# Lecture 5

## Mixture models

$$p(\vec{y}|\vec{\theta}) = \sum_{k=1}^{K} \pi_k \underbrace{p_k(\vec{y})}_{\text{prob. distribution of the } k^{th} \text{ mixture component}},$$

$$\begin{cases} 0 \le \pi_k \le 1, \\ \sum_{k=1}^{K} \pi_k = 1. \end{cases}$$

One can define $p_k(\vec{y}) = p(\vec{y}|\vec{\theta}_k) = p(\vec{y}|z=k, \vec{\theta})$, where ($\vec{\theta}_k$: model prms)

$z = \{1, 2, \dots, K\}$ is a latent variable which labels mixture components and

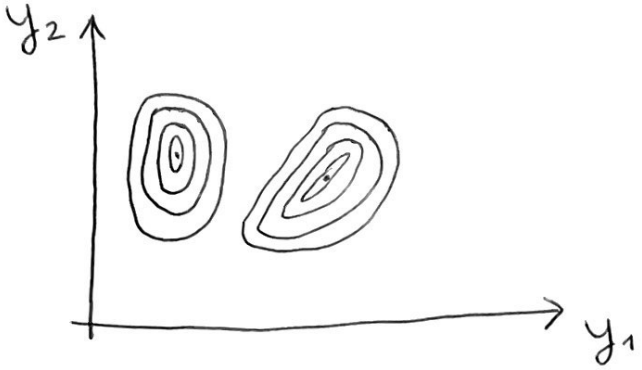
$\vec{\theta} = \{\pi_1, \dots, \pi_K, \vec{\theta}_1, \dots, \vec{\theta}_K\}$ is a vector of all model prms.

Now, $p(\vec{y}|\vec{\theta}) = \sum_{k=1}^{K} p(\vec{y}|z=k, \vec{\theta}) \underbrace{p(z=k|\vec{\theta})}_{\text{Cat}(z|\vec{\theta}) = \pi_k} \Rightarrow$

$\Rightarrow \sum_k \pi_k p_k(\vec{y})$, as above

### Gaussian mixture models (GMM):

$$p(\vec{y}|\vec{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{y}|\vec{\mu}_k, \Sigma_k)$$




Mixture of 2 gaussians in 2D

Bernoulli mixture models (BMM):

If $y_i = \{0, 1\}$ and $\vec{y} = \{y_1, y_2, \ldots, y_D\}$ is a set of D binary observations,

$$\begin{cases} p(\vec{y} \mid z = k, \vec{\theta}) = \prod_{d=1}^{D} \text{Ber}(y_d \mid \mu_{d,k}) = \\ \qquad = \prod_{d=1}^{D} \mu_{d,k}^{y_d} (1 - \mu_{d,k})^{1 - y_d}, \\ p(z = k \mid \vec{\theta}) = \text{Cat}(z = k \mid \vec{\theta}) = \pi_k \text{ as above.} \end{cases}$$

Here, $\vec{\theta} = \{\pi_1, \ldots, \pi_K, \vec{\mu}_1, \ldots, \vec{\mu}_K\}$, where

each $\vec{\mu}_j = \{\mu_{1,j}, \mu_{2,j}, \ldots, \mu_{D,j}\}$

$j = 1, \ldots, K$ is a D-dim vector

BMMs can be used to fit black-and-white images, e.g. 0-9 digits in the MNIST dataset.

# Statistics

**Model training**: estimate $\vec{\theta}$ (model prms) from $D$ (data)

Maximum likelihood estimate (MLE):

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \underbrace{p(D|\vec{\theta})}_{\text{likelihood}}$$

One commonly assumes in supervised learning

$$p(D|\vec{\theta}) = \prod_{n=1}^{\tilde{N}} p(\vec{y}_n | \vec{x}_n, \vec{\theta})$$

($\vec{y}_n$: target variables; $\vec{x}_n$: input pattern)

Log-likelihood: $\mathcal{L}(\vec{\theta}) = \log p(D|\vec{\theta}) =$

$$= \sum_{n=1}^{\tilde{N}} \log p(\vec{y}_n | \vec{x}_n, \vec{\theta}).$$

Connection between MLE & MAP (max posterior prob.):

$$\vec{\theta}_{MAP} = \arg\max_{\vec{\theta}} [\log \underbrace{p(\vec{\theta}|D)}_{\text{posterior prob.}}] \ (=)$$

$$(=) \arg\max_{\vec{\theta}} [\log \underbrace{p(D|\vec{\theta})}_{\text{likelihood}} + \log \underbrace{p(\vec{\theta})}_{\text{prior}}]$$

as $N \uparrow$, $\log p(D|\vec{\theta})$ grows and

dominates $\log p(\vec{\theta})$, leading to

$$\vec{\theta}_{MAP} \to \vec{\theta}^*.$$

This is equiv. to $p(\vec{\theta}|\mathcal{D}) = \delta(\vec{\theta} - \vec{\theta}_{MAP})$.

