Advanced Probabilistic Machine Learning and Applications

Caterina De Bacco and Nicoló Ruggeri

1 Tutorial 12: Variational Inference

Exercise 1: implementing Gaussian Mixture Model with VI

In this tutorial we will implement the CAVI algorithm for the GMM model and replicate the figure of Lecture 12 notes.

(a) Draw synthetic dataset of GMM with $K = 3$ and $N = 1000$ datapoints.

(b) Derive the ELBO for this model.

(c) Implement the CAVI updates for this model using what learned in the Lecture.

(d) Run the algorithm and plot the results at convergence.

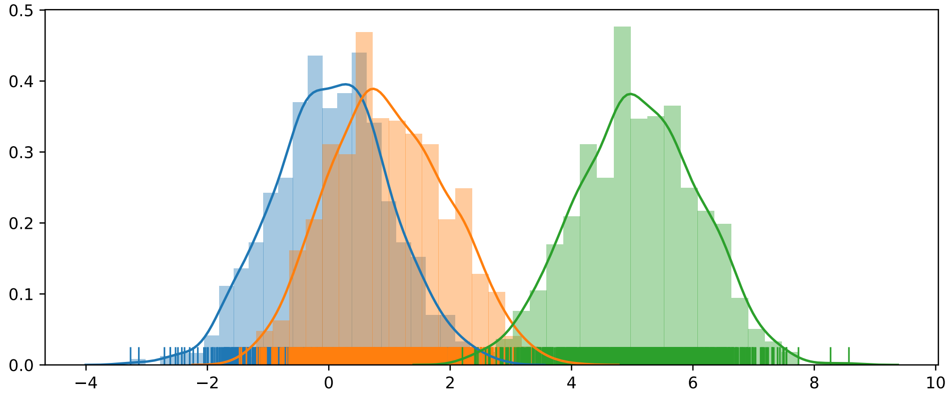
(e) Plot the behavior of the ELBO over iteration time. Comment.

(f) Plot the state of the variational result for the mixture mean parameters at the iteration times where the ELBO changes significantly. Comment.

(g) Calculate accuracy in labelling the datapoints by cluster.

Solution.

(a) (see Jupyter Notebook). We draw 1000 points for every cluster, for a total of $K \cdot 1000$ datapoints. The densities of the three clusters look like the following:



(b) Recall from the lecture that we are dealing with the following probabilistic model and variational approximation

$$
\begin{aligned}
p(\mu, c) &= \prod_k p(\mu_k) \prod_i p(c_i) \\
&= \prod_k N(\mu_k\,;\,0, \sigma^2) \prod_i Cat\left(\frac{1}{K}\right) && \textbf{prior} \\
p(X|\mu, c) &= \prod_i p(x_i|c_i, \mu) \\
&= \prod_i \prod_k p(x_i|\mu_k)^{c_{ik}} \\
&= \prod_i \prod_k N(x_i\,;\,\mu_k, 1)^{c_{ik}} && \textbf{likelihood} \\
q(\mu, c\,;\,m, s^2, \rho) &= \prod_k q(\mu_k\,;\,m_k, s_k^2) \prod_i q(c_i\,;\,\rho_i) \\
&= \prod_k N(\mu_k\,;\,m_k, s_k^2) \prod_i Cat(\rho_i) && \textbf{variational approx.}
\end{aligned}
$$

and the ELBO is defined as

$$
\textbf{ELBO} = \underbrace{\mathbb{E}_q[\log p(x|\mu, c)]}_{\textcircled{1}} + \underbrace{\mathbb{E}_q[\log p(\mu, c)]}_{\textcircled{2}} \underbrace{- \mathbb{E}_q[\log q(\mu, c)]}_{\textcircled{3}}\,. \tag{1}
$$

We now compute the three terms explicitly:

$$
\begin{aligned}
\textcircled{1} &= \sum_i \sum_k \mathbb{E}_q\left[c_{ik} \log p(x_i|\mu_k)\right] && (c, \mu \text{ independent for } q) \\
&= \sum_i \sum_k \mathbb{E}_q[c_{ik}] \mathbb{E}_q\left[\log p(x_i|\mu_k)\right] && \left(\mathbb{E}_q[c_{ik}] = \rho_{ik}\right) \\
&\propto \sum_i \sum_k -\frac{1}{2} \rho_{ik} \mathbb{E}_q\left[(x_i - \mu_k)^2\right] \\
&= \sum_i \sum_k -\frac{1}{2} \rho_{ik} (s_k^2 + (x_i - m_k)^2)\,,
\end{aligned}
$$

