CS280 Fall 2018 Assignment 1
Part A

ML Background

Due in class, October 12, 2018

Name: 张相宜

Student ID: 2018233134

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.

Given a set of data points, $\{x_1 \dots x_n\}$

underlying distribution : $q(x)$, let $\tilde{p}(x)$ be the empirical distribution

$$\tilde{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)$$

KL-divergence from the empirical distribution $\tilde{p}(x)$ to the model distribution ~~$p(x|\theta)$~~ $q(x|\theta)$

$$KL(\tilde{p}(x) || q(x|\theta)) = \int \tilde{p}(x) \log \frac{\tilde{p}(x)}{q(x|\theta)} dx$$

$$= -\int \tilde{p}(x) \log \tilde{p}(x) dx - \int \tilde{p}(x) \log q(x|\theta)\, dx$$

so,

$$\arg\min_q KL(\tilde{p}(x) || q(x|\theta)) = \arg\max_q \log q(x|\theta)$$

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?

① F.

$$\frac{\partial J}{\partial w} = -\frac{1}{|D|}\sum_{i\in D}\underbrace{\frac{\partial \log\sigma(aw)}{\partial w}}_{f_1} + \lambda\,\underbrace{\frac{\partial(\|w\|_2^2)}{\partial w}}_{f_2}, \quad a = y_i x_i^T$$

当 $\lambda > 0$, $\lambda\frac{\partial\|w\|_2^2}{\partial w} > 0$

$$-\frac{\partial\log\sigma(aw)}{\partial w} = \frac{a\sigma'(aw)}{\sigma(aw)} = \cancel{-\frac{-\frac{-a^2e^{-aw}}{(1+e^{-aw})^2}}{\frac{1}{1+e^{-aw}}}} = -a^2\frac{\frac{-e^{-aw}}{(1+e^{-aw})^2}}{\frac{1}{1+e^{-aw}}}$$

$$f_1' = \frac{\partial\log(1+e^{-aw})}{\partial w} = \frac{-e^{-aw}}{1+e^{-aw}}$$

$$f_1'' = \frac{e^{-x}}{(1+e^{-aw})^2} \geq 0$$

$f_2'' > 0$

$\therefore$ $J(w)$ 为 convex ($\lambda>0$), $J(w)$ has only one globally optimal solution

② F

l_2 regularization will penalize the larger weight but will not penalize many weights to zero.

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.)

① $l(\theta) = \sum_{n=1}^{N} \log p(X|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$

$\frac{\partial l(\theta)}{\partial \mu_k} = -\frac{1}{2} \sum_{n=1}^{N} -2\Sigma^{-1}(X-\mu) = \sum_n \Sigma_k^{-1}(X_n - \mu_k)$

$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_k \|X_n - \mu_k\| \\ 0 \end{cases}$

$\therefore \frac{d}{d\mu_k} l(\theta) = \sum_n r_{nk} \Sigma_k^{-1}(X_n - \mu_k)$

② $\frac{\partial l(\theta)}{\partial \pi_k} = 0, \forall k, \ \text{s.t.} \sum_k \pi_k = 1 \Rightarrow \pi_k = \frac{\sum_n Z_n^k}{n}$

$l(\theta; X, Z) = \sum_n \langle \log P(Z_n|\pi) \rangle_{p(Z|X)} + \sum_n \langle \log P(X_n|Z_n, \mu, \Sigma) \rangle_{p(Z|X)}$

$= \sum_n \sum_k \langle Z_n^k \rangle \log \pi_k - \frac{1}{2} \sum_n \sum_k \langle Z_n^k \rangle (X_n - \mu_k)^T \Sigma_k^{-1}(X_n - \mu_k) + \log|\Sigma_k| + C$

$(Z_n^{(t)} = \arg\max_k (X_n - \mu_k^{(t)})^T \Sigma^{-1}(X_n - \mu_k^{(t)}))$

$\frac{\partial l}{\partial \pi_k} = \sum_n \sum_k \frac{Z_n^k}{\pi_k} \overset{\sum_k \pi_k = 1}{=\!=\!=} \sum_n \frac{\sum_k Z_n^k}{1} = \sum_n \sum_k Z_n^k$

③ 在后页

③ $\frac{d}{d\Sigma^{-1}} \log P(X;\mu,\Sigma) = \sum_{n=1}^{N} \frac{\partial\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) - \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma^{-1}|\right)}{\partial\Sigma^{-1}}$

$\frac{d(a^T X a)}{dX} = a a^T \qquad \frac{d \log|X|}{dX} = X^{-T} \Rightarrow \frac{d \log|\Sigma^{-1}|}{d\Sigma^{-1}} = (\Sigma^{-1})^{-T} = \Sigma$

(Σ is symmetric positive definite)

$\Rightarrow \frac{d}{d\Sigma^{-1}} \log P(X;\mu,\Sigma) = -\frac{1}{2}(X-\mu)(X-\mu)^T + \frac{1}{2}\Sigma$

$\frac{d}{d\Sigma} = \frac{d\lg P}{d\Sigma^{-1}} \cdot \frac{d\Sigma^{-1}}{d\Sigma} = \left(-\frac{1}{2}(X-\mu)(X-\mu)^T + \frac{1}{2}\Sigma\right) \cdot -\frac{1}{\Sigma^2}$

$= \frac{1}{2\Sigma^2}(X-\mu)(X-\mu)^T - \frac{1}{2\Sigma}$