Introduction and learning objectives

In Lecture 9, k-means assigned each point to its nearest centre and updated each centre to its group's mean. This can fail when groups have different spreads (amounts of variability around their centres). A Gaussian mixture model (GMM) uses a Gaussian distribution for each group and gives probabilities for the component labels.

By the end of the tutorial, you should be able to:

- Construct a GMM using an unobserved component label and derive its marginal density.
- Derive its likelihood and posterior component probabilities, and relate them to k-means assignments.
- Explain the advantages and limitations of GMMs compared with k-means.

Optional exercise: Derive gradient descent and SGD updates in Problem 3.

1 Gaussian Mixture Models

Problem 1. A probability model for groups. In a GMM, each observation is generated in two steps:

1. Draw a component label $z_n \in \{1, \dots, K\}$ with $p(z_n = k) = \pi_k$.
2. Draw $\mathbf{x}_n$ from $\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for the selected component k.

We observe $\mathbf{x}_n \in \mathbb{R}^D$ but not z_n; z_n is a *latent variable*. The parameters are $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$.

(a) What constraints must the mixing weights $\pi_1, \dots, \pi_K$ satisfy?

Hint: Each π_k is a probability. How are the events $z_n = 1, \dots, z_n = K$ related?

(b) What properties must each covariance matrix $\boldsymbol{\Sigma}_k$ satisfy for the multivariate Gaussian density to be well defined?

Hint: Recall the definition of covariance. For a nonzero vector $\mathbf{v}$, what does $\mathbf{v}^\top \boldsymbol{\Sigma}_k \mathbf{v}$ represent within component k? What matrix operations appear in the Gaussian density?

(c) Show that

$$p(\mathbf{x}_n, z_n = k \mid \theta) = \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$
$$p(\mathbf{x}_n \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Explain why the marginal is a valid probability density.

Hint: Use the product rule and the sum rule of probability.

(d) In Lecture 9's example of groups with different variability, what can $\boldsymbol{\Sigma}_k$ describe that a centre alone cannot? What does π_k describe?

Problem 2. Learning without observed labels. Assume the observations $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ are independent and identically distributed under the GMM. We learn θ by maximum likelihood.

(a) Starting from $p(\mathcal{D} \mid \theta) = \prod_n p(\mathbf{x}_n \mid \theta)$, show that maximum likelihood is equivalent to minimising

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].$$

Why can we not move the logarithm inside the sum over k?

(b) Use Bayes' rule to show that the *responsibility* is

$$\gamma_{nk} := p(z_n = k \mid \mathbf{x}_n, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

Explain why $\sum_k \gamma_{nk} = 1$. How does γ_{nk} differ from the hard assignment $r_{nk} \in \{0, 1\}$ in k-means?

(c) Suppose $\pi_k = 1/K$ and all components have the same fixed covariance $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$. Show that

$$\gamma_{nk} = \frac{\exp\bigl(-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/(2\sigma^2)\bigr)}{\sum_j \exp\bigl(-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/(2\sigma^2)\bigr)}.$$

Which component has the largest responsibility? Compare choosing that component with the k-means assignment step. Assume the nearest centre is unique.

(d) Now allow the mixing weights and covariances to differ. Is the nearest centre always the most probable component? Explain using part (b), and relate your answer to the groups with different variability in Lecture 9.

Optimisation. The mixture likelihood and its derivatives can be evaluated exactly. The logarithm of the sum makes direct analytic maximisation difficult, but we can still use gradient descent.

Optional extension. In part (c), what happens to the responsibilities as $\sigma^2 \to 0$ with the centres fixed? What changes when two centres tie for the smallest distance?

GMMs and k-means

	k-means	**GMM**
Assignments	Assigns each point to one cluster.	Gives a posterior probability for each component.
Group shapes	Works best for compact, roughly spherical groups with similar variability.	Allows different variability; full covariances also allow oriented elliptical groups.
Output	Centres and cluster assignments.	A density, component probabilities and a model for sampling new observations.
Cost	Fewer parameters; usually cheaper to fit.	More parameters, data and computation, especially with full covariances.

Both require a choice of K, suitable features, and initialisation. A GMM's probabilities depend on its Gaussian assumptions; one component cannot describe every possible group shape. When fitting covariances, a component can collapse onto one point and make the likelihood unbounded. We therefore constrain or regularise the covariances.

