CS280 Fall 2022 Assignment 1
Part A

ML Background

October 16, 2022

Name: Dai ZiJia

Student ID:2022233158

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.

Proof.

$$\begin{aligned}
\arg\min_\theta KL(p_{emp}||q) &= arg\min_\theta \int (p_{emp}\log p_{emp} - p_{emp}\log q)dx \\
&= arg\min_\theta (\int (p_{emp}\log p_{emp})dx - \int (p_{emp}\log q)dx) \\
&= arg\min_\theta -\int (p_{emp}\log q)dx \\
&= arg\max_\theta \int (p_{emp}\log q)dx \\
&= arg\max_\theta \int \frac{1}{n}\sum[\log q(x_i|\theta)\delta(x, x_i)]dx \\
&= arg\max_\theta \frac{1}{n}\sum \log q(x_i|\theta)\int \delta(x, x_i)dx \\
&= arg\max_\theta \frac{1}{n}\sum \log q(x_i|\theta)
\end{aligned}$$

[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 π_k without considering any constraint on π_k. (bonus 2 points: with constraint $\sum_k \pi_k = 1$.)

$$\begin{aligned}
r_{nk} &= \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k'} \pi_{k'} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}\right)} \\
\sum_n r_{nk} &= \frac{\pi_k \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k'} \pi_{k'} \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}\right)} \\
\frac{dl(\theta)}{d\boldsymbol{\mu}_k} &= \frac{dl(\theta)}{dp(x_n|\theta)} \cdot \frac{dp(x_n|\theta)}{d\mathcal{N}(x_n|\boldsymbol{\mu}_k, \sum_k)} \cdot \frac{d\mathcal{N}(x_n|\boldsymbol{\mu}_k, \sum_k)}{d\boldsymbol{\mu}_k} \\
&= \frac{\pi_k \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k'} \pi_{k'} \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}\right)} \frac{d}{d\boldsymbol{\mu}_k} \ln\left[\pi_k \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right] \\
&= \frac{\pi_k \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k'} \pi_{k'} \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}\right)} \left[\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x_n} - \boldsymbol{\mu}_k\right)\right] \\
&= \sum_n r_{nk} \left[\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x_n} - \boldsymbol{\mu}_k\right)\right]
\end{aligned}$$