GAUSSIAN MIXTURES

(z, x) jointly distributed, x categorical and latent (unobservable)

$$p(x) = \sum_{k=1}^{K} p(z, x) = \sum_{k=1}^{K} p(z = k) p(x|z = k)$$

Example:

$$z \sim \text{Categorical}(\pi_1, \dots, \pi_K) \quad \text{where} \quad \sum \pi_k = 1,$$
$$x|z = k \sim N(\mu_k, \sigma_k^2)$$

(1)
$$p(x) = \sum_{k=1}^{K} \pi_k G(x|\mu_k, \sigma_k^2)$$

To do: Find maximum likelihood estimates of π_k, μ_k, σ_k^2.

Suppose θ is a parameter of $p(x)$.

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log p(x) &= \frac{\partial}{\partial\theta} \log \left(\sum_{k=1}^{N} p(k, x) \right) \\
&= \frac{\sum_{k=1}^{N} \frac{\partial}{\partial\theta} p(k, x)}{\sum_{j=1}^{K} p(j, x)} \\
&= \frac{\sum_{k=1}^{K} p(k, x) \frac{\partial}{\partial\theta} \log p(k, x)}{\sum_{j=1}^{K} p(j, x)} \\
&= \sum_{k=1}^{K} \left(\frac{p(k, x)}{\sum_{j=1}^{K} p(j, x)} \right) \frac{\partial}{\partial\theta} \log p(k, x) \\
&= \sum_{k=1}^{K} p(k|x) \frac{\partial}{\partial\theta} \big(\log p(k) + \log p(x|k) \big)
\end{aligned}$$

With $p(x)$ as in (1) and $k \in \{1, \dots, K\}$,

$$\begin{aligned}
\frac{\partial}{\partial\mu_k} \log p(x) &= p(k|x) \frac{x - \mu_k}{\sigma_k^2} \\
\frac{\partial}{\partial\sigma_k^2} \log p(x) &= p(k|x) \frac{1}{2\sigma_k^2} \left(\frac{(x - \mu_k)^2}{\sigma_k^2} - 1 \right)
\end{aligned}$$

Let $(z^{(1)}, x^{(1)}), \ldots, (z^{(n)}, x^{(n)})$ be a random sample. Set

$$r_k^{(i)} = p(k|x^{(i)}).$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log\prod_{i=1}^n p(x^{(i)}) &= \sum_{i=1}^n \frac{\partial}{\partial\theta}\log p(x^{(i)}) \\ &= \sum_{i=1}^n\sum_{k=1}^K r_k^{(i)}\frac{\partial}{\partial\theta}\big(\log p(k) + \log p(x^{(i)}|k)\big)\end{aligned}$$

With $p(x)$ as in (1) and $k \in \{1, \ldots, K\}$,

$$\begin{aligned}\frac{\partial}{\partial\mu_k}\log\prod_{i=1}^n p(x^{(i)}) &= \sum_{i=1}^n r_k^{(i)}\frac{x^{(i)}-\mu_k}{\sigma_k^2} \\ &= \frac{1}{\sigma_k^2}\left(\sum_{i=1}^n r_k^{(i)}x^{(i)} - \mu_k\sum_{i=1}^n r_k^{(i)}\right) \\ \frac{\partial}{\partial\sigma_k^2}\prod_{i=1}^n \log p(x^{(i)}) &= \frac{1}{2(\sigma_k^2)^2}\left(\sum_{i=1}^n r_k^{(i)}(x^{(i)}-\mu_k)^2 - \sigma_k^2\sum_{i=1}^n r_k^{(i)}\right)\end{aligned}$$

Setting these expressions equal to zero and solving, we get

$$\widehat{\mu}_k = \frac{\sum_{i=1}^n r_k^{(i)}x^{(i)}}{\sum_{i=1}^n r_k^{(i)}}, \quad \widehat{\sigma}_k^2 = \frac{\sum_{i=1}^n r_k^{(i)}(x^{(i)}-\widehat{\mu}_k)^2}{\sum_{i=1}^n r_k^{(i)}}.$$

