CS280 Fall 2018 Assignment 1
Part A

ML Background

October 23, 2020

Name:

Student ID:

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:
The key is to show that the two objectives have the same solution.
Let $L(q) \doteq \text{KL}(p_{emp}||q)$ where q is a shorthand for $q(x; \theta)$.

$$
\begin{aligned}
L(q) &= \int p_{emp}(x)\left[\log p_{emp}(x) - \log q(x)\right] && (1)\\
&= \left(\int_x p_{emp}(x)\left[-\log q(x)\right]\right) - \mathcal{H}(p_{emp}(x)) && (2)
\end{aligned}
$$

1

where $\mathcal{H}(p_{emp}(x)) = \mathbb{E}_{p_{emp}(x)}\left[-\log p_{emp}(x)\right]$ is constant wrt q.
Since $\delta(\cdot)$ is the discrete delta function, we have $p_{emp}(x_i) = \frac{1}{n}, \forall i \in [1, n]$. (Or generally, do manipulation using the property $\int_x \delta(x, x_0)f(x) = f(x_0)$.)
It follows

$$
L(q) = \frac{1}{n}\sum_{i=1}^{n} -\log q(x_i) + C
$$

if the answer does not explicitly state how the dirac delta is manipulated, at least 1 point is not given 2

where $C = -\mathcal{H}(p_{emp}(x))$.
Recall that for MLE, we are maximizing:

$$
\begin{aligned}
J(q) &= \log\prod_{i=1}^{n} q(x_i) && (3)\\
&= \sum_{i=1}^{n}\log q(x_i) && (4)
\end{aligned}
$$

1

Obviously, the solution

$$
q* = q(x; \hat{\theta}) = \max_q J(q) = \min_q L(q)
$$

1

is also the solution of minimizing the KL divergence.

[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 $\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\phi})$, $\phi_k \geq 0$, $\sum_{j=1}^{k} \phi_j = 1$. (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 μ_k is

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

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

Solution: Using the chain rule we can obtain the gradients.
For the gradient wrt $\boldsymbol{\mu_k}$:

$$\frac{dl}{d\boldsymbol{\mu_k}} = \sum_n \frac{dl_n}{dP_n} \frac{dP_n}{d\boldsymbol{\mu_k}}$$

where $P_n \doteq p(\mathbf{x_n}|\theta)$ and

$$\frac{dl_n}{dP_n} = \frac{1}{P_n}$$

Let $P_{nk} = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ such that $P_n \doteq p(\mathbf{x_n}|\theta) = \sum_{k=1}^{K} \pi_k P_{nk}$:

$$\frac{dP_n}{d\boldsymbol{\mu_k}} = \pi_k \frac{dP_{nk}}{d\boldsymbol{\mu_k}}$$

where

$$\frac{dP_{nk}}{d\boldsymbol{\mu_k}} = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \frac{d\left(-\frac{1}{2}(\mathbf{x_n} - \boldsymbol{\mu_k})\boldsymbol{\Sigma_k^{-1}}(\mathbf{x_n} - \boldsymbol{\mu_k})^\intercal\right)}{d\boldsymbol{\mu_k}} \tag{5}$$

$$= P_{nk} \boldsymbol{\Sigma_k^{-1}} (\mathbf{x_n} - \boldsymbol{\mu_k}) \tag{6}$$

Combining the above, we have

$$\frac{dl(\theta)}{d\boldsymbol{\mu_k}} = \sum_n \frac{1}{P_n} \cdot \pi_k \cdot P_k \cdot \boldsymbol{\Sigma_k^{-1}} (\mathbf{x_n} - \boldsymbol{\mu_k}) = \sum_n r_{nk} \boldsymbol{\Sigma_k^{-1}} (\mathbf{x_n} - \boldsymbol{\mu_k})$$

5 points or 0 point

For the gradient wrt π_k:

$$\frac{dl}{d\pi_k} = \sum_n \frac{dl_n}{dP_n}\frac{dP_n}{d\pi_k} \tag{7}$$

$$= \sum_n \frac{1}{P_n}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{8}$$

$$= \sum_n \frac{r_{nk}}{\pi_k} \quad \text{2 points} \tag{9}$$

To handle the constraint on π_k, the key is how to convert a constrained optimization to unconstrained optimization.

