Discussion 4

Note: *Your TA will probably not cover all the problems on this worksheet. The discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, discussions, and homework.*

This Week's Cool AI Demo/Video:
`https://oasis.decart.ai/starting-point`

1 Gaussian Mixture Clustering with Kangaroos and Berkeley Students (1D, Two Gaussians)

Consider a Gaussian Mixture Model (GMM) with two components in one dimension. Let $z_n \in \{K, B\}$ be the latent (unobserved) class of the n-th jumper (K = Kangaroo and B = Berkeley student) and $x_n \in \mathbb{R}$ be the observed jump height (in meters).

Assume that the classes occur with equal probability and that, conditional on the class, jump heights are Gaussian with class-dependent means μ_{K} and μ_{B} (the average jump heights of Kangaroos and Berkeley Students, respectively), and a shared variance σ^2 (the variability in jump heights).

(a) Write down the following quantities explicitly.

- (i) The **prior** probability that a jumper is a Kangaroo or Berkeley student before observing the jump height: $p(z_n)$.
- (ii) The **likelihood** of the jump height conditioned on whether the jumper is a Kangaroo or Berkeley student: $p(x_n \mid z_n)$.
- (iii) The the **marginal** probability that a certain jump height is observed: $p(x_n)$.
- (iv) The **posterior** probability of a jumper being a Kangaroo or Berkeley student after observing the jump height: $p(z_n \mid x_n)$.

Solution:

(i) **Prior:**

$$p(z_n = \text{K}) = p(z_n = \text{B}) = \tfrac{1}{2}$$

(ii) **Conditional:**

$$p(x_n \mid z_n = k) = \mathcal{N}(x_n; \mu_k, \sigma^2)$$

(iii) **Marginal:**

$$p(x_n) = \sum_{k\in\{\mathrm{K},\mathrm{B}\}} p(x_n \mid z_n = k)p(z_n = k) = \tfrac{1}{2}\mathcal{N}(x_n; \mu_{\mathrm{K}}, \sigma^2) + \tfrac{1}{2}\mathcal{N}(x_n; \mu_{\mathrm{B}}, \sigma^2)$$

(iv) **Posterior:**

$$p(z_n = k \mid x_n) = \frac{p(x_n \mid z_n = k)p(z_n = k)}{p(x_n)}$$

(b) Recall the K-means objective:

$$J(\{r_{nk}\}, \{\mu_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$$

where μ_k is the mean of cluster k and $r_{nk} \in \{0, 1\}$ is a binary indicator variable that describes which of the K clusters the data point x_n is assigned to. The K-means algorithm alternates between updating the cluster assignments and updating the cluster centers.

We modify this algorithm to generate GMMs in the following way:

(1) **Update assignments**: Replace the hard assignments r_{nk} with soft assignments γ_{nk}, representing the probability that a point x_n belongs to cluster k.

(2) **Update parameters**: Rather than selecting cluster centers that minimize the squared distances of data points to their assigned cluster centers, select the parameters θ that maximize the Q function:

$$Q(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \ln p(x_n, z_n = k \mid \theta).$$

(i) Identify which probability from part (a) corresponds to γ_{nk}. Verify $\sum_k \gamma_{nk} = 1$.

(ii) Show that the Q function is actually the expected log-likelihood of our observed data, $X = (x_1, \dots, x_N)$, and the (unobserved) latent vector, $Z = (z_1, \dots, z_N)$ under the posterior probability of Z.[1].

(iii) Derive the maximum likelihood estimate for μ_k using the *expected* log-likelihood, $Q(\theta)$, treating γ_{nk} as constants.

Solution:

(i) The probability of assigning point x_n to class k is equivalent to the posterior probability $p(z_n = k|x_n)$. Since $p(z_n = k|x_n)$ is a valid probability distribution, summing over its domain $\{k\}_{k=1}^{K}$ will yield 1.

(ii) Using the probability we identified in part (i), the Q function is

$$Q(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|x_n) \ln p(x_n, z_n = k \mid \theta) \tag{1}$$

[1] We optimize the joint likelihood because it is much more tractable to compute while still optimizing the marginal likelihood. See this page for more information.

