CS280 Fall 2017 Assignment 1
Part A

ML Background

Due in class, October 13, 2017

Name: 陈宏宇

Student ID:87522273

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.
$\arg\min_q KL(p_{emp}||q) = \arg\min_q \int p_{emp}(x)(\log p_{emp}(x) - \log q(x))dx$

so it is equals to minimize

$-\arg\min_q \int p_{emp}(x)(\log q(x))dx$

$= -\arg\min_q \sum_x [\sum_{i=1}^{n} \delta(x, x_i)] \log q(x)$

$= -\arg\min_q \sum_{i=1}^{n} \log q(x_i)$

$= -\arg\min_q \log q(x)$

if we want to minimize the $-\arg\min_q \log q(x)$ when $q(x) = q(x; \hat{\theta})$, then we get the minimal value.

■

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. (1):F

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

$\frac{\partial^2 J(w)}{\partial w^2} = \frac{1}{|D|}\sum_{i \in D}(y_i \mathbf{x}_i^T)^2 + 2\lambda >= 0$

So it just have only one locally optimal solution.
(2) F.

We assume that the loss function

$J(\mathbf{w}) = L(\mathbf{w}) + \lambda ||\mathbf{w}||^2$

$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} + 2\lambda \mathbf{w}$

if we adopt the gradient decent algorithm we update the parameter $\mathbf{w}$ by the following :

$\mathbf{w} \longrightarrow \mathbf{w} - 2\lambda \mathbf{w} - \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$

$\mathbf{w} \longrightarrow (1 - 2\lambda)\mathbf{w} - \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$

When $\mathbf{w} < 1$,the effect of the parameter $\mathbf{w}$ is small so $\hat{\mathbf{w}}$ cannot be 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.)

Proof. 1:

$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{|2\pi|^D |\sum_k|}} exp((\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)))$

$\log(\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k))$

$= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log(|\Sigma_k^{-1}|) - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)$

$\frac{d}{d\mu_k} l(\theta) = \frac{1}{p(\mathbf{x}|\theta)} \frac{\partial p(\mathbf{x}|\theta)}{\partial \mu_k}$

$= \frac{1}{p(\mathbf{x}|\theta)} \frac{\partial \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\partial \mu_k} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_k k')} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)$

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

2: $\frac{d}{d\pi_k} l(\theta) = \frac{1}{p(\mathbf{x}|\theta)} \frac{\partial p(\mathbf{x}|\theta)}{\partial \pi_k}$

$= \frac{\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_k k')}$

3:

$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{|2\pi|^D |\sum_k|}} exp((\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k)))$

$\log(\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k))$

$= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log(|\Sigma_k^{-1}|) - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k)$

$\frac{d \log(\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{d(\Sigma_k)}$
$= -\frac{N}{2} \Sigma_k^{-1} - \frac{1}{2} \frac{d(Tr[\Sigma_k^{-1} S])}{d(\Sigma_k)}$

$= -\frac{N}{2} \Sigma_k^{-1} + \frac{1}{2} \Sigma_k^{-1} S \Sigma_k^{-1}$

with $S = (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$

■

4. Residual error for PCA (10 points)

Given $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, and $\mathbf{x}_i \in R^d$, the principle components and their corresponding eigenvalues are denoted by $\{(\mathbf{v}_j, \lambda_j)\}_{j=1}^d$.

- Denote $z_{ij} = \mathbf{x}_i^T \mathbf{v}_j$. Prove that

$$\|\mathbf{x}_i - \sum_{j=1}^K z_{ij}\mathbf{v}_j\|^2 = \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^K \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j$$

- Show that

$$J_K := \frac{1}{n}\sum_{i=1}^n \left(\mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^K \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^K \lambda_j = \sum_{j=K+1}^d \lambda_j$$

Proof. because x_i and v_j are both vectors,so we can get that $x_i^T v_j = v_j^T x_i$.

So now we can simplify that

$\|x_i - \sum_{j=1}^K z_{ij}v_j\|^2$

$= (x_i - \sum_{j=1}^K z_{ij}v_j)^T(x_i - \sum_{j=1}^K z_{ij}v_j)$

$= x_i^T x_j - x_i^T \sum_{j=1}^K z_{ij}v_j - (\sum_{j=1}^K z_{ij}v_j)^T x_i + (\sum_{j=1}^K z_{ij}\mathbf{v}_j)^T(\sum_{j=1}^K z_{ij}v_j)$

$= x_i^T x_j - \sum_{j=1}^K x_i^T v_j v_j^T x_i - \sum_{j=1}^K x_i^T v_j v_j^T x_i + \sum_{j=1}^K z_{ij}\mathbf{v}_j^T(\sum_{j=1}^K z_{ij}v_j)$

$= x_i^T x_j - \sum_{j=1}^K x_i^T v_j v_j^T x_i - \sum_{j=1}^K x_i^T v_j v_j^T x_i + \sum_{j=1}^K z_{ij}\mathbf{v}_j^T(\sum_{j=1}^K z_{ij}v_j)$

$= x_i^T x_j - \sum_{j=1}^K x_i^T v_j v_j^T x_i - \sum_{j=1}^K x_i^T v_j v_j^T x_i + \sum_{j=1}^K z_{ij}z_{ij}v_j^T v_j$

$= x_i^T x_j - \sum_{j=1}^K x_i^T v_j v_j^T x_i - \sum_{j=1}^K x_i^T v_j v_j^T x_i + \sum_{j=1}^K x_i^T v_j v_j^T x_i$

$= x_i^T x_j - \sum_{j=1}^K x_i^T v_j v_j^T x_i$

$= x_i^T x_i - \sum_{j=1}^K v_j^T x_i x_i^T v_j$ ■

Proof.
$J_K := \frac{1}{n}\sum_{i=1}^n \left(\mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^K \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j\right)$

$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{K}\mathbf{v}_j^T[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T]\mathbf{v}_j$

$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{K}\lambda_j$

$= \sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i\mathbf{v}_j^T\mathbf{v}_j - \sum_{j=1}^{K}\lambda_j$

$= \sum_{j=1}^{d}\mathbf{v}_j^T[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i]\mathbf{v}_j - \sum_{j=1}^{K}\lambda_j$

$= \sum_{j=1}^{d}\lambda_j - \sum_{j=1}^{K}\lambda_j$
$= \sum_{j=k+1}^{d}\lambda_j$

■