Follow Torch the triceratops, our new class mascot on instagram: @torchthetriceratops

1 One Last Definition

Gaussian Mixture Models (GMM)

1. **GMM**: Probabilistic models used for clustering data. The algorithm assumes that the data is generated by a mixture of K Gaussian distributions, where K is a hyperparameter. GMM works by iteratively estimating the parameters of the Gaussian distributions and the weights of the mixture components using the Expectation-Maximization (EM) algorithm.

2. **EM**: An iterative method used for estimating the parameters of statistical models.

   Let Z be a multinomial random variable with components $z_1, z_2, \dots, z_k$, where each component is 0 or 1 i.e. $P(z_j = 1)$ is the probability that a point comes from the Gaussian distribution j.

   Let $\theta = \mu_1, \mu_2, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k, \pi_1, \dots, \pi_k$, where $\pi_j = P(z_j = 1)$.

   The likelihood is $\prod_{i=1}^{N} P(x_i \mid \theta)$.
   Hence, the log-likelihood is $l = \sum_{i=1}^{N} \log P(x_i \mid \theta) = \sum_{i=1}^{m} \log \sum_{j=1}^{k} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$.

   E-Step: Calculate $P(z_j = 1 \mid x_i, \theta)\ \forall i, j$.

$$P(z_j = 1 \mid x_i, \theta)$$
$$= \frac{p(x_i \mid z_j = 1, \mu_j, \Sigma_j) p(z_j = 1 \mid \pi_j)}{p(x_i \mid \theta)}$$
$$= \frac{\mathcal{N}(x_i \mid \mu_j, \Sigma_j)\pi_j}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)}$$

   M-Step: Apply MLE and update the parameters $\pi_j, \mu_j, \Sigma_j\ \forall j$.
   Let's find the MLE for μ_j.

$$\frac{\partial l}{\partial \mu_j} \sum_{i=1}^{N} \log \sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)$$
$$= \sum_{i=1}^{N} \frac{1}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)} \frac{\partial l}{\partial \mu_j} \sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)$$
$$= \sum_{i=1}^{N} \frac{1}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)} \frac{\partial l}{\partial \mu_j} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$$
$$= \sum_{i=1}^{N} \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)} \frac{\partial l}{\partial \mu_j} \frac{(x_i - \mu_j)^2}{2\Sigma_j}$$
$$= \sum_{i=1}^{N} P(z_j = 1 \mid x_i, \theta) \Sigma_j^{-1} (x_i - \mu_j)$$

   We can set this to 0, and solve for μ_j to get $\mu_j = \frac{\sum_{i=1}^{N} P(z_j=1|x_i,\theta) x_i}{\sum_{i=1}^{N} P(z_j=1|x_i,\theta)}$.

   We can do similar calculations for the other two parameters π_j and Σ_j.
   $\pi_j = \frac{\sum_{i=1}^{N} P(z_j=1|x_i,\theta)}{N}$
   $\Sigma_j = \frac{\sum_{i=1}^{N} P(z_j=1|x_i,\theta)(x_i-\mu_j)(x_i-\mu_j)^\top}{\sum_{i=1}^{N} P(z_j=1|x_i,\theta)}$

2 Gaussian Mixture Models

2.1 GMM vs K-means

What is the key difference between mixture modeling and K-means?

K-means is hard assignment (each point belongs to only one cluster) while mixture modeling is soft assignment (calculates probability that a point belongs to a cluster).

In GMM, we aim to maximize our likelihood. That is, $\arg\max_\theta \prod_{i=1}^N P(x_i \mid \theta)$.

$$\begin{aligned}\arg\max_\theta \prod_{i=1}^N P(x_i \mid \theta) &= \arg\max_\theta \prod_{i=1}^N \sum_{j=1}^k P(x_i, z_i = j \mid \theta) \\ &= \arg\max_\theta \prod_{i=1}^N \sum_{j=1}^k P(z_i = j) P(x_i \mid z_i = j, \theta)\end{aligned}$$

What happens to this expression if we assume a hard-assignment? Simplify using the assumption.

Hard assignment means that $P(z_i = j) = 1$ if the point belongs to the jth cluster.

Thus we have:
$\arg\max_\theta \prod_{i=1}^N \sum_{j=1}^k P(z_i = j) P(x_i \mid z_i = j, \theta)$

Our points are from a gaussian distribution, thus we have:
$\arg\max_\theta \prod_{i=1}^N \sum_{j=1}^k P(z_i = j) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2} ||x_i - \mu_i||_2^2)$
$= \arg\max_\theta \prod_{i=1}^N \exp(\frac{-1}{2\sigma^2} ||x_i - \mu_i||_2^2)$

Taking the log, we get
$= \arg\max_\theta \log \prod_{i=1}^N \exp(\frac{-1}{2\sigma^2} ||x_i - \mu_i||_2^2)$
$= \arg\max_\theta \sum_{i=1}^N \log \exp(\frac{-1}{2\sigma^2} ||x_i - \mu_i||_2^2)$
$= \arg\max_\theta \sum_{i=1}^N \frac{-1}{2\sigma^2} ||x_i - \mu_i||_2^2$
$= \arg\min_\mu \sum_{i=1}^N ||x_i - \mu_i||_2^2$

This looks familiar....it's K-means!