STATISTICAL MODELLING AND INFERENCE

Sergio-Yersi Villegas Pelegrín

SEMINAR 4: Problems 4 & 5

Problem 4

We want to derive the Bayesian EM algorithm for maximizing the posterior probability $P(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{x})$ of the studied Bernoulli mixture model, together its corresponding priors for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. To begin with, we summarize the Bayesian Bernoulli mixture model we have: for a random sample $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, for each $\boldsymbol{x}_i$ we have:

$$\boldsymbol{x}_i|\boldsymbol{\theta}_k, z_i = k \sim Bern(\boldsymbol{\theta}_k)$$

$$z_i|\boldsymbol{\pi} \sim Multinomial(\boldsymbol{\pi})$$

$$\boldsymbol{\pi}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$$

$$\boldsymbol{\theta}_k|\boldsymbol{a}_k, \boldsymbol{b}_k \sim Beta(\boldsymbol{a}_k, \boldsymbol{b}_k)$$

Now, in order to be able to analyze our posterior distribution of interest, we must know the conditional independence relationships between all parameters:

- $\boldsymbol{\pi}$ just depends on $\boldsymbol{\alpha}$ as a Dirichlet distribution (one of the added priors in this problem).
- $\boldsymbol{z}$ depends on $\boldsymbol{\pi}$ as a Multinomial distribution and then, it also depends on $\boldsymbol{\alpha}$, since $\boldsymbol{\pi}$ does.
- $\boldsymbol{\theta}$ depends on the $\{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K$ parameters of the Beta distribution (the other prior added in this problem).
- Finally, $\boldsymbol{x}$ depends on all the aforementioned parameters

Clearly, directly trying to compute $P(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{x})$ is infeasible. However, we can use Gibbs sampling and construct a Markov chain that allows us to approximate this distribution arbitrarily accurate. Specifically, we will iterate between all the parameters and sample each one of them given the rest. In the following steps, we will keep using both Bayes rule (to alternate between conditional probabilities and the probability of the intersection) and the conditional independence property (simplifying conditional probabilities by excluding the independent parameters in each case), in order to get to the solution.

Step 1: sample $\boldsymbol{\pi}$ given $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\theta}_k, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K$:

$$P(\boldsymbol{\pi}|\boldsymbol{z}, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\theta}_k) = P(\boldsymbol{\pi}|\boldsymbol{z}, \boldsymbol{\alpha}) \sim P(\boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\alpha}) = P(\boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\alpha}) \cdot P(\boldsymbol{\pi}, \boldsymbol{\alpha}) \sim P(\boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\alpha}) \cdot P(\boldsymbol{\pi}|\boldsymbol{\alpha}) \sim$$

$$\sim Multinomial(\boldsymbol{\pi}) \cdot Dir(\boldsymbol{\alpha}) \sim Dir(\boldsymbol{\alpha} + \boldsymbol{n})$$

where $n_j = \sum_{j=1}^{n} \mathbf{1}(z_i = j)$, meaning that it sums 1 whenever $z_i = j$.

Step 2: sample $\boldsymbol{\theta}_k$ given $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K$:

$$P(\boldsymbol{\theta}_k|\boldsymbol{z}, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\pi}) = P(\boldsymbol{\theta}_k|z_i, \boldsymbol{a}_k, \boldsymbol{b}_k, \boldsymbol{x}_i) \sim P(\boldsymbol{\theta}_k, z_i, \boldsymbol{a}_k, \boldsymbol{b}_k, \boldsymbol{x}_i) = P(\boldsymbol{x}_i|\boldsymbol{\theta}_k, z_i, \boldsymbol{a}_k, \boldsymbol{b}_k) \cdot P(\boldsymbol{\theta}_k, z_i, \boldsymbol{a}_k, \boldsymbol{b}_k) =$$

$$= P(\boldsymbol{x}_i|\boldsymbol{\theta}_k, z_i) \cdot P(\boldsymbol{\theta}_k, z_i, \boldsymbol{a}_k, \boldsymbol{b}_k) \sim P(\boldsymbol{x}_i|\boldsymbol{\theta}_k, z_i) \cdot P(\boldsymbol{\theta}_k|z_i, \boldsymbol{a}_k, \boldsymbol{b}_k) = P(\boldsymbol{x}_i|\boldsymbol{\theta}_k, z_i) \cdot P(\boldsymbol{\theta}_k|\boldsymbol{a}_k, \boldsymbol{b}_k) \sim Bern(\boldsymbol{\theta}_k) \cdot Beta(\boldsymbol{a}_k, \boldsymbol{b}_k) \sim$$

$$\sim \prod_{j=1}^{d} \theta_{kj}^{x_{ij}} \cdot (1-\theta_{kj})^{1-x_{ij}} \cdot \theta_{kj}^{a_{kj}-1} \cdot (1-\theta_{kj})^{b_{kj}-1} \sim \prod_{j=1}^{d} \theta_{kj}^{a_{kj}+x_{ij}-1} \cdot (1-\theta_{kj})^{b_{kj}-x_{ij}} \sim Beta(\boldsymbol{a}_k + \boldsymbol{x}_i, 1 + \boldsymbol{b}_k - \boldsymbol{x}_i)$$

Step 3: sample $\boldsymbol{z}$ given $\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\theta}_k, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K$:

For each i, $z_i \in \{1, ..., K\}$ given $\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\theta}_k, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K$ follows a multinomial distribution such that:

$$P(z_i|\boldsymbol{\theta}_k, \{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\pi}) = \frac{\pi_k \cdot Bern(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \pi_k \cdot Bern(\boldsymbol{x}_j|\theta_{kj})}$$

Now that we know the posterior distributions in steps $1-3$, we can run the Gibbs sampling algorithm, where the approximate posterior probabilities can be obtained by Monte Carlo approximation.

