APMLA: Lecture 2

Caterina De Bacco and Isabel Valera

1 Gaussian Mixture Models (GMMs): An example of a simple but expressive generative model

While, as seen in the previous lecture, the Gaussian distribution has some important analytical properties, it may be not expressive enough to capture the underlaying distribution of real-world data, which often require of multimodal, heavy-tailed, or asymmetric distributions. Bishop (2006) provides an example of the 'Ol Faithful' dataset, shown here in Figure 1, where it is clear that a a Gaussian distribution is not expressive enough to accurately fit the data, but instead a *mixture* of two Gaussian distributions are necessary to better fit the data.

Gaussian Mixture Distribution. In general, the superposition of K Gaussian distributions can be formulated as the following probabilistic model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

which is called *Gaussian mixture distribution*. Here, each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \sigma_x)$ is often referred as a *component* of the mixture and is characterized by its mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$; and the parameters π_k are called *mixing coefficients* and should fulfil that $\sum_{k=1}^{K} \pi_k = 1$ (with $0 \leq \pi_k \leq 1$) in order for the resulting $p(\mathbf{x})$ to be a valid probability density function (pdf), i.e., $p(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ and $\int p(\mathbf{x}) d\mathbf{x} = 1$ (refer to Section 2.3.9 of Bishop (2006) for details on the proof). We also remark that the mixing coefficients π_k correspond to the prior probability of *picking* the k-th component in the mixture.

Gaussian Mixture Model (GMM). Let us know introduce a categorical latent variable $z \in \{1, \dots, K\}$, such that the joint distribution of the observed variable $\mathbf{x}$ and the latent variable z factorizes as:

$$p(\mathbf{x}, z) = p(\mathbf{x}|z)p(z),$$

where $p(z = k) = \pi_k$ and $p(\mathbf{x}|z = k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Alternatively, we can write $p(\mathbf{x}|z) = \prod_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{[z=k]}$, where $[z = k]$ returns one iff z takes value k.

The marginal distribution of the observed variable $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \sum_{z=1,\dots,K} p(\mathbf{x}, z) = \sum_{z=1,\dots,K} p(z)p(\mathbf{x}|z) = \sum_{k=1,\dots,K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

which shows that the marginal distribution of $\mathbf{x}$ is indeed a Gaussian mixture distribution.

As a result, one can easily generate samples $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ from a Gaussian mixture distribution by using the generative process of the GMM, where we first sample each latent variable z_n from a Categorical distribution with category probabilities $\boldsymbol{\pi} = (\pi_1, \dots \pi_k)$, and then sample the corresponding observation $\mathbf{x}_n$ from $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n})$. The graphial model corresponding to this generative model is shown in Figure 3.

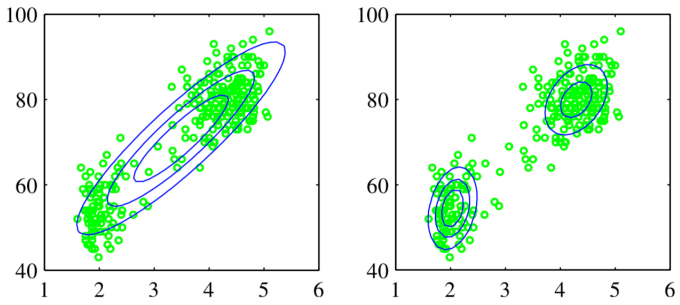

Figure 1: The 'Old Faithful' dataset (Figure 2.21 from Bishop (2006)).

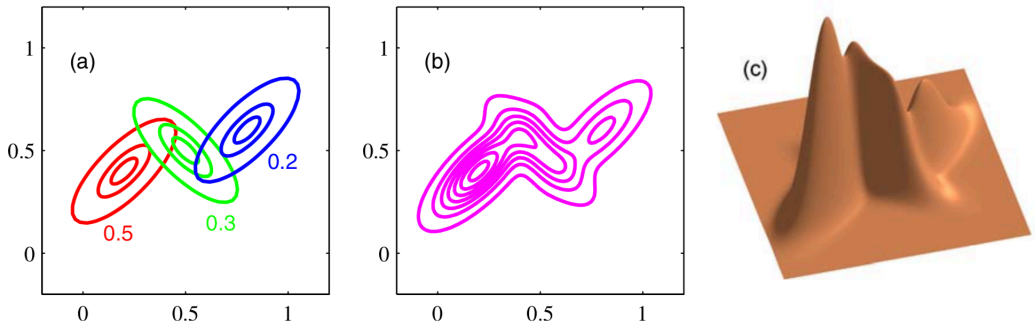

Figure 2: Illustration of a mixture of 3 Gaussians in a two-dimensional space (Figure 2.23 from Bishop (2006)).

2 MLE solution for GMMs: Introduction of the E-M algorithm

In the following, we show how to obtain the MLE solution for the parameters of the GMM.

