text{Bayes rule, ignore constant} \tag{4}$$

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

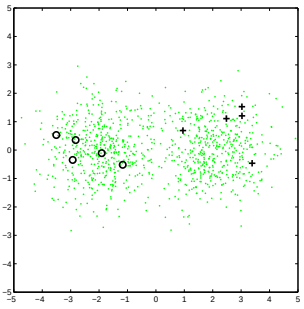
$$= \arg \max_k \log \pi_k + \sum_{w=1}^V x_w \log \theta_{kw} \quad ; \log \text{ is monotonic.} \tag{6}$$

2 When Some or All Labels are Missing: The EM Algorithm

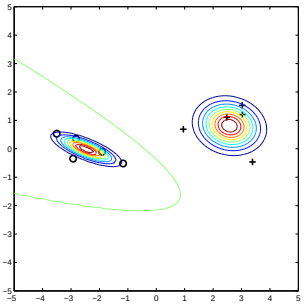
We do not know the true label on unlabeled points. However, we can start from some initial Θ, for example those trained on labeled data only. We can then assign an unlabeled point x_i fractional class labels $p(y_i = k|x_i, \Theta)$ for $k = 1 \ldots K$. Intuitively, each x_i is split into K copies, but copy k has weight $\gamma_{ik} = p(y_i = k|x_i, \Theta)$



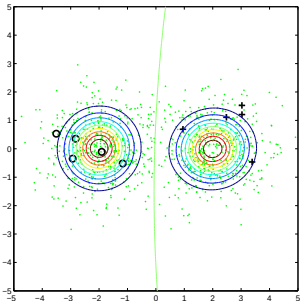
(a) labeled data



(b) labeled and unlabeled data (small dots)



(c) model learned from labeled data



(d) model learned from labeled and unlabeled data

Figure 1: In a binary classification problem, if we assume each class has a Gaussian distribution, then we can use unlabeled data to help parameter estimation.

instead of one. We now have a dataset with fractional counts, but this is not a problem. One can show that the new MLE for Θ is a weighted version of (2)

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

This is the EM (Expectation Maximization) algorithm on a mixture model of multinomial distributions with 2 components:

1. Estimate $\Theta^{(t=0)}$ from labeled examples only. It should be point out that EM is sensitive to the initial parameter $\Theta^{(0)}$. One better strategy is random-restart, by trying multiple different initial parameters.
2. Repeat until convergence:
   (a) **E-step**[1]: For $i = 1 \ldots n, k = 1 \ldots K$, compute $\gamma_{ik} = p(y_i = k | x_i, \Theta^{(t)})$.
   (b) **M-step**: Compute $\Theta^{(t+1)}$ from (7). Let $t = t + 1$.

3 EM Optimizes a Lower Bound of the Log Likelihood

EM might look like a heuristic method. However, it is not – 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 Θ is

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

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

$$\begin{aligned} \ell(\Theta) &= \log p(x_{1:n} | \Theta) && (9) \\ &= \sum_{i=1}^n \log p(x_i | \Theta) && (10) \\ &= \sum_{i=1}^n \log \sum_{y=1}^K p(x_i, y | \Theta) && (11) \\ &= \sum_{i=1}^n \log \sum_{y=1}^K p(y | \Theta) p(x_i | y, \Theta). && (12) \end{aligned}$$

We note that there is a summation *inside* the log. This couples the Θ 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 Θ 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 ℓ, we have $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$.

[1]For simplicity, we adopt the version where one may change the label on labeled points. It is also possible to "pin down" those labels. The overall log likelihood would then be a sum of log marginal probability $\log p(x)$ on unlabeled points, plus log joint probability $\log p(x, y)$ on labeled points.

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

$$\log \sum_i p_i f_i \geq \sum_i p_i \log f_i, \tag{13}$$

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.

$$\begin{aligned}
\ell(\Theta) &= \sum_{i=1}^{n} \log \sum_{y=1}^{K} p(x_i, y|\Theta) && (14)\\
&= \sum_{i=1}^{n} \log \sum_{y=1}^{K} p(y|x_i, \Theta^{(t)}) \frac{p(x_i, y|\Theta)}{p(y|x_i, \Theta^{(t)})} && (15)\\
&\geq \sum_{i=1}^{n} \sum_{y=1}^{K} p(y|x_i, \Theta^{(t)}) \log \frac{p(x_i, y|\Theta)}{p(y|x_i, \Theta^{(t)})} && (16)\\
&\equiv Q(\Theta, \Theta^{(t)}). && (17)
\end{aligned}$$

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

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

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

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

On the other hand, Q lower bounds ℓ. Therefore

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

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

4 A More General View of EM

You might have noticed that we never referred to the specific model $p(x|y), p(y)$ (in this case Naive Bayes) in the above analysis, except when we find the solution (7). EM is general and applies to joint probability models whenever some random variables are missing. *EM is advantageous when the marginal is difficult 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

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

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

$$\begin{aligned}
\ell(\Theta) &= \sum_Z q(Z) \log P(X|\Theta) && (22)\\
&= \sum_Z q(Z) \log \frac{P(X|\Theta)q(Z)P(X,Z|\Theta)}{P(X,Z|\Theta)q(Z)} && (23)\\
&= \sum_Z q(Z) \log \frac{P(X,Z|\Theta)}{q(Z)} + \sum_Z q(Z) \log \frac{P(X|\Theta)q(Z)}{P(X,Z|\Theta)} && (24)\\
&= \sum_Z q(Z) \log \frac{P(X,Z|\Theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{P(Z|X,\Theta)} && (25)\\
&= F(\Theta, q) + KL(q(Z)\|p(Z|X,\Theta)). && (26)
\end{aligned}$$

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 attends 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 Θ 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 Θ to maximize F, a lower bound of ℓ.

There are several variations of EM:

- Generalized EM (GEM) finds Θ 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.