Problem 5

As seen in *Problem 3*, the K-means algorithm arises as a particular limit of the Gaussian mixture EM-algorithm. We want to show that, for $\epsilon \to 0$, maximizing the expected complete-data log likelihood (ECL) for this model is equivalent to minimizing the distortion measure (J) for the K-means algorithm. Therefore, we will compute both derivatives and see that, when $\epsilon \to 0$, we get the same result:

Distortion measure (J)

J can be written as the following:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$$

Therefore, we can minimize it in respect of $\boldsymbol{\mu}_k$, by fixing $k = k'$ (therefore not computing the sum over all k) and obtain the following result:

$$\left.\frac{\partial J}{\partial \boldsymbol{\mu}_k}\right|_{\boldsymbol{\mu}_{k'}^J} = \sum_{i=1}^{n} r_{ik'} \cdot 2 \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}^J\| \cdot (-1) = 0 \Rightarrow \sum_{i=1}^{n} r_{ik'} \cdot \boldsymbol{x}_i = \boldsymbol{\mu}_{k'}^J \cdot \sum_{i=1}^{n} r_{ik'} \Rightarrow \boldsymbol{\mu}_{k'}^J = \frac{\sum_{i=1}^{n} r_{ik'} \cdot \boldsymbol{x}_i}{\sum_{i=1}^{n} r_{ik'}}$$

Expected complete-data log likelihood (ECL)

We can write our distribution function, since we are working with a Gaussian Mixture Model, as the following:

$$P(\boldsymbol{x}_i, \boldsymbol{z_i}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \cdot N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the parameters of the normal distribution. Then, maximizing the expected complete-data log likelihood can be expressed as:

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{z_i}|\boldsymbol{x}_i} \cdot log(P(\boldsymbol{x}_i, \boldsymbol{z_i}|\boldsymbol{\theta})) = \arg\max_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \cdot log(\sum_{k=1}^{K} \pi_k \cdot N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

However, as seen in *Problem 3*, the $\mathbf{\Sigma}_k$ parameter can be written as $\mathbf{\Sigma}_k = \epsilon \mathbb{I}_d$, where $\mathbb{I}_d$ is the identity matrix for d dimensions. Therefore, $\mathbf{\Sigma}_k$ is no longer a parameter, but a constant $\mathbf{\Sigma}_k = \epsilon^d$. Together with the fact that, again, we can compute the corresponding derivative by fixing $k = k'$ (therefore not computing the sum over all k), we could now write the maximization expression, just in respect of $\boldsymbol{\mu}_k$, as:

$$\boldsymbol{\mu}_{k'} = \arg\max_{\boldsymbol{\mu}_{k'}} \sum_{i=1}^{n} \gamma_{ik'} \cdot (log(\frac{\pi_{k'}}{\sqrt{2\pi\epsilon^d}}) - \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}\|^2}{2\epsilon^d}) \equiv \arg\max_{\boldsymbol{\mu}_{k'}} \textit{ECL}$$

Then, we obtain the following result:

$$\left.\frac{\partial \textit{ECL}}{\partial \boldsymbol{\mu}_k}\right|_{\boldsymbol{\mu}_{k'}^{ECL}} = \sum_{i=1}^{n} -\frac{\gamma_{ik'} \cdot 2 \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}^{ECL}\|}{2\epsilon^d} = \sum_{i=1}^{n} -\frac{\gamma_{ik'} \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}^{ECL}\|}{\epsilon^d} = 0 \Rightarrow \sum_{i=1}^{n} \gamma_{ik'} \cdot \boldsymbol{x}_i = \boldsymbol{\mu}_{k'}^{ECL} \cdot \sum_{i=1}^{n} \gamma_{ik'}$$

$$\boldsymbol{\mu}_{k'}^{ECL} = \frac{\sum_{i=1}^{n} \gamma_{ik'} \cdot \boldsymbol{x}_i}{\sum_{i=1}^{n} \gamma_{ik'}}$$

As we saw in *Problem 3*: $\epsilon \to 0 \Rightarrow \ \gamma_{ik} \to r_{ik}$. Therefore, $\boldsymbol{\mu}_{k'}^{ECL} = \boldsymbol{\mu}_{k'}^{J}$, proving that maximizing the expected complete-data log likelihood for this model is equivalent to minimizing the distortion measure for the K-means algorithm.