Machine Learning, advanced course

Assignment 3

- *Author* -
Nadir Ait Lahmouch

January 22, 2021

Contents

3.1 Easier EM for Advertisements

We will design an EM-algorithm to obtain MLE for all the model parameters:

$$\Omega = (\theta^r, \theta^a, \theta^e, \{\psi_{c,d,f} : c \in [C], d \in [D], f \in [F]\})$$

And in the sum we want to maximize, we have X is Bernoulli distributed :

$$X_{n,m,l} = \begin{cases} 1 & \text{if the person n click on the advertisement m in the journal l} \\ 0 & \text{otherwise} \end{cases}$$

and where the Z^r_n, Z^e_i, Z^a_m are the latent class variables and have categorical distributions.

E-step:
We have the log likelihood LL is:

$$\begin{aligned}
\mathbf{LL} &= \sum_{\substack{n\in[N],\\ l\in[L],m\in A(l)}} \log(\boldsymbol{p}(X_{n,m,l}, Z^r_n, Z^e_l, Z^a_m) \\
&= \sum_{\substack{n\in[N],\\ l\in[L],m\in A(l)}} \sum_{\substack{c\in[C],\\ d\in[D],f\in[F]}} \left\{\log\theta^{r,e,a}_{c,d,f} + X_{n,l,m}\log\psi_{c,d,f} + (1 - X_{n,m,l})\cdot\log\left(1-\psi_{c,d,f}\right)\right\}
\end{aligned}$$

And therefore the expectation of the log likelihood is:

$$\begin{aligned}
\boldsymbol{q}(\Omega|\Omega^{(t)}) &= \sum_{n,l,m=1}^{N,L,A(l)} \mathbf{E}_{(Z^r,Z^e,Z^a|X,\Omega^{(t)})}[\log(\boldsymbol{p}(X_{n,m,l}, Z^r_n, Z^e_l, Z^a_m)] \\
&= \sum_{n,l,m=1}^{N,L,A(l)} \sum_{\substack{c\in[C],\\ d\in[D],f\in[F]}} \boldsymbol{p}(Z^r_n = c, Z^e_l = d, Z^a_m = f|X_{n,m,l})\log(\boldsymbol{p}(X_{n,m,l}, Z^r_n = c, Z^e_l = d, Z^a_m = f)) \\
&= \sum_{n,l,m=1}^{N,L,A(l)} \sum_{\substack{c\in[C],\\ d\in[D],f\in[F]}} \mathbf{R}^{c,d,f}_{n,m,l}[\theta^{r,e,a}_{c,d,f} + X_{n,m,l}\log(\psi_{c,d,f}) + (1 - X_{n,m,l})(\log(1-\psi_{c,d,f}))]
\end{aligned}$$

Where $\theta^{r,e,a}_{c,d,f} = \boldsymbol{p}\left(Z^r_n = c, Z^e_l = d, Z^a_m = f\right)$ and the responsibilities are:

$$\mathbf{R}^{c,d,f}_{n,m,l} = \boldsymbol{p}\left(Z^r_n = c, Z^e_i = d, Z^a_m = f | X_{n,m,l}\right)$$

The responsibilities are the only entities than we don't priory know at each step, so let's compute them:

$$\mathbf{R}^{c,d,f}_{n,m,l} = \frac{\theta^{r,e,a}_{c,d,f}[\psi^{X_{n,m,l}}_{c,d,f} + (1-\psi_{c,d,f})^{(1-X_{n,m,l})}]}{\sum\limits_{c'\in[C],d'\in[D],f'\in[F]} \theta^{r,e,a}_{c',d',f'}[\psi^{X_{n,m,l}}_{c',d',f'} + (1-\psi_{c',d',f'})^{(1-X_{n,m,l})}]}$$

M-step:

For both ψ and θ we maximize using Lagrange multiplier and get:

$$\psi_{c,d,f}(t+1) = \frac{\sum\limits_{n,l,m=1}^{N,L,A(l)} X_{n,m,l}\mathbf{R}^{c,d,f}_{n,m,l}}{\sum\limits_{n,l,m=1}^{N,L,A(l)} \mathbf{R}^{c,d,f}_{n,m,l}}$$

$$\theta^{r,e,a}_{c,d,f}(t+1) = \frac{\sum\limits_{c'\in[C],d'\in[D],f'\in[F]} X_{n,m,l}\mathbf{R}^{c,d,f}_{n,m,l}}{N \times \sum\limits_{l=1}^{L} Card(A(l))}$$

With Card(A(l))= $\sum\limits_{a\in A(l)} 1$.

3.3 Complicated likelihood for leaky units on a tree

The goal of this exercise is to find a linear algorithm that computes $\boldsymbol{p}(X|T, M, \sigma, \alpha, \pi)$. First, let's remind that:

$X_{\downarrow u}$ represents the subtree that u is the root of

We have:

$$\boldsymbol{p}(X_{\downarrow root}|T, M, \sigma, \alpha, \pi) = \sum_{c\in[C]} \pi_c \boldsymbol{p}(\bigcap_{u\in V(T)} X_u | Z_{root} = c, T, M, \sigma, \alpha, \pi)$$

Therefore:

$$\begin{aligned}&= \boldsymbol{p}(\bigcap_{u \in V(T)} X_u | T, M, \sigma, \alpha, \pi)\\&= \sum_{c \in [C]} \pi_c \boldsymbol{p}(X_{\downarrow root} | Z_{root} = c, T, M, \sigma, \alpha, \pi)\end{aligned}$$

If u is a node in the tree, we note v_1 is the parent of u, and v_2, v_3 its childs. Of course, for $u = root$, $v_1 = None$ and for the tree's leafs, $v_2 = None, v_3 = None$. For a node u, we therefore have:

$$\boldsymbol{p}(X_{\downarrow u} | Z_{v_1} = c_1, Z_u = c_2, T, M, \sigma, \alpha, \pi) = \sum_{c, c\prime \in [C]} \pi_c \pi_{c\prime} \boldsymbol{p}(X_{\downarrow u} | Z_{v_1} = c_1, Z_u = c_2, Z_{v_2} = c, Z_{v_3} = c\prime, T, M, \sigma, \alpha, \pi) \quad \textbf{(*)}$$

Given Z_{v_1}, Z_u, Z_{v_2} and Z_{v_3}, we have $X_{\downarrow v_2}$ and $X_{\downarrow v_3}$ are independent (the childs subtrees are independant by d-separation) therefore (we ommit the $T, M, \sigma, \alpha, \pi$ in this following derivation due to lack to space):

$$\begin{aligned}\boldsymbol{p}(X_{\downarrow u} | Z_{v_1} = c_1, Z_u = c_2, Z_{v_2} = c, Z_{v_3} = c\prime) &= \boldsymbol{p}(X_u | Z_{v_1} = c_1, Z_u = c_2, Z_{v_2} = c, Z_{v_3} = c\prime)\\&\cdot \boldsymbol{p}(X_{\downarrow v_2} | Z_u = c_2, Z_{v_2} = c) \boldsymbol{p}(X_{\downarrow v_3} | Z_u = c_2, Z_{v_3} = c\prime)\end{aligned}$$

To facilitate the reading, let's put:

$\mathbf{A}_{v_1, u, c_1, c_2} = \boldsymbol{p}(X_{\downarrow u} | Z_{v_1} = c_1, Z_u = c_2, T, M, \sigma, \alpha, \pi)$ with v_1 the parent of the node u.