Maximum Likelihood Estimation. As shown before, the Gaussian mixture distribution is characterized by the parameters $\pi = (\pi_1, \dots \pi_k)$, $\{\mu_k, \Sigma_k\}_{k=1}^K$, which one may think to estimate given the observed dataset $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ by *maximum likelihood estimation* (MLE) as:

$$\pi, \{\mu_k, \Sigma_k\}_{k=1}^K = \underset{\pi, \{\mu_k, \Sigma_k\}_{k=1}^K}{\arg\max} \; \mathcal{L}(\pi, \{\mu_k, \Sigma_k\}_{k=1}^K), \tag{2}$$

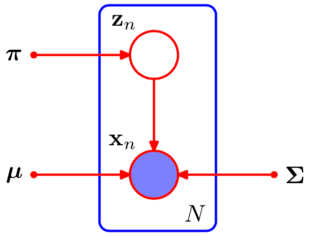

Figure 3: Graphical model for the GMM (Figure 9.6 from Bishop (2006)).

where

$$\mathcal{L}(\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) = \ln p(\mathbf{X}|\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) = \sum_{n=1}^N \ln \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right). \tag{3}$$

Observations: Unfortunately, in contrast to the MLE solution for the Gaussian distribution, the above optimization problem does not have closed-form solution (since there is a sum inside the logarithm). Also, there exist more than one equivalent MLE solutions. Specifically, for a K-component mixture we will have a total of $K!$ equivalent MLE solutions corresponding to the $K!$ ways of assigning K sets of parameters to K components.

Nevertheless, as done for the Gaussian distribution, we set the derivative with respect to the mean parameter $\boldsymbol{\mu}_k$ to zero, i.e.,

$$0 = \sum_{n=1}^N \gamma_k(\mathbf{x}_n) \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k),$$

where we have defined

$$\gamma_k(\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{4}$$

which corresponds to the posterior probability of $z_n = k$, i.e., the posterior probability that the observation $\mathbf{x}_n$ has been sampled from the component (cluster) k.

Then we can write the MLE result for $\boldsymbol{\mu}_k$ as

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{z_n}(\mathbf{x}_n) \mathbf{x}_n, \tag{5}$$

where we have defined

$$N_k = \sum_{n=1}^N \gamma_k(\mathbf{x}_n), \tag{6}$$

which can be interpreted as the effective number of points assigned to cluster k.

Following a similar procedure for the covariance matrix, we obtain:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{z_n}(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top, \tag{7}$$

Finally, we also would like to find the mixing coefficients $\boldsymbol{\pi}$ that maximize the log-likelihood, however in this case we need to ensure that $\sum_{k=1}^K \pi_k = 1$. We do so by using a Lagrange multiplier to account for the later constraint directly in the objective function as:

$$\mathcal{L}(\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right),$$

for which we now take the derivative with respect to π_k that is set to zero, i.e.,

$$0 = -\sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda.$$

If we now multiply both sides of the above expression by π_k and make the sum over k to enforce the constraint $\sum_{k=1}^K \pi_k = 1$, we find that $\lambda = -N$ and:

$$\pi_k = \frac{N_k}{N}. \tag{8}$$

Observation: It is important to notice that Equations 5, 7 and 8 do not lead to a closed-form solution since they all depend on the posterior probability of $z_n = k$ given $\mathbf{x}_n$, $\gamma_k(\mathbf{x}_n)$, which in turn depends on the likelihood and therefore on the parameters we are trying to find.

Expectation-Maximization (EM) algorithm. However, one may alternatively propose an iterative algorithms for finding a solution to the MLE problem, which informally iterates between estimating the posterior probabilities $\{\gamma_k(\mathbf{x}_n)\}_{k=1}^K$ conditioned on the current values of the parameters $\boldsymbol{\pi} = (\pi_1, \dots \pi_k)$, $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, and then maximize the paramteres $\boldsymbol{\pi} = (\pi_1, \dots \pi_k)$, $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ conditioned on the current estimates of $\{\gamma_k(\mathbf{x}_n)\}_{k=1}^K$. This algorithm results indeed in the particularization of the Expectation-Maximization (EM) algorithm for the GMM.

More specifically, one may find an MLE solution for the GMM by the following Algorithm:

1. Initialize the GMM parameters $\boldsymbol{\pi} = (\pi_1, \dots \pi_k)$, $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and evaluate the log-likelihood $\mathscr{L}(\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K)$.
2. **E-step.** Evaluate the posterior probabilities for $z_n = k$ for all k, i.e., $\{\gamma_k(\mathbf{x}_n)\}_{k=1}^K$ according to Eq. 4.
3. **M-step.** For the new value of $\{\gamma_k(\mathbf{x}_n)\}_{k=1}^K$, re-estimate the GMM parameters as:
   (a) Update $\{\boldsymbol{\mu}_k^{new}\}_{k=1}^K$ as in Eq. 5.
   (b) Update $\{\boldsymbol{\Sigma}_k^{new}\}_{k=1}^K$ as in Eq. 7, using the new values for the mean parameters $\{\boldsymbol{\mu}_k^{new}\}_{k=1}^K$.
   (c) Update the probabilities $\{\pi_k^{new}\}_{k=1}^K$ as in Eq. 8.
4. Evaluate the log-likelihood $\mathscr{L}(\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K)$ and check for convergence (in log-likelihood or parameter estimates). If the convergence criterium is not achieved, return to step 2.