1. Use the reparameterization trick. 5 pts in total
   Let $\pi_k \doteq \frac{e^{w_k}}{\sum_{k'}^K e^{w_{k'}}}$ where $\sum_k^K \pi_k = 1$ but $w_k \in \mathbb{R}$ is unconstrained.
   Then find the derivative wrt w_k.

$$\frac{dl}{dw_k} = \sum_{n=1}^N \left(\frac{dl_n}{dP_n} \cdot \sum_{j=1}^K \frac{dP_n}{d\pi_j} \cdot \frac{d\pi_j}{dw_k}\right)$$

where

$$\frac{dl_n}{dP_n} = \frac{1}{P_n}, \frac{dP_n}{d\pi_j} = P_{nj} \quad 1$$

For $\frac{d\pi_j}{dw_k}$:
If $j = k$,

$$\frac{d\pi_j}{dw_k} = \frac{e^{w_k} \cdot \sum_{k'} e^{w_{k'}} - e^{w_k} \cdot e^{w_k}}{(\sum_{k'} e^{w_{k'}})^2} = \pi_k(1 - \pi_k)$$

else,

$$\frac{d\pi_j}{dw_k} = \frac{0 - e^{w_k} \cdot e^{w_j}}{(\sum_{k'} e^{w_{k'}})^2} = -\pi_j\pi_k \quad 2$$

Hence:

$$\frac{dl}{dw_k} = \sum_{n=1}^N \frac{1}{P_n} \cdot \left(P_{nk} \cdot \pi_k(1 - \pi_k) - \sum_{j \neq k} P_{nj} \cdot \pi_j\pi_k\right) \tag{10}$$

$$= \sum_{n=1}^N \frac{\pi_k}{P_n} \cdot \left(P_{nk} - \sum_j P_{nj} \cdot \pi_j\right) \tag{11}$$

$$= \sum_{n=1}^N \frac{\pi_k}{P_n} \cdot (P_{nk} - P_n) \tag{12}$$

$$= \sum_{n=1}^N r_{nk} - \pi_k \quad 1 \tag{13}$$

On the one hand, for gradient descent, we can use the above gradient to obtain

$$w'_k = w_k - \alpha\frac{dl}{dw_k}$$

and then plug back into $\pi_k \doteq \frac{e^{w_k}}{\sum_{k'}^{K} e^{w_{k'}}}$ to obtain π_k'.

On the other hand, we know the optimal solution to π_k by setting the above gradient to zero:

$$\sum_{n=1}^{N} r_{nk} - \pi_k = 0 \tag{14}$$

$$\Rightarrow \pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \tag{15}$$

1

2. Use the Lagrangian duality.

   The Lagrangian function is

$$L(\pi_k) = l(\theta; \pi_k) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

   where λ is the dual variable.

   Due to the KKT condition, set $\frac{dL}{d\pi_k} = 0$:

$$\frac{dL}{d\pi_k} = \frac{dl}{d\pi_k} + \lambda = 0 \Rightarrow \sum_n \frac{r_{nk}}{\pi_k} + \lambda = 0 \tag{16}$$

1

   which should also satisfy:

$$\sum_{k=1}^{K} \pi_k - 1 = 0 \tag{17}$$

1

   Rearrange (16) we get

$$\pi_k = -\frac{1}{\lambda} \sum_n r_{nk} \tag{18}$$

   Sum over k on both sides to apply (17):

$$1 = \sum_k \pi_k = -\frac{1}{\lambda} \sum_k \sum_n r_{nk} = -\frac{1}{\lambda} \sum_n \sum_k r_{nk} = -\frac{1}{\lambda} \cdot N \tag{19}$$

   We get

$$\lambda = -N$$

1

   Plug back into (18), we get the optimal solution to π_k:

$$\pi_k = \frac{1}{N} \sum_n r_{nk}$$

1

Note that using lagragian duality we are not dealing with gradient descent for GMM, so there are at most 4 points for this solution