Then, in the expression **(*)** above, we have:

$$\mathbf{A}_{v_1, u, c_1, c_2} = \sum_{c, c\prime \in [C]} \pi_c \pi_{c\prime} \mathbf{A}_{u, v_2, c_2, c} \mathbf{A}_{u, v_3, c_2, c\prime} \boldsymbol{p}(X_u | Z_{v_1} = c_1, Z_u = c_2, Z_{v_2} = c, Z_{v_3} = c\prime, T, M, \sigma, \alpha, \pi)$$

We can see in the formula above that the **A**s can be computed recusively (as the probabilites in the formula are known, following the given normal distribution). For a node u, such as the cardinal is $\#U$ is number of nodes of the subtree that u is the root of, the complexity of computing its **A** will be:

$$\text{Complexity} = \#U \times C^2$$

In conclusion, to compute $\boldsymbol{p}(X|T, M, \sigma, \alpha, \pi)$, we need to compute $\boldsymbol{p}(X_{\downarrow root}|T, M, \sigma, \alpha, \pi)$ and the complexity will be equal to $|V(T)| \times C^2$.

3.6 Spectral Graph Analysis

First subproblem

Let $G = (V, E)$ be an undirected d-regular graph and L be the normalized Laplacian of G. And let $\mathbf{x} = (x_u)_{u \in V}$, therefore we have:

$$L\mathbf{x} = \left(I - \frac{1}{d}A\right)\mathbf{x} = \mathbf{x} - \frac{1}{d}\left(\sum_{v \in V, (u,v) \in E} x_v\right)_{u \in V} = \left(x_u - \frac{1}{d}\sum_{v \in V, (u,v) \in E} x_v\right)_{u \in V}$$

And therefore we have:

$$\begin{aligned}
\mathbf{x}^T L\mathbf{x} &= \sum_{u \in V} x_u \cdot \left(x_u - \frac{1}{d}\sum_{v \in V, (u,v) \in E} x_v\right) \\
&= \sum_{u \in V} x_u^2 - \frac{1}{d}\sum_{u \in V}\sum_{v \in V, (u,v) \in E} x_u x_v
\end{aligned}$$

Now we know that G is and undirected graph $\implies$ The edges are only counted once and therefore:

$$\textstyle\sum_{u \in V}\sum_{v \in V, (u,v) \in E} x_u x_v = 2\sum_{(u,v) \in E} x_u x_v$$

Also, since the graph is d-regular:

$$\textstyle\sum_{u \in V} x_u^2 = \frac{2}{d}\sum_{(u,v) \in E} x_u^2$$

And now if we replace in $\mathbf{x}^T L\mathbf{x}$ above we get:

$$\begin{aligned}
\mathbf{x}^T L\mathbf{x} = \frac{2}{d}\sum_{(u,v)\in E} x_u^2 - \frac{2}{d}\sum_{(u,v)\in E} x_u x_v &= \frac{1}{d}\sum_{(u,v)\in E} 2x_u^2 - 2x_u x_v \\
&= \sum_{(u,v)\in E} x_u^2 + x_v^2 - 2x_u x_v \\
&= \sum_{(u,v)\in E} (x_u - x_v)^2
\end{aligned}$$

Second subproblem

To show that a matrice A is positive semidefinite, we have to prove that for all $\mathbf{x}$, we have $\mathbf{x}^T A\mathbf{x} \geq 0$.

For the normalized Laplacian L of the graph G, we proved above that:

$$\mathbf{x}^T L\mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^{|V|} \implies \text{L a is positive semidefinite matrix.}$$

Third subproblem

A non trivial minimizer would be a vector $\mathbf{x}$ such that for every $u, v \in V, \quad |x_u - x_v| > 0$.
While a trivial one would be $\mathbf{x} = (\alpha), \alpha \in \mathbb{R}$ (a vector with only the same constant), in this case $\mathbf{x}^T L\mathbf{x} = 0$, but the embedding created with this vector won't be effective.
Therefore we choose to define a non trivial vector.

As seen in the lecture, the vector x_* can be seen as a 1-dimensional embedding, i.e. a mapping of the vertices into the real line. Choosing the vector x_* that minimize the equation (2) in the question sheet (that minimizes the distance between two vertices linked in the graph) is meaningful because it embeds the graph in such a way that the closest vertices in the graph are close in the embedding too.