where the last passage can be obtained by expanding $(x_i - \mu_k)^2$ and taking expectations or by noticing that

$$\mu_k \sim N(m_k, s_k^2) \implies x_i - \mu_k \sim N(x_i - m_k, s_k^2)$$

and

$$\mathbb{E}[(x_i - \mu_k)^2] = Var(x_i - \mu_k) + \mathbb{E}[(x_i - \mu_k)]^2 = s_k^2 + (x_i - m_k)^2 \,.$$

$$\begin{aligned}
\textcircled{2} &= \sum_k \mathbb{E}_q[\log p(\mu_k)] + \sum_i \mathbb{E}_q[\log p(c_i)] \\
&\propto \sum_k -\frac{1}{2\sigma^2}\mathbb{E}_q[\mu_k^2] + \sum_i \sum_k \log \frac{1}{K}\mathbb{E}_q[c_{ik}] \\
&= \sum_k -\frac{1}{2\sigma^2}(s_k^2 + m_k^2) + \sum_i \log \frac{1}{K}\sum_k \rho_{ik} & \left(\sum_k \rho_{ik} = 1\right) \\
&\propto \sum_k -\frac{1}{2\sigma^2}(s_k^2 + m_k^2) \,.
\end{aligned}$$

Finally

$$\begin{aligned}
\textcircled{3} &= -\sum_k \mathbb{E}_q[\log q(\mu_k)] - \sum_i \mathbb{E}_q[\log q(c_i)] \\
&\quad \left(\text{entropy of gaussian } -\mathbb{E}_q[\log q(\mu_k)] \propto \frac{1}{2}\log(s_k^2)\right) \\
&= \frac{1}{2}\sum_k \log s_k^2 - \sum_i \sum_k \mathbb{E}_q[c_{ik}] \log \rho_{ik} \\
&= \frac{1}{2}\sum_k \log s_k^2 - \sum_i \sum_k \rho_{ik} \log \rho_{ik} \,.
\end{aligned}$$

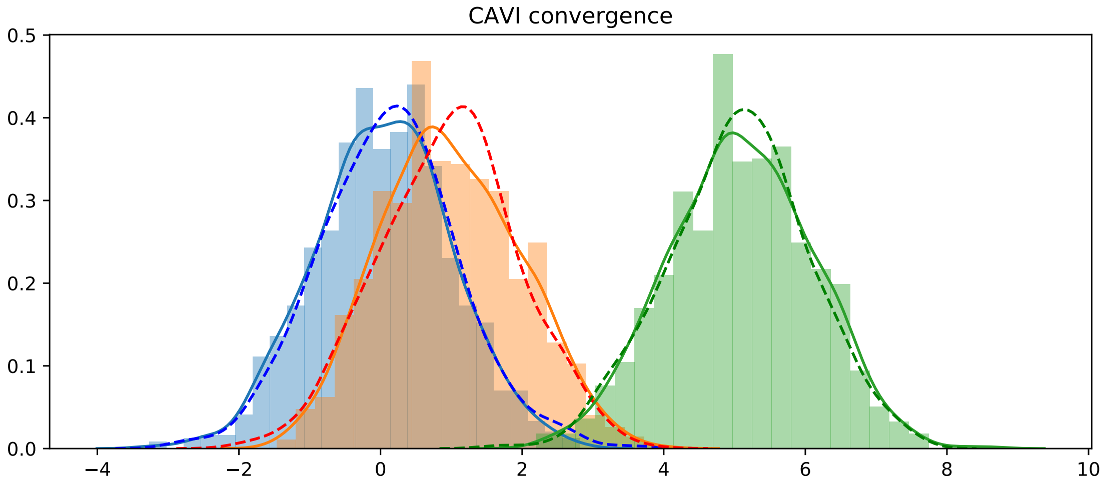
Comment: one could wonder why $\mathbb{E}[c_{ik}] = \rho_{ik}$. Since for a mean field q the r.v. c_i is independent from all the other latent variables, expectations can be marginalized:

$$\begin{aligned}
\mathbb{E}_q[c_{ik}] &= \mathbb{E}_{q(c_i;\rho_i)}[c_{ik}] & \text{(definition of expected value)} \\
&= \sum_{c_i} \underbrace{(\prod_m \rho_{im}^{c_{im}})}_{\mathbb{P}(c_i)} c_{ik} \\
&= \rho_{ik}
\end{aligned}$$

since the sum $\sum_{c_i}$ runs over all the possible values of the one-hot vector c_i, i.e. for a specific realization of c_i only one of the terms $c_{ij} = 1$ and for all the others $c_{im} = 0 \, \forall m \neq j$. Only the realization for which $j = k$ yields a non zero value of c_{ik}, in which case $c_{ik} = 1$ and $(\prod_m \rho_{im}^{c_{im}})c_{ik} = \rho_{ik}^1 \cdot 1 = \rho_{ik}$.

