CS280 Fall 2018 Assignment 1
Part A

ML Background

Due in class, October 12, 2018

Name:MinJie

Student ID:10109867

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.

Solution:
Give $KL(p||q) = \int p(x)(\log p(x) - \log q(x))dx$ and $p_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i)$

$$
\begin{aligned}
\text{KL(p}_{emp}||q) &= \int p_{emp}(x)(\log p_{emp}(x) - \log q(x))dx \\
&= \int p_{emp}(x)\log p_{emp}(x)dx - \int p_{emp}(x)\log q(x))dx \\
&= \int p_{emp}(x)\log p_{emp}(x)dx - \int (\frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i))\log q(x)dx \\
&= \int p_{emp}(x)\log p_{emp}(x)dx - \sum_{x\in D}(\frac{1}{n}\sum_{i=1}^{n}\delta(x, x_i))\log q(x) \\
&= \int p_{emp}(x)\log p_{emp}(x)dx - \frac{1}{n}\sum_{i=1}^{n}\log q(x_i|\theta)
\end{aligned}
$$

$program@e$

We want to argmin q on $KL(p_{emp}||q)$, from above we know that we can transform to calcutate the argmax q on $\frac{1}{n}\sum_{i=1}^{n}\log q(x_i|\theta)$

$\frac{1}{n}\sum_{i=1}^{n}\log q(x_i|\theta)$ is the log likelihood estimator representation of D, and we know $\hat{\theta}$ is the Maximum Likelihood Estimator, so $\frac{1}{n}\sum_{i=1}^{n}\log q(x_i|\hat{\theta})$ can minimize the whole equation $KL(p_{emp}||q)$.

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?

Question1:

$J(\mathbf{w})$ has multiple locally optimal solutions $\Longrightarrow$ False

The most convenient way to determine this problem is to test the convexity of $J(\mathbf{w})$, and to determine the convexity of $J(\mathbf{w})$, we should test whether the Hessian matrix is positive definite.

$y_i \in -1, +1$, we know $\sigma(x) = \frac{1}{1+e^{-x}}$

So Frist derivatives of $J(\mathbf{w})$, let $y_i x_i^T w = L_i$:

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

Then second derivatives of $J(\mathbf{w})$:

$\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^2} = \frac{\partial}{\partial \mathbf{w}} \cdot \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{|D|}\sum_{i\in D} \frac{e^{y_i x_i^T \mathbf{w}}}{(1+e^{y_i x_i^T \mathbf{w}})^2} x_i x_i^T + 2\lambda \mathbf{I}$

$\frac{1}{|D|}\sum_{i\in D} \frac{e^{y_i x_i^T \mathbf{w}}}{(1+e^{y_i x_i^T \mathbf{w}})^2} x_i x_i^T$ is positive definite and $\lambda > 0$ So the Hessian matrix of $J(\mathbf{w})$ is positive definite, so according to the property of convex function, $J(\mathbf{w})$ is convex so it just have only one local optimal.

Question2 :

Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is sparse (has many zeros entries) $\Longrightarrow$ False

I found a answer in the book Deep-Learning, but it has 2 pages proof.

So in short, just as in the title of this question, l2 norm will reduce the lenght of the weight, not many of it weight will become 0, so weight matrix $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ don't have too much 0, so $\hat{\mathbf{w}}$ is not sparse.

3. Gradient descent for fitting GMM (15 points)

Consider the Gaussian mixture model

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

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.)

Question 1 the gradient of the log-likelihood wrt σ_k:

$\frac{d}{d\mu_k} l(\theta) = \frac{d}{d\mu_k} \sum_{i=1}^N \log p(\mathbf{x}|\theta)$

$= \frac{d}{d\mu_k} \sum_{i=1}^N \log \sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})$

$= \sum_{i=1}^N \frac{1}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})} \frac{d}{d\mu_k} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k)$

$= \sum_{i=1}^N \frac{1}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})} \frac{d}{d\mu_k} \pi_k (-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k))$

$= \sum_{i=1}^N \frac{1}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})} \Sigma_k^{-1}(x_n - \mu_k)$

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

Question 2

2.1 Derive the gradient of the log-likelihood wrt π_k without considering any constraint on π_k:

$\frac{d}{d\pi_k} l(\theta) = \frac{d}{d\pi_k} \sum_{i=1}^N \log p(\mathbf{x}|\theta)$

$= \frac{d}{d\pi_k} \sum_{i=1}^N \log \sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})$

$= \sum_{i=1}^N \frac{\mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})} = \sum_{i=1}^N \frac{r_{nk}}{\pi_k}$

2.2 **Bouns** with constraint$\Sigma_k \pi_k = 1$

After add gradient, the equation become $\frac{d}{d\pi_k} l(\theta) + \frac{d}{d\pi_k} \lambda(\Sigma_{k'} \pi_{k'} - 1)$, which is the result of 2.1 add λ, so result is $\sum_{i=1}^N \frac{r_{nk}}{\pi_k} + \lambda$

Question 3

Derive the gradient of the log-likelihood wrt Σ_k without considering any constraint on Σ_k:

$\frac{d}{d\Sigma_k} l(\theta) = \frac{d}{d\Sigma_k} \sum_{i=1}^N \log p(\mathbf{x}|\theta)$

$= \frac{d}{d\Sigma_k} \sum_{i=1}^N \log \sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})$

$= \sum_{i=1}^N \frac{\pi_k}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(\mathbf{x}|\mu_{k'}, \Sigma_{k'})} \frac{d}{d\Sigma_k} \pi_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Sigma_k)$

$= \sum_{i=1}^N r_{nk} (\frac{1}{2}(x - \mu_k)(x - \mu_k)^T (\Sigma^{-1})^2)$