DS4400 HW2

Xin Guan

1. **Linear Regression:** Consider the modified linear regression problem

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} (\theta^T \phi(x_i) - y_i)^2 + \lambda \|\theta - \mathbf{a}\|_2^2$$

where a is a known and given vector of the same dimension as that of θ. Derive the closed-form solution. Provide all steps of the derivation.

Solution:

$f(\theta) = \sum_{i=1}^{N} (\theta^T \phi(x_i) - y_i)^2 + \lambda \|\theta - \mathbf{a}\|_2^2$

$\frac{\partial f(\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{N} (\theta^T \phi(x_i) - y_i)^2}{\partial \theta} + \frac{\partial \lambda \|\theta - a\|_2^2}{\partial \theta}$

$= \sum_{i=1}^{N} [2(\theta^T \phi(x_i) - y_i) \frac{\partial(\theta^T \phi(x_i) - y_i)}{\partial \theta}] + \lambda \frac{\partial \|\theta - a\|_2^2}{\partial \theta}$

$= \sum_{i=1}^{N} [2(\theta^T \phi(x_i) - y_i)\phi(x_i)] + \lambda \frac{\partial(\theta - a)^2}{\partial \theta}$

$\lambda \frac{\partial(\theta - a)^2}{\partial \theta} = \lambda \frac{\partial(\theta^2 - 2\theta a + a^2)}{\partial \theta} = \lambda(2\theta - 2a)$

Therefore, $\frac{\partial f(\theta)}{\partial \theta} = \sum_{i=1}^{N} [2(\theta^T \phi(x_i) - y_i)\phi(x_i)] + 2\lambda(\theta - a)$

Write all data $\phi(x_1), \phi(x_2) \dots \phi(x_N)$ as a matrix:

$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \dots \\ \phi(x_N)^T \end{bmatrix}$ the dimension is $N \times d$

$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$ the dimension is $N \times d$

Then $\frac{\partial f(\theta)}{\partial \theta} = 2(\Phi^T \Phi \theta - \Phi^T Y) - 2\lambda(\theta - a)$

Let $\frac{\partial f(\theta)}{\partial \theta} = 0$

$$\Phi^T \Phi \theta - \Phi^T Y = \lambda(\theta - a)$$

$$\Phi^T \Phi \theta - \Phi^T Y = \lambda I_d \theta - \lambda I_d a$$

$$\Phi^T \Phi \theta - \lambda I_d \theta = \Phi^T Y - \lambda I_d a$$

$$(\Phi^T \Phi - \lambda I_d)\theta = \Phi^T Y - \lambda I_d a$$

$$\theta = (\Phi^T \Phi - \lambda I_d)^{-1}(\Phi^T Y - \lambda I_d a)$$

Therefore, $\hat{\theta}$ is $(\Phi^T \Phi - \lambda I_d)^{-1}(\Phi^T Y - \lambda I_d a)$

2. **Robust Regression using Huber Loss:** In the class, we defined the Huber loss as

$$\ell_\delta(e) = \begin{cases} \frac{1}{2}e^2 & |e| \le \delta \\ \delta|e| - \frac{\delta^2}{2} & |e| \ge \delta \end{cases}$$

Consider the robust regression model

$$\min_\theta \sum_{i=1}^{N} \ell_\delta(y_i - \theta^T\phi(x_i))$$

where $\phi(x_i)$ and y_i denote the i-th input sample and output/response, respectivly and unknown parameter vector.

a) Provide the steps of the batch gradient descent in order to obtain the solution for θ.

Solution:

Let $J(\theta) = \sum_{i=1}^{N} \ell_\delta(y_i - \theta^T\phi(x_i))$

We have : $\frac{\partial \ell_\delta(e)}{\partial e} = \begin{cases} e & |e| \le \delta \\ \delta & e \ge \delta \\ -\delta & e \le -\delta \end{cases}$

Therefore, $\frac{\partial J(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{N} \partial \ell_\delta(y_i - \theta^T\phi(x_i))}{\partial\theta} = \sum_{i=1}^{N} \begin{cases} [y_i - \theta^T\phi(x_i)]\cdot\phi(x_i) & |y_i - \theta^T\phi(x_i)| \le \delta \\ \delta\cdot\phi(x_i) & y_i - \theta^T\phi(x_i) \ge \delta \\ -\delta\cdot\phi(x_i) & y_i - \theta^T\phi(x_i) \le -\delta \end{cases}$

Gradient Descent Steps:

Assuming we have a Maximum iteration number T_{max}, threshold ϵ and Learning rate ρ.