Training in practice and a first look at EM

Gradient descent and SGD. Minimise $\mathcal{L}$ from Problem 2. Initialise distinct means, for example with k-means centres. Parameterise weights and covariances to satisfy the constraints in Problem 1; an example follows the optional exercise. For each minibatch, evaluate the loss, compute gradients by automatic differentiation, and update the parameters. Each observation's density still uses all K components. Problem 3 derives the mean update.

Expectation–maximization (EM). EM alternates two steps, like k-means:

1. **E-step.** Compute the responsibilities using the current parameters.
2. **M-step.** Hold these responsibilities fixed and fit weighted Gaussians. The mean update is

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_n \gamma_{nk}^{\text{old}} \mathbf{x}_n}{\sum_n \gamma_{nk}^{\text{old}}}.$$

The k-means update has the same weighted-average form, with hard assignments r_{nk} in place of responsibilities. Full EM also updates weights and covariances. Exact EM steps do not decrease the data likelihood, but do not guarantee a global optimum.

Reading. In Bishop's *Pattern Recognition and Machine Learning* (PRML), read Sections 9.2–9.2.1 for GMMs and maximum likelihood, and Section 9.2.2 for EM. Section 9.3.2 explains the k-means connection. Section 9.4 gives the general EM derivation and is optional.

Problem 3 (optional). Learning the means with gradients. Keep $\pi_k > 0$ and the shared covariance $\sigma^2\mathbf{I}$ fixed, and learn only the means. The responsibilities depend on the current means.

For one observation, write

$$\ell_n = -\log s_n, \qquad s_n = \sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \sigma^2\mathbf{I}), \qquad \mathcal{L} = \frac{1}{N}\sum_n \ell_n.$$

(a) Use the chain rule and the identity

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \sigma^2\mathbf{I})\frac{\mathbf{x}_n - \boldsymbol{\mu}_k}{\sigma^2}$$

to show that

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{L} = \frac{1}{N\sigma^2}\sum_n \gamma_{nk}(\boldsymbol{\mu}_k - \mathbf{x}_n).$$

Show your intermediate calculations for $\nabla_{\boldsymbol{\mu}_k} s_n$ and $\nabla_{\boldsymbol{\mu}_k}\ell_n$, and identify the responsibility before averaging over observations.

(b) Substitute part (a) into the gradient descent rule from Tutorial 6 to show that, with step size $\eta > 0$, the update is

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \frac{\eta}{N\sigma^2}\sum_n \gamma_{nk}^{\text{old}}(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}).$$

Here γ_{nk}^{old} is evaluated at the old parameters.

Explain how γ_{nk} weights each point's contribution. Compare with Lecture 9's k-means gradient:

$$\nabla_{\boldsymbol{\mu}_k} J = 2\sum_n r_{nk}(\boldsymbol{\mu}_k - \mathbf{x}_n).$$

(c) Let $\mathcal{B}$ be a uniformly sampled minibatch of B observations. Starting from its average loss $\mathcal{L}_{\mathcal{B}} = B^{-1}\sum_{n\in\mathcal{B}}\ell_n$, derive the update

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \frac{\eta}{B\sigma^2}\sum_{n\in\mathcal{B}} \gamma_{nk}^{\text{old}}(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}).$$

Show the minibatch gradient you use. What changes for $B = 1$ or for the entire dataset? Do we need to sample a component label z_n?

Learning the other parameters. To learn weights and covariances as well, parameterise a diagonal-covariance GMM using unconstrained real numbers a_k and b_{kd}:

$$\pi_k = \frac{e^{a_k}}{\sum_j e^{a_j}}, \qquad \boldsymbol{\Sigma}_k = \operatorname{diag}(v_{k1}, \ldots, v_{kD}), \qquad v_{kd} = \varepsilon + \log(1 + e^{b_{kd}}), \quad \varepsilon > 0.$$

Softmax keeps the weights positive and summing to one. Softplus gives each variance a positive lower bound. Automatic differentiation computes gradients of the same negative log-likelihood with respect to the means, a_k and b_{kd}. Evaluate mixture log-densities using log-sum-exp for numerical stability.