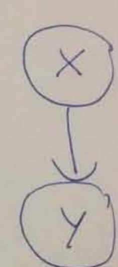
6.438 : Recitation - 12

EM-algorithm:

I. **Problem Setting:** We have some hidden variables which we call X (can be a vector), and some observed variables which we call Y.

G:



We have a known graph structure for variables X and Y, such that the joint distribution of X and Y factorizes according to the graph. Thus the joint probability of X and Y can be written as a function of the (unknown) graph parameters θ.

① $P(X,Y) = P(X,Y;\theta)$

For simplicity, we look at the special case where $\theta = (\theta_X ; \theta_{Y|X})$, so that the joint probability factorizes as:

② $P_{X,Y}(x,y;\theta) = P_X(x;\theta_X) \cdot P_{Y|X}(y|x;\theta_{Y|X})$

In other words, the variables X and Y satisfy the graph structure of G. Note that ~~many~~ many popular models, such as Naive Bayes and HMMs, satisfy this assumption.

II. <u>Objective</u>: We are given a large number of samples of the Y variable, which we represent as $\{y^{(i)}\}$, $i \in \{1, 2, \ldots S\}$.

These samples provide us with a ~~chara~~ characterization of the marginal distribution of Y i.e. $P_Y(\cdot)$.

Our aim is to find parameters $\hat{\theta} = (\hat{\theta}_X, \hat{\theta}_{Y|X})$ which produce a marginal distribution for Y close to what we obtain from samples.

More concretely, we look at the following ML objective: find the value $\hat{\theta}$ of θ which maximizes the probability of generating the observed samples. That is,

$$(3) \quad \hat{\theta}_{ML} \triangleq \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{S} P_Y(y^{(i)}; \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{S} \left(\sum_{x} P_{X,Y}(x, y^{(i)}; \theta) \right)$$

$$(4) \quad \therefore \hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{S} \left(\sum_{x} P_X(x; \theta_x) \cdot P_{Y|X}(y^{(i)} | x ; \theta_{Y|X}) \right)$$

It should be noted that (4) is a difficult problem to solve in general. We look at one way of solving it approximately viz. EM-algorithm

III. EM Algorithm:

The EM algorithm is a meta-algorithm for estimating graphical model parameters from partial observations viz solving problem (4). Since the equations are already derived in the lecture notes, we will go over them quickly.

Consider problem (4), which we write in a slightly shorter form.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{S} \left(\sum_{x} P_{X,Y}(x, y^{(i)}; \theta_{X,Y}) \right)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{S} \log \left(\sum_{x} P_{X,Y}(x, y^{(i)}; \theta_{X,Y}) \right)$$

(Taking Logarithms, since $\log(\cdot)$ is a monotone function)

Now consider any set of S distributions over X, which we denote as $q_{X|Y}(\cdot|y^{(i)})$ $\forall i \in \{1,2,\dots S\}$

Note that this distribution is defined as a function of the sample $y^{(i)}$. Then, we can write our objective as:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{S} \log\left(\sum_{x} q_{X|Y}(x|y^{(i)}) \cdot \frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})}\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{S} \log\left(E_{q_{X|Y}(\cdot|y^{(i)})}\left\{\frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})}\right\}\right)$$

$$= \underset{\theta}{\operatorname{argmax}}\ f(\theta),$$

where $f(\theta) = \sum_{i=1}^{S} \log\left(E_{q_{X|Y}(\cdot|y^{(i)})}\left[\frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})}\right]\right)$

By Jensen's inequality, since $\log(\cdot)$ is concave,

(5) $$f(\theta) \geq \sum_{i=1}^{S} E_{q_{X|Y}(\cdot|y^{(i)})}\left[\cancel{P_{X,Y}} \log\left[\frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})}\right]\right]$$

with equality iff $\frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})}$ = constant, [Not a function of x]

(6) i.e. iff $q_{X|Y}(x|y^{(i)}) = P_{X|Y}(x, y^{(i)}; \theta_{X,Y})$

⑤

[The ~~equality~~ last statement follows from the observation that $q_{X|Y}(x \mid y^{(i)}) \propto P_{X,Y}(x, y^{(i)})$, and $q_{X|Y}(\cdot)$ is a probability distribution]

We are now ready to give the EM algorithm. Let $g(\theta, q)$ be the RHS in equation (5)

i.e. $g(\theta, q) = \sum_{i=1}^{S} E_{q_{X|Y}(\cdot \mid y^{(i)})} \log \left[\frac{P_{X,Y}(x, y^{(i)}, \theta_{X,Y})}{q_{X|Y}(x \mid y^{(i)})} \right]$

As we know, $g(\theta, q)$ is a lower bound on the true objective function viz. $f(\theta)$. We also know that $\max_{q(\cdot|y)} g(\theta, q) = f(\theta)$, ~~achieved when~~ achieved when $q(\cdot \mid y^{(i)}) = P_{X,Y}(\cdot \mid y^{(i)})$. EM algorithm has 2 steps, one where we maximize over $q(\cdot \mid y^{(i)})$, and the other where we maximize over θ. It is an ~~Em~~ iterative algorithm to maximize the ~~E-step:~~ objective function $g(\theta, q)$.