3 The E-M algorithm, in general

More in general, the E-M algorithm may be applied to find the MLE solution $\theta_{MLE} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{X}|\boldsymbol{\theta})$, where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ and $\boldsymbol{\theta}$ are the model parameters. To this end, we just need to assume a generative model $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, where $\mathbf{Z}$ is the set of latent variables (e.g. in a mixture model, the component/cluster assignments $\mathbf{Z} = (z_1, \dots, z_N)$), such that the log-likelihood is given by

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

Note though that the set of latent variables $\mathbf{Z}$ is unknown and thus we can only access them through their posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.

Therefore, as in the GMM case, we can simply iterate between the following two steps:

1. **E-step.** Estimate the log-likelihood of some general parameters $\boldsymbol{\theta}$ by taking the expectation with respect to the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ as

   $$\ln p(\mathbf{X}|\boldsymbol{\theta}) \approx \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}),$$

   where $\boldsymbol{\theta}^{old}$ are the parameters from the previous iteration.
2. **M-step.** Update the model parameters via log-likelihood maximization as

   $$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}),$$

   and set $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$.

Obs. 1: Note that the E-M algorithm assumes a tractable E-step.

Obs 2: The E-M algorithm, unless it is already at a local maximum, ensures that at each iteration the log-likelihood is increased.

Obs 3: The E-M algorithm can be also used to find the MAP solution of the model assuming a prior distribution $p(\boldsymbol{\theta})$. In such case, the M-step maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + \ln p(\boldsymbol{\theta})$ (Note: $p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$).

4 The probabilistic modeling pipeline

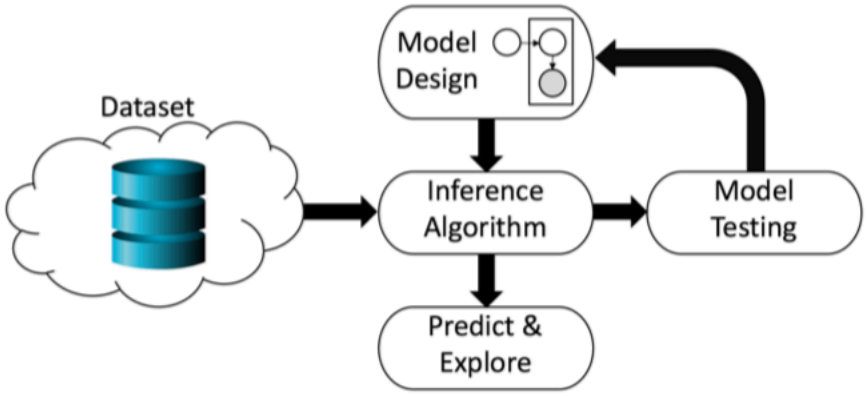

Figure 4: The probabilistic modeling pipeline (Figure by Francisco Rodriguez Ruiz)

The above figure shows the general framework of probabilistic modeling, which allows us to:

- Translate domain prior knowledge into a generative process with hidden and observed variables. As an example, we may assume that our data are sampled according to the following generative model for a GMM:

$$p(\mathbf{X}, \mathbf{Z}, \pi, \{\boldsymbol{\mu}_k\}_{k=1}^K) = Dir(\pi|\alpha) \prod_k \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_n Cat(z_n|\pi) \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_x),$$

  where $\mathbf{X}$ is the observed dataset, and $\mathbf{Z}$ is teh set of local latent variables (one per observation), $\pi = (\pi_1, \dots \pi_k)$ and $\boldsymbol{\mu}$ are the global latent variables, and α, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_x$ are the hyperparameter of the model.

  The generative model introduces our prior knowledge on the latent structure of the observed data (likelihood model and dependencies between observed and latent variables) and on the model latent variables (prior distribution).

- Infer hidden variables by computing (or approximating) the posterior distribution of the latent variables given the observed data, by "reversing" the generative process. In our example, $p(\mathbf{Z}, \pi, \{\boldsymbol{\mu}_k\}_{k=1}^K|\mathbf{Z})$.
- Use the inferred hidden variables (structure) to make predictions, explore the dataset, etc.
- **Separate assumptions (generative model design) from computations (latent variable inference).**

References

C. M. Bishop, *Pattern recognition and machine learning* (Springer, 2006).