CS280 Fall 2018 Assignment 1
Part A

ML Background

Due in class, October 12, 2018

Name:Zhijuan Hu

Student ID:67856754

1. MLE (5 points)

Given a dataset $\mathcal{D} = \{x_1, \cdots, x_n\}$. Let $p_{emp}(x)$ be the empirical distribution, i.e., $p_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i)$ and let $q(x|\theta)$ be some 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.

Since when $n \to \infty$, the samples $q(x_i|\theta)$ will close to the empirical distribution $p_{emp}(x)$

$$\begin{aligned}
&\min_q \quad KL(p_{emp}||q) \\
&= \min_q \quad \int p_{emp}[log(p_{emp}) - logq(x)]dx \\
&= \min_q \quad -\int p_{emp}logq(x)dx \\
&= \max_q \quad \int p_{emp}logq(x)dx \\
&= \max_q E[logq(x)]
\end{aligned}$$

where Maximum Likelihood Estimator $\hat{\theta} = \arg max_q E[logq(x)]$

■

2. Properties of l_2 regularized logistic regression (10 points)

Consider minimizing

$$J(\mathbf{w}) = -\frac{1}{|D|}\sum_{i \in D} \log \sigma(y_i \mathbf{x}_i^T \mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$$

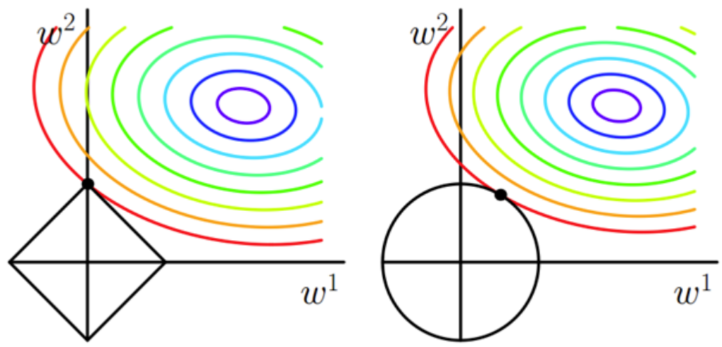
where $y_i \in -1, +1$. Answer the following true/false questions and **explain why**.

- $J(\mathbf{w})$ has multiple locally optimal solutions: T/F?
- Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is sparse (has many zeros entries): T/F?

Proof.

- False. Since $J(w)$ is convex (where $-log\sigma(y_i \mathbf{x}_i^T \mathbf{w})$, $\|\mathbf{w}\|_2^2$ is convex).

- False. Since l_2 is more smooth. From the Figure1 below, we can find that l_2 will prefer select more features. And for those features which are close to origin, l_2 norm will make them close to 0 not equal to 0 like l_1 norm.



(a) **ℓ_1-ball meets quadratic function.** ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

(b) **ℓ_2-ball meets quadratic function.** ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes.

Figure 1: Comparison between l_1 norm and l_2 norm

∎

3. Gradient descent for fitting GMM (15 points)

Consider the Gaussian mixture model

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

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:

$$r_{nk} := p(z_n = k|\mathbf{x}_n, \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_k k')}$$

- Show that the gradient of the log-likelihood wrt μ_k is

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

- Derive the gradient of the log-likelihood wrt π_k without considering any constraint on π_k. (bonus: with constraint $\sum_k \pi_k = 1$.)
- Derive the gradient of the log-likelihood wrt Σ_k without considering any constraint on Σ_k. (bonus: with constraint Σ_k be a symmetric positive definite matrix.)

  Proof.

  Suppose there are K Gauss distribution, and x is a sample which obeys multi-Gauss distribution. Denoted the probability x_i fall into model k as:

$$p(\mathbf{x_i}|z_k) = \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

  and

$$\begin{aligned} p(\mathbf{x}_i|\theta) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \\ &= \sum_{k=1}^{K} p(z_i = k) p(\mathbf{x}_i|z_i = k) \end{aligned}$$

  – Since

$$\begin{aligned} \nabla_{\mu_k} p(\mathbf{x_n}|\theta) &= \pi_k \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x_n} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x_n} - \mu_k)\} \Sigma^{-1}(\mathbf{x_n} - \mu_k) \\ &= \pi_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k) \Sigma_k^{-1}(\mathbf{x_n} - \mu_k) \end{aligned}$$

  we have

$$\begin{aligned} \frac{d}{d\mu_k} l(\theta) &= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k) \Sigma_k^{-1}(\mathbf{x_n} - \mu_k) \\ &= \sum_n r_{nk} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k) \end{aligned}$$

- Consider the MLE problem:

$$\max l(\theta)$$
$$s.t. \sum_{k=1}^{K} \pi_k = 1$$

Using Lagrange Multiplier method, construct Lagrange function:

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

Since

$$\frac{p(\mathbf{x}_n|\theta)}{d\pi_k} = \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \tag{1}$$
$$\frac{\mathcal{L}(\pi_k)}{d\pi_k} = \frac{d}{d\pi_k} l(\theta) - \lambda = 0 \tag{2}$$

we have

$$\begin{aligned}
\frac{d}{d\pi_k} l(\theta) &= \frac{d}{d\pi_k} \sum_{i=1}^{n} logp(\mathbf{x}_n|\theta) \\
&= \sum_{i=1}^{n} \frac{1}{p(\mathbf{x}_n|\theta)} \frac{dp(\mathbf{x}_n|\theta)}{d\pi_k} \\
&= \sum_{i=1}^{n} \frac{1}{p(\mathbf{x}_n|\theta)} \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \\
&= \sum_{i=1}^{n} \frac{r_{nk}}{\pi_k}
\end{aligned}$$

Using (2) and $\sum_{k=1}^{K} \pi_k = 1$, we have

$$\begin{aligned}
\lambda &= \sum_{i=1}^{n} \frac{r_{nk}}{\pi_k} \\
&= N
\end{aligned}$$

- From the equation

$$\frac{\partial log(f(x))}{\partial x} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial x} \tag{3}$$
$$\Rightarrow \quad \frac{\partial f(x)}{\partial x} = f(x) \frac{\partial log(f(x))}{\partial x} \tag{4}$$

Using (4), we have

$$
\begin{aligned}
\frac{d}{d\Sigma_k}p(\mathbf{x_i}|\theta) &= \frac{d}{d\Sigma_k}\sum_{i=1}^{k} p(z_k)p(\mathbf{x}_i|z_k) && (5)\\
&= \frac{d}{d\Sigma_k}p(z_k)p(\mathbf{x}_i|z_k) && (6)\\
&= p(z_k)\frac{d}{d\Sigma_k}p(\mathbf{x}_i|z_k) && (7)\\
&= p(z_k)p(\mathbf{x}_i|z_k)\frac{dlog}{d\Sigma_k}p(\mathbf{x}_i|\mu_k,\Sigma_k) && (8)
\end{aligned}
$$

Since

$$
\begin{aligned}
tr(ABC) &= tr(CAB) && (9)\\
\frac{dtr(BA)}{dA} &= B^T && (10)\\
\frac{dtr(A^{-1}B)}{dA} &= -(A^{-1})^T B (A^{-1})^T && (11)
\end{aligned}
$$

For

$$
\frac{dlog}{d\Sigma_k}p(\mathbf{x}_i|\mu_k,\Sigma_k) = \frac{dlog}{d\Sigma_k}\left[\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_k|^{\frac{1}{2}}}exp\{-\frac{1}{2}(\mathbf{x_n}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x_n}-\mu_k)\}\Sigma^{-1}(\mathbf{x_n}-\mu_k)\right] \quad (12)
$$

$$
= \frac{d}{d\Sigma_k}\left[-\frac{1}{2}log(|\Sigma_k|)-\frac{1}{2}(\mathbf{x_n}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x_n}-\mu_k)\}\Sigma^{-1}(\mathbf{x_n}-\mu_k)\right] \quad (13)
$$

$$
= (-\frac{1}{2})\left[\Sigma_k^{-1}-\Sigma_k^{-1}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^T\Sigma_k^{-1}\right] \quad (14)
$$

Hence, combine (8) and (14), equation (5) has

$$
\begin{aligned}
\frac{d}{d\Sigma_k}p(\mathbf{x_i}|\theta) &= p(z_k)p(\mathbf{x}_i|z_k)(-\frac{1}{2})\left[\Sigma_k^{-1}-\Sigma_k^{-1}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^T\Sigma_k^{-1}\right] && (15)\\
&= \pi_k\mathcal{N}(\mathbf{x}_i|\mu_k,\Sigma_k)(-\frac{1}{2})\left[\Sigma_k^{-1}-\Sigma_k^{-1}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^T\Sigma_k^{-1}\right] && (16)
\end{aligned}
$$

At last, since

$$
\begin{aligned}
\frac{d}{d\Sigma_k}l(\theta) &= \frac{d}{d\Sigma_k}\sum_{i=1}^{N} logp(\mathbf{x_i}|\theta) && (17)\\
&= \frac{1}{p(\mathbf{x}|\theta)}\frac{d}{d\Sigma_k}p(\mathbf{x}_k|\theta) && (18)\\
&= \frac{\pi_k\mathcal{N}(\mathbf{x}_i|\mu_k,\Sigma_k)}{\sum_{i=1}^{k}\pi_k\mathcal{N}(\mathbf{x}_i|\mu_k,\Sigma_k)}(-\frac{1}{2})\left[\Sigma_k^{-1}-\Sigma_k^{-1}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)^T\Sigma_k^{-1}\right] && (19)
\end{aligned}
$$

■