[For those interested, this procedure of iteratively maximizing a function with 2 sets of parameters, by alternately holding one set of parameters fixed and maximizing over the other, is called 'alternating maximization'.]

EM in a nutshell:

Objective function. $g(\theta, q_{X|Y}(\cdot|y^{(i)}))$

$$= \sum_{i=1}^{S} E_{q_{X|Y}(\cdot|y^{(i)})} \log \left[\frac{P_{X,Y}(x, y^{(i)}; \theta_{X,Y})}{q_{X|Y}(x|y^{(i)})} \right]$$

Initialization: Assign values to parameters $\theta_{X,Y}$, call it $\hat{\theta}^{(0)}_{X,Y}$.

E-step: Given $\hat{\theta}^{(t)}_{X,Y}$, set $\boxed{\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)}) = \underset{q_{X|Y}(\cdot|y^{(i)})}{\operatorname{argmax}} \, g(\hat{\theta}^{(t)}, q_{X|Y}(\cdot|y^{(i)})) \quad \forall i \in \{1, \ldots S\}}$

From equation (6), we already know this optimal value viz.,

$$\boxed{\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)}) = P_{X|Y}(\cdot|y^{(i)}; \hat{\theta}^{(t)}_{X,Y}) \quad \forall i \in \{1, 2, \ldots S\}}$$

M-step: Given $\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)})$, set $\boxed{\hat{\theta}^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \, g(\theta, \hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)}))}$

We can get a slightly simplified expression by leaving out the term in the denominator of the $\log(\cdot)$ in $g(\theta, q)$, since it does not depend on θ.

$$\boxed{\hat{\theta}^{(t+1)}_{X,Y} = \underset{\theta_{X,Y}}{\operatorname{argmax}} \sum_{i=1}^{S} E_{\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)})} \operatorname{Log}\left(P_{X,Y}(x, y^{(i)}; \theta_{X,Y}) \right)}$$

IV. Solution to EM via Sampling

EM is, strictly speaking, a meta-algorithm. It converts our original hard problem into 2 simpler problems viz. E-step and M-step, but does not tell us how to solve these steps

One fairly general method (but not necessarily efficient) of ~~solv~~ solving EM subproblems is via $\boxed{\text{Sampling}}$. Recall the M-step of EM:

$$\hat{\theta}^{(t+1)} = \underset{\theta_{X,Y}}{\arg\max} \sum_{i=1}^{S} E_{\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)})} \log\left(P_{X,Y}(x, y^{(i)}; \theta_{X,Y})\right) \quad (7)$$

Note that the E-step simply sets $\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)})$ to $P_{X|Y}(x|y^{(i)}; \hat{\theta}^{(t)}_{X,Y})$, so we do not write it explicitly. Thus, our aim is to solve M-step where q is described as above.

We will do this by generating a large number of samples of X, for each observed sample $y^{(i)}$ $\forall i \in \{1, 2, \dots S\}$. These samples of X, when juxtaposed with the original samples of Y, will give us a large number of samples of (X, Y), which are representative of the ~~original~~ distribution $P_{X,Y}(x, y; \hat{\theta}^{(t)}, D)$ *. Then, our task of determining $\hat{\theta}^{(t+1)}$ simply boils down to finding Maximum Likelihood estimates in the fully observed setting, a problem which we have already solved.

* ~~Note:~~ It seems difficult to give a formal definition of the quantity $P_{X,Y}(x, y; \hat{\theta}^{(t)}, D)$. Thus, it is for intuition only.

Let us do the above formally. For each observation $y^{(i)}$ $\forall i \in \{1, 2, \dots, S\}$, we generate N_i samples of x from the distribution $q^{(t+1)}_{X|Y}(\cdot|y^{(i)})$. Call these samples $\{X^{(i,1)}, X^{(i,2)}, \dots, X^{(i,N_i)}\}$. If N_i is sufficiently large, then we can write

$$E_{\hat{q}^{(t+1)}_{X|Y}(\cdot|y^{(i)})} \log\left(P_{X,Y}(x, y^{(i)}; \theta_{X,Y})\right)$$

$$\approx \frac{1}{N_i} \sum_{j=1}^{N_i} \log\left(P_{X,Y}(x^{(i,j)}, y^{(i)}; \theta)\right)$$

The above statement merely says that the empirical expectation approximately equals the true expectation, when N_i is large. This is basically the Law of Large Numbers (LLN).