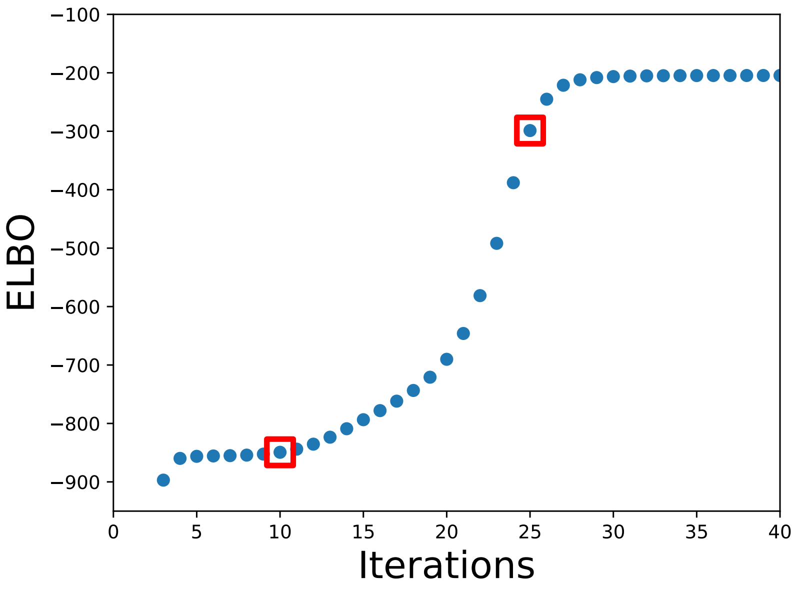
(c) (see Jupyter Notebook)

(d) results at convergence are plotted below:



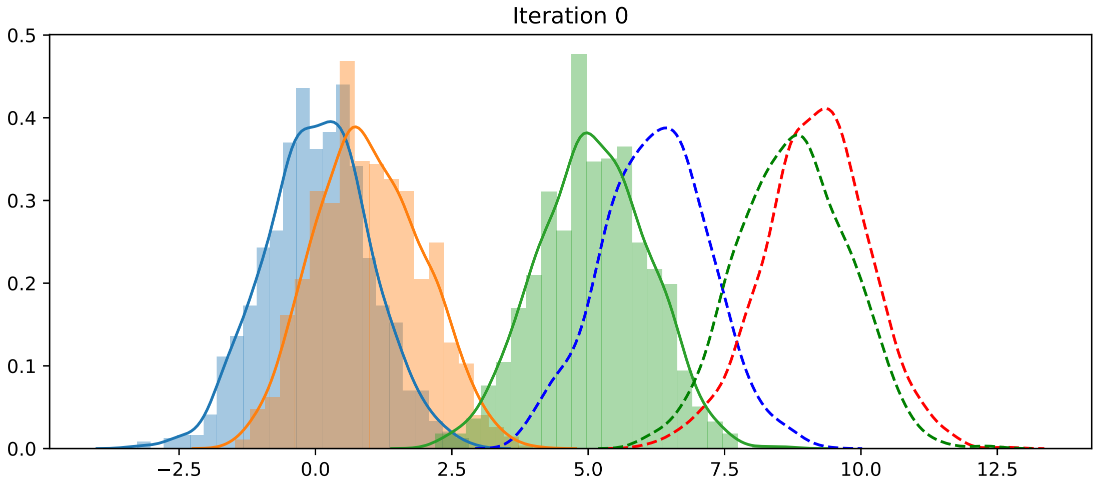
where the continuous lines represent the ground truth clusters' estimated densities, and the dashed lines densities estimated from drawings from the variational approximations of the clusters' distributions. As we can see cluster means and assignments are approximated pretty well.

