CS280 Fall 2021 Assignment 1
Part A

ML Background

September 25, 2021

Name: 樊雨鑫

Student ID: 2020233246

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.

$$KL(p\|q) = \int p(x)(\log p(x) - \log q(x))\,dx$$

$$= \int p(x)\log p(x)\,dx - \int p(x)\log q(x)\,dx$$

$$KL(p_{emp}(x)\|q) = \int p_{emp}(x)\log p_{emp}(x)\,dx - \int p_{emp}(x)\,q(x)\,dx$$

$$\arg\min_q KL(p_{emp}(x)\|q) = \arg\min_q\left(\int p_{emp}(x)\log p_{emp}(x)\,dx - \int p_{emp}(x)\log q(x)\,dx\right)$$

$$= \int p_{emp}(x)\log p_{emp}(x)\,dx - \arg\max_q \int p_{emp}(x)\log q(x)\,dx$$

$$= \int p_{emp}(x)\log p_{emp}(x)\,dx - \int p_{emp}(x)\log \arg\max_\theta q(x)\,dx$$

$\hat{\theta}$ is the maximum likelihood estimator.

$$\hat{\theta} = \arg\max_\theta q(x)$$

So, $\arg\min_q KL(p_{emp}(x)\|q)$ is obtained by $q(x) = q(x;\hat{\theta})$.

[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 $\sum_k \pi_k = 1$.)

① $\frac{dl(\theta)}{dp(x_n|\theta)} = \sum_{n=1}^{N} \frac{1}{P(x_n|\theta)}$ $\frac{dp(x_n|\theta)}{d\mu_k} = \frac{d\sum_{k=1}^{K} \pi_{k'} N(x_n|\mu_{k'}, \Sigma_{k'})}{d\mu_k} = \sum_{k'=1}^{K} \pi_{k'} \frac{dN(x_n|\mu_{k'}, \Sigma_{k'})}{d\mu_k}$

$= \pi_k \frac{dN(x_n|\mu_k, \Sigma_k)}{d\mu_k}$

$\frac{dN(x_n|\mu_k, \Sigma_k)}{d\mu_k} = N(x_n|\mu_k, \Sigma_k)\Sigma_k^{-1}(x_n - \mu_k)$

$\frac{dl(\theta)}{d\mu_k} = \sum_{n=1}^{N} \frac{1}{P(x_n|\theta)} \frac{dP(x_n|\theta)}{d\mu_k} = \sum_{n=1}^{N} \frac{1}{P(x_n|\theta)} \cdot \pi_k N(x_n|\mu_k, \Sigma_k)\Sigma_k^{-1}(x_n - \mu_k) = \sum_n r_{nk}\Sigma_k^{-1}(x_n - \mu_k)$

② $\frac{dl(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{1}{P(x_n|\theta)} \frac{dP(x_n|\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} N(x_n|\mu_{k'}, \Sigma_{k'})}$

when $\sum_k \pi_k = 1$, $\frac{d\pi_{k'}}{d\pi_k} = -1$ if $k \neq k'$. $\frac{dp(x_n|\theta)}{d\pi_k} = N(x_n|\mu_k, \Sigma_k) - \sum_{k' \neq k} N(x_n|\mu_{k'}, \Sigma_{k'})$

So:

$\frac{dl(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k) - \sum_{k' \neq k} N(x_n|\mu_{k'}, \Sigma_{k'})}{\sum_{k'} \pi_{k'} N(x_n|\mu_{k'}, \Sigma_{k'})}$ when $\sum_k \pi_k = 1$.