# CS280 Fall 2021 Assignment 1
# Part A

ML Background

October 17, 2021

**Name:**

**Student ID:2021233240**

**1. MLE (5 points)**

Given a dataset $\mathcal{D} = \{x_1, \cdots, x_n\}$. Let $p_{emp}(x)$ be the empirical distribution, i.e., $p_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i)$ where $\delta(x, a)$ is the Dirac delta function[1] centered at $a$. Assume $q(x|\theta)$ be some probabilistic model.

- Show that $\arg\min_q KL(p_{emp}||q)$ is obtained by $q(x) = q(x; \hat{\theta})$, where $\hat{\theta}$ is the Maximum Likelihood Estimator and $KL(p||q) = \int p(x)(\log p(x) - \log q(x))dx$ is the KL divergence.

# Solution 1

$$\begin{aligned}
KL(p_{emp}||q) &= \int p_{emp}(x)(\log p_{emp}(x) - \log q(x))dx && (1)\\
&= \int p_{emp}(x)\log p_{emp}(x)dx - \int p_{emp}(x)\log q(x)dx && (2)\\
&= E_{p_{emp}(x)}(\log p_{emp}(x)) - \int p_{emp}(x)\log q(x)dx && (3)
\end{aligned}$$

Since $p_{emp}(x)$ is the empirical distribution, $E_{p_{emp}(x)}(\log p_{emp}(x))$ is a constant. Here, we let it be $C$.

And, because $p_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i)$ where $\delta(x, x_i)$ is the Dirac delta function, we can get $p_{emp}(x_i) = \frac{1}{n}$.

So the equation

$$KL(p_{emp}||q) = \frac{1}{n}\sum_{i=1}^{n}(-\log q(x)) + C \tag{4}$$

According to the Maximum Likelihood Estimator, we need to maximize:

$$\begin{aligned}
F(q) &= \log\prod_{i=1}^{n} q(x_i) && (5)\\
&= \sum_{i=1}^{n}(\log q(x_i)) && (6)
\end{aligned}$$

So, we can observe equation $(4)$ and $(6)$ and get the $q*$:

$$q* = \arg\min_q KL(p_{emp}||q) = \arg\max_q F(q)$$

[1] https://en.wikipedia.org/wiki/Dirac_delta_function

**2. Gradient descent for fitting GMM (10 points)**

Consider the Gaussian mixture model

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\pi_j \geq 0, \sum_{j=1}^{K} \pi_j = 1$. (Assume $\mathbf{x}, \boldsymbol{\mu}_k \in \mathbb{R}^d, \boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$)

Define the log likelihood as

$$l(\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)$$

Denote the posterior responsibility that cluster $k$ has for datapoint $n$ as follows:

$$r_{nk} := p(z_n = k|\mathbf{x}_n, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

- Show that the gradient of the log-likelihood wrt $\boldsymbol{\mu}_k$ is

$$\frac{d}{d\boldsymbol{\mu}_k} l(\theta) = \sum_n r_{nk} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

- Derive the gradient of the log-likelihood wrt $\pi_k$ without considering any constraint on $\pi_k$. (bonus 2 points: with constraint $\sum_k \pi_k = 1$.)

# Solution 2

## 1.

According to the chain rule:

$$\frac{dl}{d\boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{dl_n}{dp(\mathbf{x}_n|\theta)} \cdot \frac{dp(\mathbf{x}_n|\theta)}{d\boldsymbol{\mu}_k} \tag{7}$$

$$= \sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n|\theta)} \cdot \pi_k \cdot \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \frac{d(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k) \sum_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T)}{d\boldsymbol{\mu}_k} \tag{8}$$

$$= \sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n|\theta)} \cdot \pi_k \cdot \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \sum_{\boldsymbol{k}}^{\boldsymbol{-1}} (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{9}$$

$$= \sum_n r_{nk} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{10}$$

## 2.

According to the chain rule:

$$\frac{dl}{d\boldsymbol{\pi}_k} = \sum_{n=1}^{N} \frac{dl_n}{dp(\mathbf{x}_n|\theta)} \cdot \frac{dp(\mathbf{x}_n|\theta)}{d\boldsymbol{\pi}_k} \tag{11}$$

$$= \sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n|\theta)} \cdot \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{12}$$

$$= \sum_{n} \frac{r_{nk}}{\boldsymbol{\pi}_k} \tag{13}$$

When we consider the constraint $\sum_k \pi_k = 1$, We can change the form of $\pi_k$ to remove the restriction: let $\pi_k = \frac{e^{w_k}}{\sum_g^G e^{w_g}}$. Then we can get the gradient of $w_k$:

$$\frac{dl}{dw_k} = \sum_{n=1}^{N} \frac{dl_n}{dp(\mathbf{x}_n|\theta)} \cdot \sum_{j=1}^{K} (\frac{dp(\mathbf{x}_n|\theta)}{d\boldsymbol{\pi}_j} \cdot \frac{d\boldsymbol{\pi}_j}{dw_k}) \tag{14}$$

$$= \sum_{n=1}^{N} \frac{1}{p(\mathbf{x}_n|\theta)} \cdot \sum_{j=1}^{K} (p(\mathbf{x}_n|\theta)_j \cdot \frac{d\boldsymbol{\pi}_j}{dw_k}) \tag{15}$$

if $j = k$,

$$\frac{d\pi_j}{dw_k} = \frac{e^{w_k} \sum_g e^{w_g} - e^{w_k} e^{w_k}}{(\sum_g e^{w_g})^2} \tag{16}$$

$$= \pi_k(1 - \pi_k) \tag{17}$$

if $j \neq k$,

$$\frac{d\pi_j}{dw_k} = \frac{-e^{w_k} e^{w_j}}{(\sum_g e^{w_g})^2} \tag{18}$$

$$= -\pi_j \pi_k \tag{19}$$

So:

$$\frac{dl}{dw_k} = \sum_{n=1}^{N} \frac{dl_n}{dp(\mathbf{x}_n|\theta)} \cdot (p(\mathbf{x}_n|\theta)_k \cdot \pi_k(1 - \pi_k) - \sum_{j \neq k} p(\mathbf{x}_n|\theta)_j \cdot \pi_j \pi_k) \tag{20}$$

$$= \sum_{n=1}^{N} \frac{\pi_k}{p(\mathbf{x}_n|\theta)} \cdot (p(\mathbf{x}_n|\theta)_k - p(\mathbf{x}_n|\theta)) \tag{21}$$

$$= \sum_{n=1}^{N} (r_{nk} - \pi_k) \tag{22}$$

And when the $\frac{dl}{dw_k} = 0$, the parameter becomes the optimal:

$$\frac{dl}{dw_k} = \sum_{n=1}^{N} (r_{nk} - \pi_k) = 0 \tag{23}$$

And:

$$\pi_k = \frac{1}{N} \cdot \sum_{n=1}^{N} r_{nk} \tag{24}$$