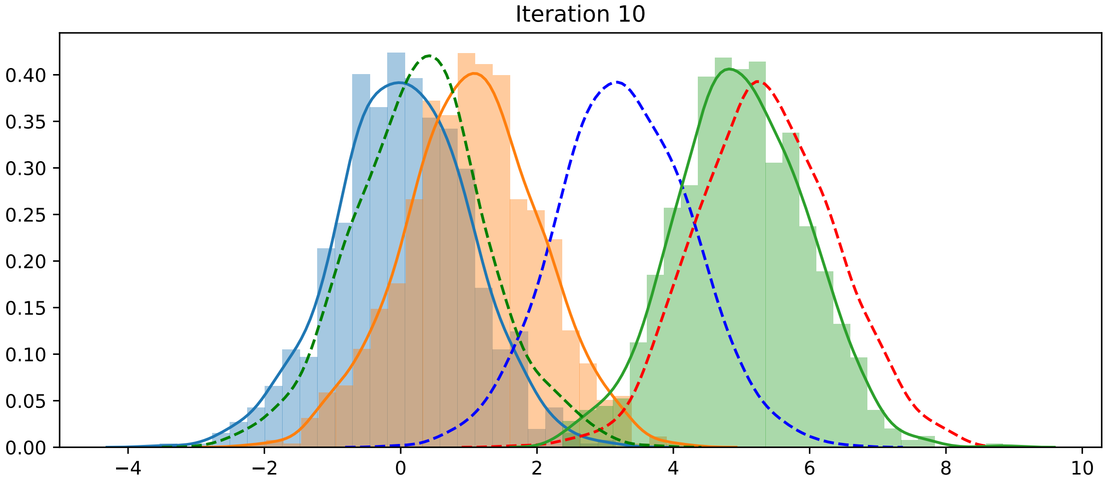
(e) plotting the ELBO (ignoring the constant terms and including only the terms from point (b)) yields

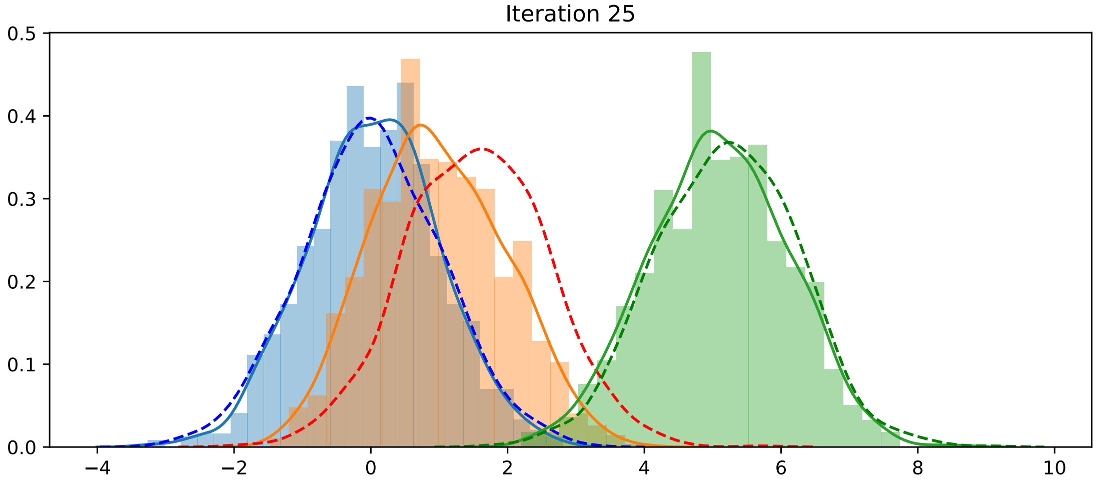


where the boxed points are the (qualitative) changing points in the ELBO trend. We will see in the next point how this qualitative points reflect in the approximation of the means.

(f) we plot the approximated density for the initialization and the two changing points in the ELBO convergence







as we can see the changing points in the ELBO convergence reflect themsleves in the clusters' means approximation. Specifically, in the first changing point we see that two of the three clusters have more or less been detected, while one is still needs to converge. This happens after the second changing points. From the second changing points to the convergence there is no major change in the variational parameters, but rather a refinement of their values.

(g) (see Jupyter Notebook). Apart from the code, playing with the cluster assignments and the model hyper-parameters (e.g. σ^2) should result in different values of accuracy. Notice that accuracy is not a well-defined notion in clustering. To overcome the identifiability problem with the clusters, we make greedy accuracy maximizing assignments between approximated and ground truth clusters.

□