CS280 Fall 2018 Assignment 1
Part A

ML Background

Due in class, October 12, 2018

Name: Ke Zhang

Student ID: 50369264

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

The likelihood function and its log form are written as

$$L(x,\theta) = \prod_{i=1}^{n} q(x_i;\theta)$$
$$logL(x,\theta) = \sum_{i=1}^{n} \log q(x_i;\theta)$$

So the likelihood estimator $\hat{\theta}$ is solved so that $\sum_{i=1}^{n}\log(q(x_i;\theta))$ is maximized, that is, $\hat{\theta}$ is the parameter that could make the model $q(x|\hat{\theta})$ generate the dataset $\mathcal{D}$ most likely.

$$\begin{aligned}
\min_q KL(p_{emp}||q) &= \min_q \Big(\int p(x)\log p(x)dx - \int p(x)\log q(x)dx \Big) \\
&\Rightarrow \max_q \int \sum_{i=1}^{n} \delta(x, x_i)\log q(x)dx \\
&= \max_q \sum_{i=1}^{n} \int \delta(x, x_i)\log q(x)dx \\
&= \max_q \sum_{i=1}^{n} \log q(x_i, \theta)
\end{aligned}$$

Therefore, $KL(p_{emp}||q)$ could be minimized by choosing $\theta = \hat{\theta}$ in $q(x,\theta)$.

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?

Solution

- False. In the loss function $J(\mathbf{w})$, both two terms are convex, so the function is convex and its optmial solution is unique which is locally as well as globally.
- False. Due to the mathematical formula of l_2, to minimize the loss function as well as constrain $\hat{\mathbf{w}}$, the term $\lambda\|\mathbf{w}\|_2^2$ could only make some cofficients as small as possible but can't eliminite them, so $\hat{\mathbf{w}}$ cannot be sparse like l_1 norm does.

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:

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

- 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.)

Solution

-

$$\begin{aligned}
l(\theta) &= \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta) \\
&= \sum_{n=1}^{N} \log \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \Big) \\
&= \sum_{n=1}^{N} \log \Big(\sum_{k=1}^{K} r_{nk} \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{r_{nk}} \Big) \\
&\geq \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{r_{nk}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \Big[\log \pi_k - \frac{1}{2} \log(2\pi|\Sigma_k|) - \frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \Big]
\end{aligned} \tag{1}$$

the last line above is derived by Jason's inequality $\log(E(X)) \geq E(\log X)$. To obtain the gradient w.r.t.μ_k, we only need to solve the lower bound of $l(\theta)$, therefore

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

- Compute the gradient w.r.t Σ_k of the function (1):

$$\frac{d}{d\pi_k} l(\theta) = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k}$$

If we have the constriant condition $\sum_{k=1} \pi_k = 1$, denote the Lagrange function be $\mathcal{L}(\theta) = l(\theta) + \lambda(\sum_{k=1} \pi_k - 1)$ where λ is a dual variable., the derivation is very close to the result in question (1) except that the extra term $\lambda(\sum_{k=1} \pi_k - 1)$ is added, hence

$$\frac{d}{d\pi_k} \mathcal{L}(\theta) = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} + \lambda \tag{2}$$

$$\frac{d}{d\lambda} \mathcal{L}(\theta) = \sum_{k=1}^{K} \pi_k - 1 \tag{3}$$

Let equation (2) and (3) equal to 0, then π_k could be solved.

- Take gradient w.r.t Σ_k in (1):

$$\frac{d}{d\Sigma_k} l(\theta) = \sum_{n=1}^{N} r_{nk} \Big(- \frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \frac{(\mathbf{x}_n - \mu_k)^T (\mathbf{x}_n - \mu_k)}{\Sigma_k^2} \Big)$$

If Σ_k is considered as symmetric PD matrix, i.e., $\Sigma_k \succ \mathbf{0}$, set the Lagrange function $\mathcal{L}(\theta) = l(\theta) + \text{Tr}(\Sigma_k \mathbf{\Lambda})$ where $\mathbf{\Lambda}$ is a dual variable.

$$\frac{d}{d\Sigma_k} l(\theta) = \sum_{n=1}^{N} r_{nk} \Big(- \frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \frac{(\mathbf{x}_n - \mu_k)^T (\mathbf{x}_n - \mu_k)}{\Sigma_k^2} \Big) + \mathbf{\Lambda}$$