Let

$$I_k = \{i : z^{(i)} = k\}, \quad n_k := \big|\{i : z^{(i)} = k\}\big|.$$

Then

$$\pi_k \approx \frac{n_k}{n} \approx \frac{1}{n}\sum_{i=1}^n p(k|x^{(i)}) = \frac{1}{n}\sum_{i=1}^n r_k^{(i)}$$

Set

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^n \widehat{r}_k^{(i)}$$

As

$$r_k^{(i)} = p(k|x^{(i)}) = \frac{p(k)p(x^{(i)}|k)}{\sum_{j=1}^K p(j)p(x^{(i)}|j)} = \frac{\pi_k p(x^{(i)}|k)}{\sum_{j=1}^K \pi_j p(x^{(i)}|j)},$$

we set

$$\widehat{r}_k^{(i)} = \frac{\widehat{\pi}_k\widehat{p}(x^{(i)}|k)}{\sum_{j=1}^K \widehat{\pi}_j\widehat{p}(x^{(i)}|j)}$$

1. The EM algorithm

Choose initial approximations $\pi_{k,0}$, $\mu_{k,0}$, $\sigma_{k,0}$. Set $\theta_{k,0} = (\mu_{k,0}, \sigma_{k,0})$.

Set

$$r_{k,0}^{(i)} = \frac{\pi_{k,0} G(x^{(i)} \mid \theta_{k,0})}{\sum_{j=1}^K \pi_{j,0} G(x^{(i)} \mid \theta_{k,0})}$$

For $t \geq 1$:

Update parameters:

$$\begin{aligned}
\pi_{k,t} &= \frac{1}{n} \sum_{i=1}^n \widehat{r}_{k,t-1}^{(i)}, \\
\mu_{k,t} &= \frac{\sum_{i=1}^n r_{k,t-1}^{(i)} x^{(i)}}{\sum_{i=1}^n r_{k,t-1}^{(i)}}, \\
\sigma_{k,t}^2 &= \frac{\sum_{i=1}^n r_{k,t-1}^{(i)} (x^{(i)} - \mu_{k,t})^2}{\sum_{i=1}^n r_{k,t-1}^{(i)}}.
\end{aligned}$$

Then update responsibilities:

$$r_{k,t}^{(i)} = \frac{\pi_{k,t} G(x^{(i)} \mid \theta_{k,t})}{\sum_{j=1}^K \pi_{j,t} G(x^{(i)} \mid \theta_{k,t})}$$

2. K-means clustering

$x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$

If $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ and, for $1 \leq k \leq K$, let

$$r_k^{(i)} = r_k^{(i)}(\mu_1, \ldots, \mu_K) = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_\ell \|x^{(i)} - \mu_\ell\|, \\ 0 & \text{otherwise.} \end{cases}$$

When $r_k^{(i)} = 1$, we say that μ_k *takes responsibility for* $x^{(i)}$.

The *cluster of* μ_k, written C_k or $C(\mu_k)$, is the set of all $x^{(i)}$ for which μ_k takes responsibility:

$$C(\mu_k) = \{x^{(i)} : r_k^{(i)} = 1\}.$$

The *covariance of clusters* C_k *and* C_ℓ is

$$\operatorname{Cov}(C_k, C_\ell) = \sum_{x^{(i)} \in C_k} \sum_{x^{(j)} \in C_\ell} \|x^{(i)} - x^{(j)}\|^2 = \sum_{i=1}^K \sum_{j=1}^K r_k^{(i)} r_\ell^{(j)} \|x^{(i)} - x^{(j)}\|^2$$

If $k = \ell$, we call this the *variance of* C_k:

$$\operatorname{Var} C_k = \sum_{x^{(i)} \in C_k} \sum_{x^{(j)} \in C_k} \|x^{(i)} - x^{(j)}\|^2 = \sum_{i=1}^K \sum_{j=1}^K r_k^{(i)} r_k^{(j)} \|x^{(i)} - x^{(j)}\|^2$$

The clustering problem: Given $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ and $K > 0$, find $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ minimizing

$$\sum_{k=1}^{K} \operatorname{Var} C(\mu_k)$$