CS280 Fall 2018 Assignment 1
Part A

ML Background

Due in class, October 12, 2018

Name: Yingying Ma

Student ID: 88678580

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

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

arg $\text{min}_q KL(p||q)$ divergent has nothing to do with the fist item

$$\begin{aligned} KL(p_{emp}||q(x;\hat{\theta})) &= \int_x p(x)\log p(x) - \int_x p(x)\log q(x;\hat{\theta}) \\ &= \int_x p(x)\log p(x) - \frac{1}{n}\int_x \sum_{i=1}^{n}\delta(x, x_i)\log q(x_i;\hat{\theta}) \\ &= \int_x p(x)\log p(x) - \frac{1}{n}\sum_{i=1}^{n}\log q(x;\hat{\theta}) \end{aligned}$$

When $\hat{\theta}$ the Maximum Likelihood Estimator

$$\begin{aligned} q(x) &= \text{arg min}_q KL(p||q) \\ &= \text{arg max}_q \sum_{i=1}^{n}\log q(x_i;\hat{\theta}) \\ &= \text{arg max}_q \log\prod_i q(x_i;\hat{\theta}) \end{aligned}$$

Thus, $\arg\min_q KL(p_{emp}||q)$ is obtained by $q(x) = q(x;\hat{\theta})$.

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

- F. Let $l(\boldsymbol{w}) = y_i\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{w}$, $\sigma'(l) = \sigma(l)(1-\sigma(l))$, let $g(l) = -\log(\sigma(l))$, then

$$g'(l) = -\frac{1}{\sigma(l)}\sigma(l)(1-\sigma(l)) = \sigma(l) - 1 < 0$$

  g is convex and l is an affine function, then J is a convex function. So $J(w)$ has one global optimal solution.

- F. L_2 regulation won't induce sparsity.

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.)

Solution

Let Q_n be some distribution over $z's(\sum_z Q_n(z) = 1, Q_n(z) \geq 0)$

$$\begin{aligned} l(\theta) &= \sum_n \log(x_n|\theta) \\ &= \sum_n \log \sum_{z_n} p(x_n, z_n|\theta) \\ &= \sum_n \log \sum_{z_n} Q_n(z_n) \frac{p(x_n, z_n|\theta)}{Q_n(z_n)} \\ &\geq \sum_n \sum_{z_n} Q_n(z_n) \log \frac{p(x_n, z_n|\theta)}{Q_n(z_n)} \end{aligned}$$

Denote $r_{nk} := p(z_n = k|x_n, \theta)$

$$\begin{aligned} \sum_n \sum_{z_n} Q_n(z_n) \log \frac{p(x_n, z_n|\theta)}{Q_n(z_n)} &= \sum_n \sum_{k=1} Q_n(z_n = k) \log \frac{p(x_n|z_n = k, \mu, \Sigma) p(z_n = k|\pi)}{Q_n(z_n = k)} \\ &= \sum_n \sum_{k=1} r_{nk} \log \frac{1}{(2\pi)|\Sigma_k|^{\frac{1}{2}}} exp(-\frac{1}{2}(x_n - \mu_n)^T \Sigma_k^{-1}(x_n - \mu_n))\pi_k \\ &\quad - \sum_n \sum_{k=1} r_{nk} \log r_{nk} \end{aligned}$$

- Compute the gradient

$$\frac{d}{d\mu_k}\sum_n\sum_{k=1} r_{nk}\log\frac{1}{(2\pi)|\Sigma_k|^{\frac{1}{2}}}exp(-\frac{1}{2}(x_n-\mu_n)^T\Sigma_k^{-1}(x_n-\mu_n))\pi_k-\sum_n\sum_{k=1} r_{nk}\log r_{nk}$$
$$=\frac{d}{d\mu_k}\sum_n\sum_{k=1} r_{nk}\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)$$
$$=\frac{1}{2}\sum_n r_{nk}\frac{d}{d\mu_k}(2\mu_k^T\Sigma_k^{-1}x_n-\mu_k^T\Sigma_k^{-1}\mu_k)$$
$$=r_{nk}\Sigma_k^{-1}(x_n-\mu_k)$$

Set the gradient to 0, then

$$\mu_k=\frac{\sum_n r_{nk}x_n}{\sum_n r_{nk}}$$

- Derive the gradient of the log-likelihood wrt π_k without considering any constraint on π_k, that is to maximize $\sum_n\sum)kr_{nk}\log\pi_k$.

With constraint $\sum_k\pi_k=1$, the Lagrangian function:

$$\begin{aligned}L(\pi) &= \sum_n\sum_k r_{nk}\log\pi_k+\beta(\sum_k\pi_k-1)\\ \frac{\partial}{\partial\pi_k}L(\pi) &= \sum_n\frac{r_{nk}}{\pi_k}+1=0\\ \longrightarrow\pi_k &= \frac{sum_{n=1}^N r_{nk}}{-\beta}\end{aligned}$$

As $-\beta=\sum_{n=1}^N\sum_k r_{nk}=\sum_{n=1}^N 1=N$

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

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

$$\frac{d}{d\mu_k}\sum_n\sum_{k=1} r_{nk}\log\frac{1}{(2\pi)|\Sigma_k|^{\frac{1}{2}}}exp(-\frac{1}{2}(x_n-\mu_n)^T\Sigma_k^{-1}(x_n-\mu_n))\pi_k-\sum_n\sum_{k=1} r_{nk}\log r_{nk}=0$$
$$\sum_n r_{nk}\Sigma_k^{-1}-\sum_n r_{nk}\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1}=0$$
$$\Sigma_k=\frac{\sum}{}(\sum_n r_{nk}\Sigma_k^{-1}-\sum_n r_{nk}\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1})\Sigma_k=0$$
$$\Sigma_k=\frac{\sum_n r_{nk}(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_n r_{nk}}$$

As Σ_k is symmetric, so with constraint Σ_k be a symmetric positive definite matrix,

$$\Sigma_k=\frac{\sum_n r_{nk}(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_n r_{nk}}$$