Introduction to Machine Learning

Project Phase 1

Out-of-Distribution Detection with GMM and EM

Instructor: Dr. S. Amini

Farnoosh Choogani - 402100691

Department of Electrical Engineering

Sharif University of Technology

Winter 2026

Contents

Theoretical Questions

I Question1

The in-distribution data is modeled by a single multivariate Gaussian:

$$p_{ID}(x) = \mathcal{N}(x; \mu, \Sigma).$$

1. Log-likelihood function

The probability density function of a d-dimensional Gaussian is:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Taking the logarithm :

$$\log p_{ID}(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu).$$

2. Likelihood threshold and ellipsoid

We apply likelihood thresholding:

$$\log p_{ID}(x) \geq \tau.$$

$$-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) \geq \tau$$

$$-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) \geq \tau + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|$$

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) \leq -2\tau - d\log(2\pi) - \log|\Sigma|.$$

The left-hand side is the squared Mahalanobis distance. Therefore, the condition $\log p_{ID}(x) \geq \tau$ is equivalent to x lying inside or on an ellipsoid.

3. Equation of the ellipsoid

Let:

$$c = -2\tau - d\log(2\pi) - \log|\Sigma|.$$

Then the ellipsoid is defined by:

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) \leq c,$$

and its boundary is given by:

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = c.$$

II Question 2

1. Log-likelihood is not a sum of individual log-likelihoods

For a Gaussian Mixture Model (GMM), the likelihood of a sample x is:

$$p_{ID}(x; \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k).$$

Taking the logarithm gives:

$$\log p_{ID}(x;\boldsymbol{\Theta}) = \log\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)\right).$$

In general, the logarithm of a sum cannot be written as the sum of logarithms:

$$\log\left(\sum_k a_k\right) \neq \sum_k \log(a_k),$$

because the logarithm is a concave function.

Therefore, the GMM log-likelihood cannot be decomposed into a sum of log-likelihoods of individual components, which makes direct maximization difficult.

2. Jensen's Inequality in Expectation Form

Let $f(\cdot)$ be a concave function and let X be a random variable. Jensen's inequality states:

$$f\big(\mathbb{E}[X]\big) \geq \mathbb{E}[f(X)].$$

Since the logarithm is a concave function, we have:

$$\log\big(\mathbb{E}[X]\big) \geq \mathbb{E}[\log X].$$

Consider the log-likelihood of a Gaussian Mixture Model for a single sample x:

$$\log p(x) = \log\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)\right).$$

Introduce a discrete random variable $Z \in \{1,\ldots,K\}$ with distribution:

$$\mathbb{P}(Z=k) = q_k, \qquad q_k \geq 0, \quad \sum_{k=1}^{K} q_k = 1.$$

Define the random variable:

$$X(Z) = \frac{\pi_Z\mathcal{N}(x;\mu_Z,\Sigma_Z)}{q_Z}.$$

Then its expectation is:

$$\mathbb{E}_{Z\sim q}[X(Z)] = \sum_{k=1}^{K} q_k\frac{\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)}{q_k} = \sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k).$$

Applying Jensen's inequality gives:

$$\log\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)\right) = \log\big(\mathbb{E}_{Z\sim q}[X(Z)]\big) \geq \mathbb{E}_{Z\sim q}[\log X(Z)].$$

Expanding the right-hand side:

$$\mathbb{E}_{Z\sim q}[\log X(Z)] = \sum_{k=1}^{K} q_k\log\left(\frac{\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)}{q_k}\right).$$

Therefore, a lower bound on the log-likelihood is:

$$\boxed{\log\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)\right) \geq \sum_{k=1}^{K} q_k\log\left(\frac{\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)}{q_k}\right)}$$

3. Relation to the EM Algorithm

The lower bound obtained using Jensen's inequality is the objective function that the EM algorithm optimizes.

The log-likelihood of a Gaussian Mixture Model has a log–sum form, which is difficult to maximize directly. By introducing hidden variables that indicate which Gaussian component generated each data point and applying Jensen's inequality, a lower bound on the log-likelihood is obtained.

In the **E-step**, EM computes the responsibilities of each component using the current parameter values. This corresponds to computing the expectation in the lower bound. In the **M-step**, EM updates the model parameters by maximizing this same lower bound.

By repeating these two steps, EM increases (or keeps) the lower bound at each iteration, which guarantees that the true log-likelihood does not decrease.

III Question 3

1. Maximizing the Expected Complete-Data Log-Likelihood

In the M-step of the EM algorithm, the parameters of the Gaussian Mixture Model are updated by maximizing the expected complete-data log-likelihood.

Using the responsibilities

$$\gamma_{ik}^{(t)} = p(z_i = k \mid x_i, \Psi^{(t)}),$$

the objective function can be written as:

$$Q(\Psi, \Psi^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \log p(x_i \mid z_i = k, \Psi).$$

For a Gaussian component, we have:

$$p(x_i \mid z_i = k, \Psi) = \mathcal{N}(x_i; \mu_k, \Sigma_k).$$

Taking the logarithm, only the quadratic term depends on the mean μ_k. Therefore, when maximizing with respect to μ_k, the objective reduces to:

$$Q_k(\mu_k) = -\frac{1}{2} \sum_{i=1}^{n} \gamma_{ik}^{(t)} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k).$$

Maximizing this expression with respect to μ_k leads to the update rule for the mean.