---

For unsupervised learning, we consider $\mathcal{L}(\vec{\theta}) = \sum_{n=1}^{N} \log p(\vec{y}_n|\vec{\theta})$,

$$\vec{\theta}^* = \underset{\vec{\theta}}{\text{argmax}}\ \mathcal{L}(\vec{\theta}), \text{ etc.}$$

---

Now, consider the empirical distribution of the data: (unsupervised case)

$$p_{\mathcal{D}}(\vec{y}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\vec{y} - \vec{y}_n)$$

Define Kullback-Leibler (KL) divergence:

$$D_{KL}(p||q) = \sum_{\vec{y}} p(\vec{y}) \log \frac{p(\vec{y})}{q(\vec{y})}$$

$D_{KL}(p||q) \geq 0$, $= 0$ iff $p = q$.

If $\begin{cases} q(\vec{y}) \to p(\vec{y}|\vec{\theta}), \\ p(\vec{y}) \to p_{\mathcal{D}}(\vec{y}) \end{cases}$ $\Rightarrow D_{KL}(p||q) \ominus$

$$\ominus \underbrace{\sum_{\vec{y}} p_{\mathcal{D}} \log p_{\mathcal{D}}}_{\text{-entropy of } p_{\mathcal{D}}, \text{ const}(\vec{\theta})} - \underbrace{\sum_{\vec{y}} p_{\mathcal{D}} \log p(\vec{y}|\vec{\theta})}_{\text{- cross-entropy between } p_{\mathcal{D}} \text{ \& } p(\vec{y}|\vec{\theta})} \diamondsuit$$

$\Leftrightarrow \text{const}(\vec{\theta}) - \frac{1}{N} \underbrace{\sum_{n=1}^{N} \log p(\vec{y}_n | \vec{\theta})}_{\mathcal{I}(\vec{\theta})}$.

Thus, maximizing $\mathcal{I}(\vec{\theta})$ is equiv. to minimizing the KL divergence.

---

Likewise, consider

$$p_{\mathcal{D}}(\vec{x}, \vec{y}) = p_{\mathcal{D}}(\vec{y}|\vec{x}) p_{\mathcal{D}}(\vec{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\vec{x}-\vec{x}_n)\delta(\vec{y}-\vec{y}_n)$$

for supervised learning

Then $E_{p_{\mathcal{D}}(\vec{x})}[D_{KL}(p_{\mathcal{D}}(\vec{y}|\vec{x}) \| p(\vec{y}|\vec{x}, \vec{\theta}))] =$

$$= \sum_{\vec{x}} p_{\mathcal{D}}(\vec{x}) \left[ \sum_{\vec{y}} p_{\mathcal{D}}(\vec{y}|\vec{x}) \log \frac{p_{\mathcal{D}}(\vec{y}|\vec{x})}{p(\vec{y}|\vec{x}, \vec{\theta})} \right] =$$

$$= \text{const}(\vec{\theta}) - \sum_{\vec{x}, \vec{y}} p_{\mathcal{D}}(\vec{x}, \vec{y}) \log p(\vec{y}|\vec{x}, \vec{\theta}) =$$

$$= \text{const}(\vec{\theta}) - \frac{1}{N} \underbrace{\sum_{n=1}^{N} \log p(\vec{y}_n | \vec{x}_n, \vec{\theta})}_{\mathcal{I}(\vec{\theta})}.$$

Thus, maximizing $\mathcal{I}(\vec{\theta})$ is equiv. to minimizing the <u>average</u> KL divergence.

<u>MLE, Bernoulli</u>:

$\mathcal{I}(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta)$, where

$y_n \in \{0,1\}$, $\quad n = 1, \ldots, N$ $\qquad \theta = p(y=1)$.

