__/__/____

(40) Gaussian Mixture Models (Soft Clustering)

A GMM assumes that the data is generated from a mixture (combination) of several Gaussian (Normal) distributions ~ each representing a cluster or latent component

→ Each point is generated probabilistically by one of several Gaussian components. (clusters)

∴ GMM = soft clustering using probabilities.

We assume, $P(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$

- K: no of Gaussian components (clusters)
- π_k: mixing coeff (weight) of component k, such that $\sum_k \pi_k = 1$ and $\pi_k \geq 0$
- $\mathcal{N}(x \mid \mu_k, \Sigma_k)$: Gaussian distribution for component k.
- μ_k: mean vector
- Σ_k: covariance matrix.

→ To fit a Gaussian Mixture Model to the data we use Expectation-Maximization (EM) algorithm which is an iterative method that optimize the parameters of the Gaussian Distribution like mean, covariance and mixing methods.

Works in two steps:

1. Expectation Step (E-Step): In this step the algorithm calculates the probability that each data point belongs to each cluster based on current parameter estimates

2. Maximization Step (M-Step): After estimating the probabilities the algorithm updates the parameters to better fit the data.

~ These 2 are repeated until the model converges.

Steps:

1. Start with initial guesses for means, covariances and mixing coeffs of each Gaussian Distribution.
2. (E - Step): For each datapoint, calculate the probability of it belonging to each cluster.
3. (M - Step): Update the parameters using the probabilities calculated in the E - step.
4. Repeat until the log - likelihood of the data (a measure of how well the model fits the data) converge.

- Covariance (Σ): The Cov Matrix describes the shape, size & orientation of the cluster. Unlike simpler clustering methods like k-Means which assume spherical clusters, the covariance allows Gaussian components of fitted depending on r/s b/w features.

- GMM is a probabilistic generalization of k-Means.

→ Monotonicity of EM : EM never decreases log - likelihood.

At each iteration: $Q(\theta, \theta^{(t)}) = E_Z(X, \theta^{(t)}[\log p(x, z|\theta)]$

M - Step maximizes Q
which guarentees: $\log p(x|\theta^{(t+1)}) \geq \log p(x|\theta^{(t)})$

So EM always move uphill - though it may get stuck in a local maximum, not necessarily the global one.

→ Singularity Problem In EM : In mixture models, the likelihood can become infinite
A Gaussian components mean = exactly one data point
Its variance = 0
Likelihood $\to \infty$ (since $N(x_i | \mu_k, \Sigma_k) \to \infty$)
Hence MLE is ill - posed.

__/__/____

Fixes: Regularization

Bayesian Priors over parameters (MAP estimation)

Early stopping

→ Choosing the Number of Components (k)

We can't increase k just like that - more components always increase likelihood.

So use penalized model selection criteria.

- AIC (Akaike Information Criterion)

↳ $2p - 2\log L$: Penalizes complexity mildly

- BIC (Bayesian Information Criterion)

↳ $p \log N - 2\log L$: Penalizes complexity strongly.

Choose K minimizing AIC & BIC

→ Certain Relations

1. K-Means: Special case of EM where covariance = identity, and hard assignments

EM → GMM (soft clustering)

2. HMM (Hidden Markov Model): Sequence model learned via EM-like algorithm, EM generalizes here too.