CS280 Fall 2021 Assignment 1
Part A

ML Background

September 25, 2021

Name: 袁润洋

Student ID: 48768008

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.

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

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

Since the dataset $D = \{x_1, \cdots, x_n\}$ is given, $p(x)$ is a fixed function.

Then $\int p(x)\log p(x)$ is a constant, named C

$$\Rightarrow KL(p\|q) = C - \int p(x)\log q(x;\hat{\theta})dx$$

$$= C - \int \frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i)\log q(x;\hat{\theta})dx$$

$$= C - \frac{1}{n}\sum_{i=1}^{n}\int \delta(x, x_i)\log q(x;\hat{\theta})dx$$

By the nature of dirac delta function, for any $f(x)$

$$\int \delta(x, a)f(x)dx = f(a)$$

Then $$KL(p\|q) = C - \frac{1}{n}\sum_{i=1}^{n}\int \delta(x, x_i)\log q(x;\hat{\theta})dx$$

$$= C - \frac{1}{n}\sum_{i=1}^{n}\log(x_i;\hat{\theta})$$

$$\operatorname{argmin}_{\hat{\theta}} KL(p\|q(x;\hat{\theta})) = \operatorname{argmin}_{\hat{\theta}} C - \frac{1}{n}\sum_{i=1}^{n}\log(x_i;\hat{\theta})$$

$$= \operatorname{argmax}_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n}\log(x_i;\hat{\theta})$$

$$= \hat{\theta}_{MLE}$$

[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 $\sum_k \pi_k = 1$.)

$$\frac{d}{d\mu_k} l(\theta) = \sum_{n=1}^{N} \frac{\frac{dp(x_n|\theta)}{d\mu_k}}{p(x_n|\theta)}$$

$$= \sum_{n=1}^{N} \frac{\frac{dp(x_n|\theta)}{d\mu_k}}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$$

$$\frac{dp(x_n|\theta)}{d\mu_k} = \frac{d\, \pi_k \frac{e^{-\frac{(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)}{2}}}{(\sqrt{2\pi})^d |\Sigma_k|^{\frac{1}{2}}}}{d\mu_k}$$

$$= \frac{d\, -\frac{(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)}{2}}{d\mu_k} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

$$= \frac{2\Sigma_k^{-1}(x_n-\mu_k)}{2} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

$$= \Sigma_k^{-1}(x_n-\mu_k)\, \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

Then $\frac{d}{d\mu_k} l(\theta) = \sum_{n=1}^{N} \frac{\Sigma_k^{-1}(x_n - \mu_k)\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)}$

Because $r_{nk} = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} N(x|\mu_{k'}, \Sigma_{k'})}$ $\Rightarrow \frac{d}{d\mu_k} l(\theta) = \sum_{n=1}^{N} r_{nk} \Sigma_k^{-1}(x_n - \mu_k)$

2. $\frac{d\, l(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{\frac{dp(x_n|\theta)}{d\pi_k}}{p(x_n|\theta)}$

$$= \sum_{n=1}^{N} \frac{\frac{d\pi_k N(x_n|\mu_k, \Sigma_k)}{d\pi_k}}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)}$$

$$= \sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_{k'}, \Sigma_{k'})}$$

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

Bonus: When there is a constraint $\sum_k \pi_k = 1$, the MLE problem is

$$\underset{\theta, \pi}{\text{argmax}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)$$

$$\text{s.t.} \quad \sum_k \pi_k = 1$$

Applying Lagrange multiplier method, the problem is transformed to

$$\underset{\theta, \pi}{\text{argmax}}\, L(\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) + \lambda\left(\sum_k \pi_k - 1\right)$$

Then $\frac{d\, L(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} + \lambda$

MLE requires the gradient at the optimal solution is 0

Then for all $k = 1, \cdots, K$,

$\sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} + \lambda = 0 \Rightarrow \sum_{n=1}^{N} r_{nk} + \pi_k \lambda = 0 \Rightarrow \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} + \sum_{k=1}^{K} \pi_k \lambda = 0 \Rightarrow \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} + \sum_{k=1}^{K} \pi_k \lambda = 0$

Because $\sum_{k=1}^{K} r_{nk} = 1$ and $\sum_{k=1}^{K} \pi_k = 1$, the above equation is $\sum_{n=1}^{N} 1 + \lambda = 0 \Rightarrow \lambda = -N$

Bring $\lambda = -N$ into $\sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} + \lambda = 0 \Rightarrow \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} = N$

Then $\frac{d\, l(\theta)}{d\mu_k} = N$ with constraint $\sum_k \pi_k = 1$