Then $\mathcal{L}(\vec{\theta}) = \sum_{n=1}^{N} \log\left[\theta^{\mathbb{I}(y_n=1)}(1-\theta)^{\mathbb{I}(y_n=0)}\right] =$

$$= \Big(\underbrace{\sum_{n=1}^{N} \mathbb{I}(y_n=1)}_{N_1}\Big)\log\theta + \Big(\underbrace{\sum_{n=1}^{N} \mathbb{I}(y_n=0)}_{N_0}\Big)\log(1-\theta).$$

$N_0 + N_1 = N$

Finally, $\left.\frac{d}{d\theta}\mathcal{L}(\vec{\theta})\right|_{\theta^*} = \frac{N_1}{\theta^*} - \frac{N_0}{1-\theta^*} = 0$ yields

$$N_1(1-\theta^*) - N_0\theta^* = 0,$$

$\theta^* = \frac{N_1}{N_0+N_1}$, fraction of $y_n=1$ observations

MLE, categorical:

$$\mathcal{L}(\vec{\theta}) = \sum_{k=1}^{K} N_k \log\theta_k,$$

$$\sum_{k=1}^{K} N_k = N; \quad \sum_{k=1}^{K} \theta_k = 1.$$

Consider $\mathcal{L}(\vec{\theta}, \lambda) = \sum_k N_k \log\theta_k + \lambda\left(1 - \sum_k \theta_k\right)$.

Then $\left.\frac{\partial\mathcal{L}}{\partial\theta_i}\right|_{\theta_i^*} = \frac{N_i}{\theta_i^*} - \lambda = 0 \Rightarrow N_i = \lambda\theta_i^*$.

Find $\lambda$ by normalization:

$$\sum_i N_i = N = \lambda \underbrace{\sum_i \theta_i^*}_{=1} \Rightarrow \lambda = N.$$

Thus, $\theta_i^* = \frac{N_i}{N}$, $i = 1, \ldots, K$.

fraction of event $i$ occurring

MLE, 1D gaussian:

$$Y \sim N(\mu, \sigma^2)$$

$$\mathcal{D} = \{y_1, \ldots, y_N\}$$

$$\mathcal{L}(\mu, \sigma^2) = \sum_{n=1}^{N} \log\left[(2\pi\sigma^2)^{-1/2} e^{-\frac{(y_n-\mu)^2}{2\sigma^2}}\right] =$$

$$= N\left(-\frac{1}{2}\right)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_n (y_n-\mu)^2.$$

Consider $\begin{cases} \frac{\partial}{\partial\mu}\mathcal{L}(\mu, \sigma^2) = 0, \\ \frac{\partial}{\partial\sigma^2}\mathcal{L}(\mu, \sigma^2) = 0. \end{cases}$

we have $\frac{\partial}{\partial\mu}\sum_n (y_n-\mu)^2\Big|_{\mu^*} = 2(N\mu^* - \sum_n y_n) = 0$, or

$$\left[\mu^* = \frac{1}{N}\sum_n y_n \equiv \bar{y} \Leftarrow \text{sample average}\right]$$

$$\frac{N}{2}\frac{1}{2\pi\sigma_*^2}2\pi - \frac{1}{2(\sigma_*^2)^2}\sum_n (y_n - \bar{y})^2 = 0,$$

"$\mu^*$" or

$$\frac{1}{\sigma_*^2}\sum_n (y_n - \bar{y})^2 = N, \quad \Leftarrow \text{sample variance}$$

$$\left[\sigma_*^2 = \frac{1}{N}\sum_n (y_n - \bar{y})^2\right]$$

## MLE, multi-D Gaussian:

Consider

$$\mathcal{L}(\vec{\mu}, \Lambda) = \frac{N}{2} \log |\Lambda| - \frac{1}{2} \sum_{n=1}^{N} (\vec{y}_n - \vec{\mu})^T \Lambda (\vec{y}_n - \vec{\mu}) + \text{const}(\vec{\mu}, \Lambda),$$

where $\Lambda = \Sigma^{-1}$.

↑ precision matrix

One can show that

$$\begin{cases} \frac{\partial}{\partial \mu} \mathcal{L}(\vec{\mu}, \Lambda) = 0 \Rightarrow \vec{\mu}^* = \frac{1}{N} \sum_n \vec{y}_n = \bar{\vec{y}}, \\ \frac{\partial}{\partial \Lambda} \mathcal{L}(\vec{\mu}, \Lambda) = 0 \Rightarrow \Sigma^* = \frac{1}{N} \sum_n (\vec{y}_n - \bar{\vec{y}})(y_n - \bar{\vec{y}})^T. \end{cases}$$

$\Sigma^* = \Lambda^{*-1}$