2. Gradient with Respect to the Mean

To find the update rule for the mean μ_k, we take the gradient of the objective function $Q_k(\mu_k)$ with respect to μ_k and set it equal to zero.

From Part 1, the objective function is:

$$Q_k(\mu_k) = -\frac{1}{2} \sum_{i=1}^{n} \gamma_{ik}^{(t)} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k).$$

Taking the gradient with respect to μ_k gives:

$$Q_k(\mu_k) = -\frac{1}{2} \sum_{i=1}^{n} \gamma_{ik}^{(t)} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k)$$

$$\nabla_{\mu_k} \Big[(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \Big] = -2\Sigma_k^{-1} (x_i - \mu_k)$$

$$\nabla_{\mu_k} Q_k(\mu_k) = -\frac{1}{2} \sum_{i=1}^{n} \gamma_{ik}^{(t)} \left(-2\Sigma_k^{-1}(x_i - \mu_k)\right)$$

$$\nabla_{\mu_k} Q_k(\mu_k) = \sum_{i=1}^{n} \gamma_{ik}^{(t)} \Sigma_k^{-1}(x_i - \mu_k)$$

Setting the gradient equal to zero:

$$\sum_{i=1}^{n} \gamma_{ik}^{(t)} \Sigma_k^{-1}(x_i - \mu_k) = 0.$$

Since Σ_k^{-1} is invertible, this simplifies to:

$$\sum_{i=1}^{n} \gamma_{ik}^{(t)} (x_i - \mu_k) = 0.$$

Expanding the terms:

$$\sum_{i=1}^{n} \gamma_{ik}^{(t)} x_i - \mu_k \sum_{i=1}^{n} \gamma_{ik}^{(t)} = 0.$$

Solving for μ_k, we obtain:

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} x_i, \quad \text{where } N_k^{(t)} = \sum_{i=1}^{n} \gamma_{ik}^{(t)}.$$

3. Interpretation of the Mean Update

The mean update obtained in the M-step is:

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \gamma_{ik}^{(t)} x_i, \quad N_k^{(t)} = \sum_{i=1}^{n} \gamma_{ik}^{(t)}.$$

This expression has the same form as a weighted average:

$$\mu = \frac{\sum_i w_i x_i}{\sum_i w_i},$$

where the weights are given by $w_i = \gamma_{ik}^{(t)}$.

Each responsibility $\gamma_{ik}^{(t)}$ represents how strongly data point x_i belongs to component k. Points with larger responsibilities contribute more to the mean, while points with small responsibilities have little effect. Hence, the mean update corresponds to a weighted average of the data points.

IV Question 4

1: Expected Log-Likelihood for an Isotropic Gaussian

Assume the in-distribution data are drawn from a d-dimensional Gaussian distribution with isotropic covariance $\Sigma = \sigma^2 I$. The log-likelihood of a sample x is:

$$\log p_{ID}(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu).$$

$|\Sigma| = (\sigma^2)^d$ and $\Sigma^{-1} = \frac{1}{\sigma^2} I$:

$$\log p_{ID}(x) = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|x-\mu\|^2.$$

Taking expectation over $x \sim \mathcal{N}(\mu, \sigma^2 I)$ and using $\mathbb{E}[\|x-\mu\|^2] = d\sigma^2$, we obtain:

$$\mathbb{E}[\log p_{ID}(x)] = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{d}{2}.$$

2: Why an OOD Sample Can Have Higher Likelihood

From the expression of the log-likelihood,

$$\log p_{ID}(x) = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|x-\mu\|^2,$$

it is clear that likelihood depends only on the squared distance $\|x-\mu\|^2$ from the mean.

In high-dimensional spaces, typical in-distribution samples are not close to the mean, but instead concentrate on a thin shell at distance approximately $\sqrt{d}\sigma$ from the mean(because $\mathbb{E}[\|x-\mu\|^2] = d\sigma^2$). Therefore, a typical in-distribution sample does not maximize likelihood.

It is possible for an out-of-distribution sample with regular or structured features to have a smaller squared distance to the mean than a typical in-distribution sample. In such cases,

$$\log p_{ID}(x_{\text{OOD}}) > \log p_{ID}(x_{\text{ID}}),$$

even though x_{OOD} does not come from the in-distribution.

3: Why Likelihood Alone Is Not Reliable for OOD Detection

The above analysis shows that likelihood measures only pointwise density and does not account for how probability mass is distributed across the space.

In high-dimensional settings, most probability mass lies in regions with moderate density but large volume, while regions near the mean have high density but very small volume. Likelihood-based methods ignore this concentration-of-measure effect.

As a result, out-of-distribution samples can receive higher likelihood than typical in-distribution samples. This demonstrates that likelihood alone is not a reliable criterion for out-of-distribution detection, especially in high-dimensional spaces.