Substituting the above in equation (7), we get

$$\hat{\Theta}^{(t+1)} = \underset{\Theta}{\arg\max} \sum_{i=1}^{S} \frac{1}{N_i} \sum_{j=1}^{N_i} \log\left(P_{X,Y}\left(x^{(i,j)}, y^{(i)}, \theta\right)\right)$$

Now let $N_i = N \;\forall i$, i.e. generate the same number of X samples for each $y^{(i)}$. Then, the above expression reduces to:

$$\hat{\Theta}^{(t+1)} = \underset{\Theta}{\arg\max} \frac{1}{N} \sum_{i=1}^{S} \sum_{j=1}^{N} \log\left(P_{X,Y}\left(x^{(i,j)}, y^{(i)}; \theta\right)\right)$$

$$= \underset{\Theta}{\arg\max} \sum_{i=1}^{S} \sum_{j=1}^{N} \log\left(P_{X,Y}\left(x^{(i,j)}, y^{(i)}, \theta\right)\right)$$

(8) $$\hat{\Theta}^{(t+1)} = \underset{\Theta}{\arg\max} \sum_{(x,y) \in \text{"Samples"}} \log\left(P_{X,Y}(x, y; \theta)\right)$$

Where "Samples" is the set of all (X, Y)-samples we have generated so far, suitably put together.

That is,

$$\boxed{\text{Samples} = \left\{ \begin{array}{l} \{x^{(1,1)}, y^{(1)}\}, \{x^{(1,2)}, y^{(1)}\}, \ldots \{x^{(1,N)}, y^{(1)}\} \\ \vdots \\ \{x^{(S,1)}, y^{(S)}\}, \{x^{(S,2)}, y^{(S)}\}, \ldots \{x^{(S,N)}, y^{(S)}\} \end{array} \right\}}$$

Now we make the striking observation — problem (8) is simply the problem of finding ML estimate of θ, when the true data given to us is "Samples"!

For the particular case of directed graphical models, the ML estimates are especially easy to compute. The parameters (θ) are simply conditional probabilities, and their ML estimates are empirical conditional probabilities obtained from the data. (in this case, data being the "Samples" set)

Only 1 important question remains viz

How to sample?

Sampling from $q_{X|Y}(\,|y^{(i)})$:

The first thing to note is that given any pair of values (x,y) of the random variables (X,Y), we can calculate the probability of this ~~value~~ pair of values, given the parameters θ i.e. $P_{X,Y}(x,y) = P_{X,Y}(x,y;\theta_{X,Y})$

[The above assumes a directed G.M.; if we had an undirected G.M., we could only calculate probabilities upto a normalization constant. But even that will be good enough]

(9) Since $\hat{q}^{(t+1)}_{X|Y}(\cdot \mid y^{(i)}) = P_{X|Y}(x \mid y^{(i)}; \hat{\theta}^{(t)}_{X,Y})$

$$\propto P_{X,Y}(x, y^{(i)}; \hat{\theta}^{(t)}_{X,Y}),$$

we know $q(\cdot \mid y^{(i)})$ upto a normalization constant. That is, given any possible value x of r.v. X, we can calculate $\hat{q}^{*}(x) = \frac{\hat{q}(x \mid Y)}{Z}$, where Z is some normalization constant.

Q. How do we sample from a distribution that we can calculate upto a normalization constant?

There are 2 methods we have learnt to handle this situation:

1. MCMC.

2. Importance Sampling: This does not actually generate samples from $q(\cdot)$, but allows us to compute expectation of any function w.r.t. $q(\cdot)$, which is almost always our final aim.

For simplicity, we focus on method 1 i.e. MCMC.
In the metropolis-hastings rule, all we needed to create a M.C. (Markov Chain) with stationary distribution $q(\cdot)$,

was the ratio of probabilities of 2 states i.e. $\frac{\hat{q}(\cancel{x}\, x_b)}{\hat{q}(x_a)}$, where x_b, x_a are any 2 possible values of X.

But this is simply equal to $\frac{\hat{q}^*(x_b)}{\hat{q}^*(x_a)}$, which we can calculate using ~~q~~ equation (9).

This shows that we can easily use MCMC to generate our samples of X corresponding to each $y^{(i)}$, which we needed in our algorithm. This concludes our discussion of Sampling.

Exercise: Can you use Gibbs Sampling for this task?

V. Solution to EM via Sum-Product (for HMMs)

If you have understood the previous section, this should be easy to follow. The key idea here is ~~to~~ that for certain well-structured graphical models, we can compute all the empirical probabilities directly, which we would obtain from sampling.