Homework # 4

TJ Bai

November 28, 2023

1. Suppose each class is parameterized by some θ_c for $c \in [1, \ldots, C]$

   θ_{cd} for $i \in [1, \ldots, D]$ is equal to $p(x_{id} = 1)$

   We seek to maximize the likelihood parameterized by $\theta \ldots$

   $$p(X; \theta) = \prod_{i=1}^{N} \prod_{j=1}^{k} (p(x_i|\theta_j))^{z_{ij}} = \prod_{i=1}^{N} \prod_{j=1}^{k} \left(\prod_{p=1}^{d} \theta_{jp}^{x_{ip}} (1 - \theta_{jp})^{1-x_{ip}} \right)^{z_{ij}}$$

   This is equivalent to maximizing the log-likelihood. . .

   $$\log p(X; \theta) = \sum_{i=1}^{N} \sum_{j=1}^{l} \sum_{p=1}^{d} z_{ij} \left(x_{ip} \log \theta_{jp} + (1 - x_{ip}) \log(1 - \theta_{jp}) \right)$$

   We proceed by taking the partial derivative with respect to $\theta_{jp} \ldots$

   $$\frac{\partial}{\partial \theta_{jp}} \log p(X; \theta) = \sum_{i=1}^{N} z_{ij} \left(\frac{x_{ip}}{\theta_{jp}} - \frac{1 - x_{ip}}{1 - \theta_{jp}} \right) = \sum_{i=1}^{N} z_{ij} \left(\frac{x_{ip} - x_{ip}\theta_{jp} - \theta_{jp} + x_{ip}\theta_{jp}}{\theta_{jp}(1 - \theta_{jp})} \right) = 0$$

   $$\implies \sum_{i=1}^{N} z_{ij} x_{ip} - \sum_{i=1}^{N} z_{ij} \theta_{jp} = 0 \implies \boxed{\hat{\theta}_{jp} = \frac{\sum_{i=1}^{N} z_{ij} x_{ip}}{\sum_{i=1}^{N} z_{ij}}}$$

   This result is intuitive, as it means that the MLE estimator for the pth dimension of the jth parameter is simply the proportion of times that the pth dimension is observed for the jth parameter.

   Let $X \in \mathbb{R}^{N \times d}$ represent the data where the ith column is x_i.

   Let $z_i \in \mathbb{R}^N$ represent the vector of indicators where z_{ij} is 1 if x_j belongs to class i and 0 otherwise.

   If we fix $N_i = \sum_{i=1}^{N} z_{ij}$ as the number of times class i is observed in the data, then we can express the MLE estimator for the entire parameter θ_i as $\boxed{\frac{X^\intercal z_i}{N_i}}$.

2. $\gamma_{ic} = p(z_i = c|x_i; \theta, \pi) = \dfrac{\pi_c p(x_i|\theta_c)}{\sum_{l=1}^{k} \pi_l p(x_i|\theta_l)} = \dfrac{\pi_c \prod_{j=1}^{d} \left[\theta_{cj}^{x_j}(1-\theta_{cj})^{1-x_j}\right]}{\sum_{l=1}^{k} \pi_l \prod_{j=1}^{d} \left[\theta_{lj}^{x_j}(1-\theta_{lj})^{1-x_j}\right]}$

3. Given γ_{ic} computed in the preceding E-step, we derive the M-step updates.

$$p(X_N, Z_N; \theta, \pi) = \prod_{i=1}^{N}\prod_{j=1}^{k}\left(\pi_j p(x_i|\theta_j)\right)^{z_{ij}}$$

$$\implies \log p(X_N, Z_N; \theta, \pi) = \sum_{i=1}^{N}\sum_{j=1}^{k} z_{ij}\left(\log \pi_j + \log p(x_i|\theta_j)\right)$$

$$\implies \mathbb{E}_{z_{ij}}\left[\log p(X_n, Z_n; \theta, \pi)\right] = \sum_{i=1}^{N}\sum_{j=1}^{k} \gamma_{ij}\left(\log \pi_j + \log p(x_i|\theta_j)\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{k} \gamma_{ij}\left(\log \pi_j + \log\left(\prod_{p=1}^{d} \theta_{jp}^{x_{ip}}(1-\theta_{jp})^{1-x_{ip}}\right)\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{k} \gamma_{ij}\left(\log \pi_j + \sum_{p=1}^{d}\left(x_{ip}\log\theta_{jp} + (1-x_{ip})\log(1-\theta_{jp})\right)\right)$$

We seek π and θ to maximize this quantity. We proceed by taking the derivative with respect to θ_{jp}, which is the pth dimension of the jth component of the mixture.