$$= \sum_{n=1}^{N} \mathbb{E}_{z_n \sim p(z_n|x_n)}[\ln p(x_n, z_n \mid \theta)] \tag{2}$$

$$= \mathbb{E}_{Z \sim p(Z|X)} \left[\sum_{n=1}^{N} \ln p(x_n, z_n \mid \theta) \right] \tag{3}$$

$$= \mathbb{E}_{Z \sim p(Z|X)} \left[\ln \left(\prod_{n=1}^{N} p(x_n, z_n \mid \theta) \right) \right] \tag{4}$$

$$= \mathbb{E}_{Z \sim p(Z|X)} \Big[\ln p(X, Z \mid \theta) \Big] \tag{5}$$

(iii) To find the MLE estimate $\hat{\mu_k}$, we can take the derivative of the objective function with respect to μ_k and set it equal to 0. Before actually computing the derivative, it's helpful to expand out the probabilities and remove terms that don't directly depend on μ_k:

$$\frac{\partial Q(\theta)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k'=1}^{K} \gamma_{nk'} \cdot \ln p(x_n, z_n = k' | \theta) \tag{6}$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \gamma_{nk} \cdot \ln p(x_n, z_n = k | \theta) \tag{7}$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \gamma_{nk} \cdot \ln \big(p(x_n | z_n = k, \theta) p(z_n = k) \big) \tag{8}$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \gamma_{nk} \cdot (\ln p(x_n | z_n = k, \theta) + \cancel{\ln p(z_n = k)}) \tag{9}$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \gamma_{nk} \cdot (\ln \mathcal{N}(x_n | \mu_k, \sigma^2)) \tag{10}$$

$$= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \gamma_{nk} \cdot - \frac{(x_n - \mu_k)^2}{2\sigma^2} \tag{11}$$

Taking the derivative yields

$$\frac{\partial Q(\theta)}{\partial \mu_k} = \sum_{n=1}^{N} \gamma_{nk} \frac{x_n - \mu_k}{\sigma^2} \tag{12}$$

$$\tag{13}$$

Now we set this derivative equal to zero and solve for μ_k:

$$\sum_{n=1}^{N} \gamma_{nk} \mu_k = \sum_{n=1}^{N} \gamma_{nk} x_n$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{\sum_{n=1}^{N} \gamma_{nk}}$$

Notice this is very similar to how we estimated the cluster centers before, but now our center is a sum of sample points weighted by their posterior probabilities (γ_{nk}).

2 Closed-Form Solution for ℓ_2-Regularized Least Squares (Ridge)

Consider a data matrix $X \in \mathbb{R}^{N\times D+1}$ with rows $x_n^\top$, a target vector $t \in \mathbb{R}^N$, and a parameter vector $w \in \mathbb{R}^{D+1}$. For a regularization hyperparameter $\lambda \geq 0$, define the ridge objective:

$$E(w) \;=\; \|Xw - t\|_2^2 \;+\; \lambda\,\|w\|_2^2.$$

(a) Write $E(w)$ in expanded quadratic form $w^\top Aw - 2b^\top w + c$ by identifying A, b, and c in terms of X, t, and λ.

Solution:

Expand:

$$\|Xw - t\|_2^2 \;=\; (Xw - t)^\top(Xw - t) \;=\; w^\top X^\top Xw - 2t^\top Xw + t^\top t.$$

Thus

$$E(w) \;=\; w^\top X^\top Xw - 2t^\top Xw + t^\top t \;+\; \lambda w^\top w \;=\; w^\top(X^\top X + \lambda I)w \;-\; 2(X^\top t)^\top w \;+\; t^\top t.$$

So $A = X^\top X + \lambda I$, $b = X^\top t$, and $c = t^\top t$.

(b) Compute the gradient $\nabla_w E(w)$ using the identities

$$\nabla_w\|Xw - t\|_2^2 = 2X^\top(Xw - t), \qquad \nabla_w\|w\|_2^2 = 2w.$$

Solution:

$$\nabla_w E(w) \;=\; 2X^\top(Xw - t) + 2\lambda w \;=\; 2\big(X^\top Xw - X^\top t + \lambda w\big) \;=\; 2\big((X^\top X + \lambda I)w - X^\top t\big).$$

(c) Set the gradient to zero and derive the *normal equations* for ridge regression. Solve for w.

Solution:

Stationarity $\nabla_w E(w) = 0$ gives

$$(X^\top X + \lambda I)w \;=\; X^\top t.$$

Assuming $X^\top X + \lambda I$ is invertible (see next part), the unique minimizer is

$$\boxed{\hat{w} \;=\; (X^\top X + \lambda I)^{-1}X^\top t}.$$

(d) Justify why $X^\top X + \lambda I$ is invertible for $\lambda > 0$ (and discuss when it might fail for $\lambda = 0$).

Solution:

For any $v \neq 0$,

$$v^\top(X^\top X + \lambda I)v \;=\; \|Xv\|_2^2 + \lambda\|v\|_2^2.$$

If $\lambda > 0$, this sum is strictly positive, so $X^\top X + \lambda I$ is **positive definite** and therefore invertible. If $\lambda = 0$, invertibility requires $X^\top X$ to be positive definite, i.e., X must have full column rank ($\text{rank}(X) = d$). Otherwise the solution is not unique (the Moore–Penrose pseudoinverse yields the minimum-norm solution).