CS280 Fall 2019 Assignment 1
Part A

Basic Neural Networks

Due on October 10, 2019, 23:59 UTC+8

Name:Zhang Chengrui

Student ID:2019233183

1. Hessian in Logistic Regression (10 points)

Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be an activation function, the loss function of LR is

$$f(\mathbf{w}) = -\sum_{i=1}^{n}[y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)],$$

where $\mu_i = \sigma(\mathbf{w}^\intercal \mathbf{x}_i)$.

- Show that the Hessian of f can be written as $H = X^\intercal S X$, where $S = diag(\mu_1(1-\mu_1), \cdots, \mu_n(1-\mu_n))$ and $X = [X_1, \cdots, X_n]^\intercal$

Solution:

$$\frac{\partial u}{\partial w_a} = \frac{e^{-w^T x_i}}{(1+e^{-w^T x_i})^2} \cdot \frac{\partial (w^T x_i)}{\partial w_a}$$

$$= \frac{e^{-w^T x_i}}{(1+e^{-w^T x_i})^2} \cdot x_i$$

$$\therefore \frac{\partial f(w)}{\partial w_a} = -\sum_{i=1}^{n}\left[y_i \cdot (\cancel{1+e^{-w^T x_i}}) \cdot \frac{e^{-w^T x_i}}{\cancel{(1+e^{-w^T x_i})^2}} x_i + (1-y_i)\cdot(-1)\cdot \cancel{\frac{1+e^{-w^T x_i}}{e^{-w^T x_i}}} \cdot \frac{\cancel{e^{-w^T x_i}}}{\cancel{(1+e^{-w^T x_i})^2}} \cdot x_i \right]$$

$$= -\sum_{i=1}^{n} [y_i - \mu_i] \cdot x_i$$

and
$$\frac{\partial^2 f(w)}{\partial w_a \partial w_b} = -\sum_{i=1}^{n} \frac{\partial [y_i - \mu_i]\cdot x_i}{\partial w_b}$$

$$= \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial w_b} \cdot x_i = \sum_{i=1}^{n} x_i \cdot \frac{e^{-w^T x_i}}{(1+e^{-w^T x_i})^2} \cdot x_i = \sum_{i=1}^{n} x_i \cdot \mu_i(1-\mu_i) \cdot x_j \quad \cdots \quad (1)$$

while
$$X^T \cdot S \cdot X$$
$$= [x_1, \cdots x_n] \cdot \begin{bmatrix} \mu_1(1-\mu_1) & & 0 \\ & \ddots & \\ 0 & & \mu_n(1-\mu_n) \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= [x_1 \mu_1(1-\mu_1) + \cdots \ x_n \mu_n(1-\mu_n)] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= x_1 \mu_1(1-\mu_1) x_1 + \cdots \ x_n \mu_n(1-\mu_n) x_n$$

$$= \sum_{i=1}^{n} x_i \mu_i(1-\mu_i) x_i \quad \cdots \quad (2)$$

$\because H = (1) = (2)$

Therefore, Hessian of f can be written as $H = X^T S X$

2. Linear Regression (5 points)

Linear regression has the form

$$f(x) = E[y|x] = b + \mathbf{w}^\intercal \mathbf{x}.$$

- It is possible to solve for $\mathbf{w}$ and b separately. Show that

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\intercal \mathbf{w} = \bar{y} - \bar{\mathbf{x}}^\intercal \mathbf{w}$$

Solution:

$$f(x) = b + w^T x = b + w_1 x_1 + w_2 x_2 + \cdots w_n x_n$$

$$\because y_i = b + w^T x_i$$

$$\therefore \sum_{i=1}^{n} y_i = nb + \sum_{i=1}^{n} W^T x_i$$

$$\therefore nb = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} W^T x_i$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i^T W \Rightarrow b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i^T W \cdots (1)$$

suppose $x_i = [x_{i1}, x_{i2}, \cdots x_{ia}]^T$ $W = [W_1, W_2, \cdots W_a]^T$

$$\because \frac{1}{n}\sum_{i=1}^{n} x_i^T W = \frac{1}{n}[\, x_{11} W_1 + x_{12} W_2 + \cdots x_{1a} W_a +$$
$$x_{21} W_1 + x_{22} W_2 + \cdots x_{2a} W_a +$$
$$\vdots$$
$$x_{a1} W_1 + x_{a2} W_2 + \cdots X_{aa} W_a \,]$$

$$= \frac{1}{n}[W_1(x_{11} + x_{21} + \cdots x_{a1}) + \cdots + W_a(x_{1a} + x_{2a} + \cdots X_{aa})]$$

$$= \frac{1}{n} \cdot n \cdot \bar{x}^T \cdot W = \bar{x}^T \cdot W \cdots (2)$$

put (2) into (1)

$$\therefore b = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}^T W$$

$$= \bar{y} - \bar{x}^T W$$

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:

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

- 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:

Q1:

$$\frac{d\, l(\theta)}{d\mu_k} = \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \cdot \frac{d\, p(x_n|\theta)}{d\mu_k} \quad \cdots (1)$$

$$\frac{d\, p(x_n|\theta)}{dk} = \sum_{k=1}^{K} \pi_k \frac{d\, N(x|\mu_k, \Sigma_k)}{d\mu_k} = \pi_k \frac{d\, N(x|\mu_k, \Sigma_k)}{d\mu_k} \quad \cdots (2)$$

$$\frac{d\, N(x|\mu_k, \Sigma_k)}{d\mu_k} = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right] \cdot \left(-\frac{1}{2}\right) \frac{d(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}{d\mu_k} \quad \cdots (3)$$

$$\frac{d(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}{d\mu_k} = \frac{d\, Y^T \Sigma^{-1} Y}{dY} \cdot \frac{dY}{d\mu_k} = -\left[\Sigma^{-1} + (\Sigma^{-1})^T\right] \cdot Y = -2\Sigma^{-1} Y \quad (Y = (x-\mu_k)) \quad \cdots (4)$$

put (2) (3) (4) into (1) $\Rightarrow$ $\frac{d\, l(\theta)}{d\mu_k} = \sum_n r_{nk} \Sigma^{-1} (x_n - \mu_k)$

Q2:

$$\frac{d\, l(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \cdot \frac{d\, p(x_n|\theta)}{d\pi_k} \quad \cdots (1)$$

$$\frac{d\, p(x_n|\theta)}{d\pi_k} = \frac{d \sum_{k=1}^{N} \pi_k N(x|\mu_k, \Sigma_k)}{d\pi_k} = N(x|\mu_k, \Sigma_k) \quad \cdots (2)$$

put (2) into (1) $\Rightarrow$ $\frac{d\, l(\theta)}{d\pi_k} = \sum_{n=1}^{N} \frac{N(x|\mu_k, \Sigma_k)}{p(x_n|\theta)} = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k}$

Q3:

$$\frac{d\, l(\theta)}{d\Sigma_k} = \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \cdot \frac{d\, p(x_n|\theta)}{d\Sigma_k} \quad \cdots (1)$$

suppose $A = \exp\left[-\frac{1}{2}\left[(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)\right]\right]$ then $N(x|\mu_k, \Sigma_k) = \frac{A}{(2\pi)^{D/2} \cdot |\Sigma_k|^{\frac{1}{2}}}$

$$\frac{d\, p(x_n|\theta)}{d\Sigma_k} = \Sigma_k \pi_k \cdot \frac{d \frac{1}{(2\pi)^{D/2} \cdot |\Sigma_k|^{\frac{1}{2}}} \cdot A}{d\Sigma_k}$$

$$= \sum_k \pi_k \cdot \left[\frac{1}{(2\pi)^{D/2}} \cdot \left(-\frac{1}{2} \cdot \frac{1}{|\Sigma_k|^{\frac{3}{2}}}\right) \cdot |\Sigma_k| \cdot (\Sigma_k^{-1})^T \cdot A + \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{\frac{1}{2}}} \cdot A \cdot \left[-\frac{1}{2} \cdot \left(-\Sigma_k^{-T} (x-\mu_k)(x-\mu_k)^T \cdot \Sigma_k^{-T}\right)\right]\right]$$

$$= \Sigma_k \pi_k \cdot A \cdot \left(-\frac{1}{2}\right) \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \cdot \frac{1}{(2\pi)^{D/2}} \cdot \Sigma_k^{-1} \cdot \left[I - (x-\mu_k)(x-\mu_k)^T \cdot \Sigma_k^{-T}\right]$$

$$= -\frac{1}{2}\left[\Sigma_k^{-T} - \Sigma_k^{-T} (x-\mu_k)(x-\mu_k)^T \Sigma_k^{-T}\right] \cdot N(x|\mu_k, \Sigma_k) \quad \cdots (2)$$

put (2) into (1) $\frac{d\, l(\theta)}{d\Sigma_k} = -\frac{1}{2} \sum_{n=1}^{N} \left[\Sigma_k^{-T} - \Sigma_k^{-T} (x-\mu_k)(x-\mu_k)^T \Sigma_k^{-T}\right]$

~~put (2) into (1)~~

$$\cancel{\frac{d\, l(\theta)}{d\Sigma_k} = -\frac{1}{2} \sum_{n=1}^{N} r_{nk} \cdot \left[\Sigma_k^{-T} - \Sigma_k^{-T} (x_n-\mu_k)(x_n-\mu_k)^T \Sigma_k^{-T}\right]}$$