$$\frac{\partial}{\partial\theta_{jp}}\mathbb{E}_{z_{ij}}\left[\log p(X_n, Z_n; \theta, \pi)\right] = \sum_{i=1}^{N}\gamma_{ij}\left(x_{ip}\frac{\partial}{\partial\theta_{jp}}\log\theta_{jp} + (1-x_{ip})\frac{\partial}{\partial\theta_{jp}}\log(1-\theta_{jp})\right)$$

$$= \sum_{i=1}^{N}\gamma_{ij}\left(\frac{x_{ip}}{\theta_{jp}} - \frac{1-x_{ip}}{1-\theta_{jp}}\right) = \sum_{i=1}^{N}\gamma_{ij}\left(\frac{x_{ip} - x_{ip}\theta_{jp} - \theta_{jp} + x_{ip}\theta_{jp}}{\theta_{jp}(1-\theta_{jp})}\right) = \sum_{i=1}^{N}\frac{\gamma_{ij}(x_{ip}-\theta_{jp})}{\theta_{jp}(1-\theta_{jp})}$$

To find the stationary point with respect to θ_{jp} ...

$$\implies \sum_{i=1}^{N}\frac{\gamma_{ij}(x_{ip}-\theta_{jp})}{\theta_{jp}(1-\theta_{jp})} = 0$$

$$\implies \sum_{i=1}^{N}\gamma_{ij}x_{ip} - \theta_{jp}\sum_{i=1}^{N}\gamma_{ij} = 0$$

$$\implies \boxed{\hat{\theta}_{jp} = \frac{\sum_{i=1}^{N}\gamma_{ij}x_{ip}}{\sum_{i=1}^{N}\gamma_{ij}}}$$

Let $X \in \mathbb{R}^{N\times d}$ represent the data where the ith row is x_i.

Let $\gamma_j \in \mathbb{R}^N$ represent the vector where the ithth value is γ_{ij}.

Then, we can express $\boxed{\hat{\theta}_j = \dfrac{X^\intercal\gamma_j}{\sum_{i=1}^{N}\gamma_{ij}}}$

We now maximize with respect to π, subject to the constraint $\sum_{i=1}^{k} \pi_i = 1$.

Consider the Lagrangian $\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{k} \gamma_{ij} \left(\log \pi_j + \log p(x_i|\theta_j) \right) + \lambda \left(1 - \sum_{i=1}^{k} \pi_i \right)$

Taking the derivative with respect to $\pi_c \ldots$

$$\implies \sum_{i=1}^{N} \frac{\gamma_{ic}}{\pi_c} - \lambda = 0$$

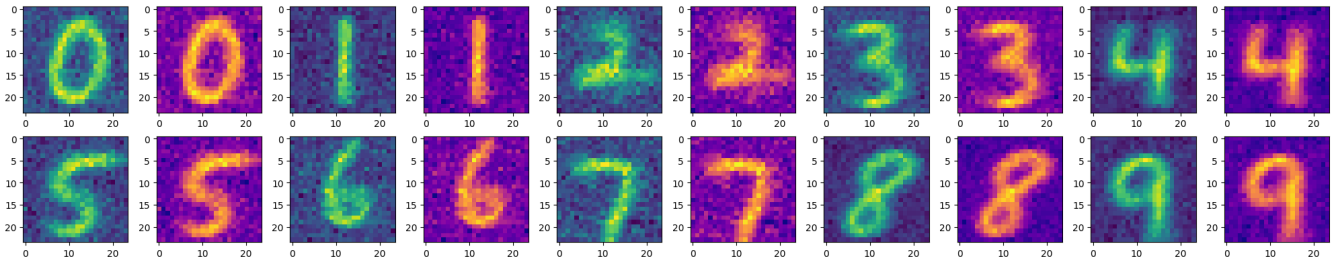
$$\implies \pi_c = \frac{\sum_{i=1}^{N} \gamma_{ic}}{\lambda}$$

Now to solve for λ, we know $\sum_{j=1}^{k} \pi_j = 1$.

$$\implies \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} \gamma_{ij}}{\lambda} = \frac{N}{\lambda} = 1 \implies \lambda = N$$

Thus, $\boxed{\hat{\pi}_c = \frac{\sum_{i=1}^{N} \gamma_{ic}}{N}}$

4. Here, we visualize θ for 2 randomly selected components from each MNIST class.



It appears that the weights differ very slightly between classes, implying that a lower number of mixture components (perhaps even 1) would be sufficient for the classification task. This also implies that the MNIST datasets tends to be relatively homogeneous and there is not a sufficiently diverse selection of contexts for each class for a large number of mixture components to be relevant.

For tuning the mixture models, I first restricted the number of mixture components to a fixed number n and then experimented with a range of values for epsilon and EM iterations. I found that slightly smaller n values tended to perform better on the valuation set. This could be justified because a large number of components would tend to overfit the training data.

I also found that increasing the EM iterations from 50 to 100 resulted in higher validation accuracy, on average. Changing the epsilon values didn't change much, in general, although I expected a slightly larger epsilon value to generalize better because it acts as a form of regularization.

Ultimately through this training procedure, I landed on 4 mixture components/class, ϵ = 1e-12, and 100 EM iterations for the sentiment analysis task, resulting in a validation accuracy of 0.752. For the MNIST task, I landed on 3 mixture components/class, ϵ = 1e-7, and 50 EM iterations for a validation accuracy of 0.774.