(i) Pick the initial point θ^0

(ii) For $t = 1, 2, \dots, T_{max}$

- for $i = 1, 2, \dots, N$, calculate $\frac{\partial \ell_\delta(y_i - \theta^T\phi(x_i))}{\partial\theta} = \begin{cases} [y_i - \theta^T\phi(x_i)]\cdot\phi(x_i) & |y_i - \theta^T\phi(x_i)| \le \delta \\ \delta\cdot\phi(x_i) & y_i - \theta^T\phi(x_i) \ge \delta \\ -\delta\cdot\phi(x_i) & y_i - \theta^T\phi(x_i) \le -\delta \end{cases}$
- sum them up to get $\frac{\partial J(\theta)}{\partial\theta}$
- If $\|\frac{\partial J(\theta)}{\partial\theta}\|_2^2 \le \epsilon$, return θ^{t-1}; else, $\theta^t = \theta^{t-1} - \rho\frac{\partial J(\theta)}{\partial\theta}|_{\theta^{t-1}}$

b) Provide the steps of the stochastic gradient descent using mini-batches of size 1, i.e., one sample in each mini-batch, inorder to obtain the solution for θ

Solution:

This step is not very different from the above process. Just add a sampling step before the calculation of $\frac{\partial J(\theta)}{\partial\theta}$. In the sampling step, just randomly pick a $x_p \in \{x_1, x_2, \dots, x_N\}$

Write down as:

Stochastic Gradient Descent Steps:

Assuming we have a Maximum iteration number T_{max}, threshold ϵ and Learning rate ρ.

(i) Pick the initial point θ^0

(ii) For $t = 1, 2, \dots, T_{max}$

- randomly pick an $x_p \in \{x_1, x_2, \dots, x_N\}$
- Calculate $\frac{\partial J(\theta)}{\partial\theta} = \frac{\partial \ell_\delta(y_p - \theta^T\phi(x_p))}{\partial\theta} = \begin{cases} [y_p - \theta^T\phi(x_p)]\cdot\phi(x_p) & |y_p - \theta^T\phi(x_p)| \le \delta \\ \delta\cdot\phi(x_p) & y_p - \theta^T\phi(x_p) \ge \delta \\ -\delta\cdot\phi(x_p) & y_p - \theta^T\phi(x_p) \le \delta \end{cases}$

- If $\|\frac{\partial J(\theta)}{\partial \theta}\|_2^2 \le \epsilon$, return θ^{t-1}; else, $\theta^t = \theta^{t-1} - \rho \frac{\partial J(\theta)}{\partial \theta}|_{\theta^{t-1}}$

3. **Probability and Random Variables:** State true or false. If true, prove it. If false, either prove or demonstrate by a counter example. Here Ω denotes the sample space and A^c denotes the complement of the event A. X and Y denote random variables.

   (a) For any $A, B \subseteq \Omega$ such that $0 < P(A) < 1, P(A|B) + P(A|B^c) = 1$

   Solution: This is **False**

   Proof. From the Question, $P(B) + P(B^c) = 1$.
   $P(A|B) = \frac{P(A \cap B)}{P(B)}$, $P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)}$
   Since $P(A \cap B) + P(A \cap B^c) = P(A)$
   $P(A|B) + P(A|B^c) = \frac{P(A \cap B)}{P(B)} + \frac{P(A) - P(A \cap B)}{1 - P(B)}$
   Then we let $P(B) = 0.5$, $P(A) = 0.4$ and $P(A \cap B) = 0.3$
   $P(A|B) + P(A|B^c) = \frac{0.3}{0.5} + \frac{0.4 - 0.3}{1 - 0.5} = 0.6 + 0.2 = 0.8 \neq 1$
   Therefore, the given term is False. □

   (b) For any $A, B \subseteq \Omega$ $P(B^c \cap (A \cup B)) + P(A^c \cup B) = 1$

   Solution: This is **True**

   Proof. $P(B^c \cap (A \cup B)) = P((B^c \cap A) \cup (B^c \cap B)) = P((B^c \cap A) \cup \emptyset) = P(B^c \cap A)$
   Therefore, $P(B^c \cap (A \cup B)) + P(A^c \cup B) = P(B^c \cap A) + P(A^c \cup B)$ Since $P(A) = P(A \cap B^c) + P(A \cap B)$,
   We can write $P(B^c \cap A) = P(A) - P(A \cap B)$
   Also, we can write $P(A^c \cup B) = P(A^c) + P(B) - P(A^c \cap B)$
   Then $P(B^c \cap (A \cup B)) + P(A^c \cup B)$
   $= P(B^c \cap A) + P(A^c \cup B)$
   $= P(A) - P(A \cap B) + P(A^c) + P(B) - P(A^c \cap B)$
   $= P(A) + P(A^c) + P(B) - (P(A \cap B) + P(A^c \cap B))$
   Since $P(A) + P(A^c) = 1$ and $P(A \cap B) + P(A^c \cap B) = P(B)$
   $P(B^c \cap (A \cup B)) + P(A^c \cup B) = 1$ □

   (c) $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$

   Solution: This is **True**

   Proof. (By induction)
   Base Case: n = 1
   When $n = 1, P(A_1, \ldots, A_n) = P(A_1)$,
   $P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1}) = P(A_1)$
   Therefore, when $n = 1$, $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$ is ture.
   Inductive Steps:
   Inductive Hypothesis:
   $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$ is true when $n = k$.
   Claim:
   $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$ is true when $n = k + 1$ **Proof of Claim:**
   When n = k+1, right hand side:
   $P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$
   $= P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{k-1})P(A_n|A_1, \cdots, A_k)$
   $= P(A_1, \ldots, A_k)P(A_{k+1}|A_1, \cdots, A_{k+1-1})$

$= P(A_1, \ldots, A_k) \frac{P(A_{k+1} \cap A_1, \cdots, A_k)}{P(A_1, \cdots, A_k)}$
$= P(A_{k+1} \cap A_1, \cdots, A_k)$
$= P(A_1, \ldots, A_{k+1})$
Therefore, the claim is true.
Thus, $P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \cdots, A_{n-1})$ is true. □

(d) If X and Y are independent discrete random variables, then $E[XY] = E[X]E[Y]$, where $E[\cdots]$ denotes expectation.
Solution:

Proof. $X \perp\!\!\!\perp Y \Rightarrow P(XY) = P(X)P(Y)$
$E(XY)$
$= \sum xyP(XY)$
$= \sum xyP(X)P(Y)$
$= \sum xP(X) \sum yP(Y)$
$= E[X]E[Y]$ □

4. **Maximum Likelihood Estimation:** Assume $X_1, X_2, \ldots, X_N$ are i.i.d. random variables each taking a real value, where

$$p_\delta(X_i = x_i) = e^{-\delta^2 + \delta x_i}$$

Here, δ is the parameter of the distribution. Assume, we observe $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$.

(a) Write down the likelihood function $L(\delta)$.
Solution:
$L(\delta) = \prod_{i=1}^{i=N} p_\delta(X_i = x_i)$
$= \prod_{i=1}^{i=N} e^{-\delta^2 + \delta x_i}$

(b) Derive the maximum likelihood or log-likelihood estimation of δ for the given observations. Provide all steps of derivations.
Solution:
$log(L(\delta)) = log(\prod_{i=1}^{i=N} e^{-\delta^2 + \delta x_i})$
$= \sum_{i=1}^{i=N} log(e^{-\delta^2 + \delta x_i})$
$= \sum_{i=1}^{i=N} -\delta^2 + \delta x_i$
$= -N\delta^2 + \delta \sum_{i=1}^{i=N} x_i$
Then $\frac{\partial log(L(\delta))}{\partial \delta} = \frac{\partial(-N\delta^2 + \delta \sum_{i=1}^{i=N} x_i)}{\partial \delta} = -2N\delta + \sum_{i=1}^{i=N} x_i$
Let $\frac{\partial log(L(\delta))}{\partial \delta} = 0$. Then, $-2N\delta + \sum_{i=1}^{i=N} x_i = 0$
$\Rightarrow 2N\delta = \sum_{i=1}^{i=N} x_i \Rightarrow \delta = \frac{\sum_{i=1}^{i=N} x_i}{2N}$
Therefore, $\hat{\delta} = \frac{\sum_{i=1}^{i=N} x_i}{2N}$

5. **Logistic Regression:** In the logistic regression for binary classification ($y \in \{0, 1\}$), we defined $p(y = 1|x) = \sigma(\omega^T x)$, where the sigmoid function is defined as

$$\sigma(z) \triangleq \frac{1}{1 + e^{-z}}$$

Assume we have trained the logistic regression model using a given dataset and have learned ω. Let x_n be a test sample.

(a) Assume $\omega^T x_n < 0.3$. To which class x_n belongs? Provide details of your derivations.

Solution:
$\omega^T x_n < 0.3 \Rightarrow e^{-\omega^T x_n} > e^{-0.3}$
$\Rightarrow 1 + e^{-\omega^T x_n} > 1 + e^{-0.3}$
$\Rightarrow \frac{1}{1+e^{-\omega^T x_n}} < \frac{1}{1+e^{-0.3}}$
$\Rightarrow p(y = 1|x) < 0.5744$

(b) Assume $\frac{1}{1+e^{\omega^T x_n}} = 0.7$. To which class x_n belongs and with what probability? Provide detaisl of your derivations.

Solution:
$\frac{1}{1+e^{\omega^T x_n}} = 0.7 \Rightarrow \frac{1}{0.7} = 1 + 1 + e^{\omega^T x_n} \Rightarrow e^{\omega^T x_n} = \frac{1}{0.7} - 1 \Rightarrow e^{-\omega^T x_n} = \frac{1}{\frac{1}{0.7}-1} = \frac{0.7}{1-0.7} = \frac{0.7}{0.3}$
$\sigma(\omega^T x_n) = \frac{1}{1+\frac{0.7}{0.3}} = \frac{0.3}{0.3+0.7} = 0.3$
Therefore, $p(y = 0|x_n) = 1 - p(y = 1|x_n) = 1 - 0.3 = 0.7$
Thus, x_n belongs to 0 with probability of 0.7