先端機械学習　前半課題

提出日：2021 年 7 月 29 日

情報工学系

学籍番号：18B14822

氏名：宮崎　直哉

本課題で作成したコードはgithub(https://github.com/naoyaaan/AdvancedMachineLearning/tree/main/Shimosaka) に公開しています。

1 PROBLEM 1

本問題で作成したソースコードはgithub(https://github.com/naoyaaan/AdvancedMachineLearning/tree/main/Shimosaka) に公開しています。

Gradient と Hessian を求める。

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{-\exp(-y_i w^T x_i)}{1+\exp(-y_i w^T x_i)} y_i x_i + 2\lambda w \tag{1}$$

$$\begin{aligned} \nabla^2 J(w) &= \frac{\partial}{\partial w} \frac{\partial J(w)}{\partial w^T}\bigg|_{w=w^{(t)}} && (2) \\ &= \sum_{i=1}^{n} \frac{\exp(-y_i w^T x_i)}{(1+\exp(-y_i w^T x_i))^2} y_i^T y_i x_i x_i^T + 2\lambda && (3) \end{aligned}$$

1 Steepest Gradient Descent

最急降下法では、以下の式で w の更新を行う。

$$d^{(t)} = -\frac{\partial J}{\partial w}\bigg|_{w=w^{(t)}} \tag{4}$$

$$w^{(t+1)} = w^{(t)} + \alpha^{(t)} d^{(t)} \tag{5}$$

なお、ここでは更新幅 α を式 (6) のように定める。

$$\alpha^{(t)} = \frac{\alpha_0}{\sqrt{t}} \tag{6}$$

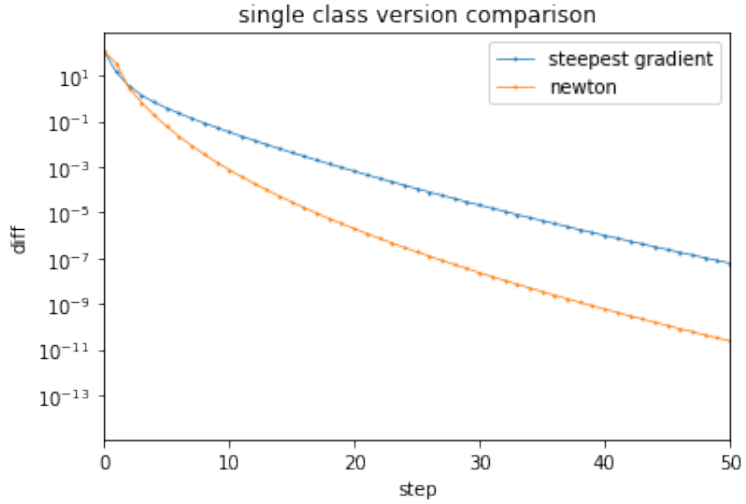
2 Newton method

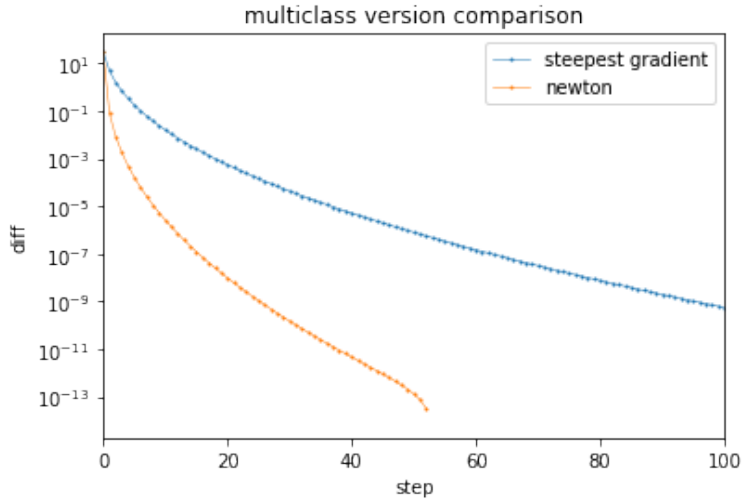
ニュートン法では、Hessian を用いてパラメータの更新を行う。

$$\nabla^2 J(w^{(t)}) d^{(t)} = -\nabla J(w^{(t)}) \tag{7}$$

3 Compare

最急降下法とニュートン法を上記の方法で実装し、比較したグラフは以下のようである。(single class version comparison)





4 Multiclass version

マルチクラスのデータセットに対しても、同様に最急降下法とニュートン法を実装した。両者を比較したグラフは上図のよう (multiclass version comparisoin)。

2 PROBLEM 2

本課題で作成したコードはgithub(`https://github.com/naoyaaan/AdvancedMachineLearning/tree/main/Shimosaka`) に公開しています。

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left((w-\mu)^T A(w-\mu) + \lambda\|w\|_1\right) \tag{8}$$

近接勾配法を用いて $\hat{w}$ を求める。$J(w) = \left((w-\mu)^T A(w-\mu) + \lambda\|w\|_1\right)$ とおくと、その勾配は以下のよう。

$$\nabla J(w) = 2A(w-\mu) \tag{9}$$

また、w の更新は以下の式のように行う。

$$w^{(t+1)} = prox_{\eta_t}\left(w^{(t)} - \eta\nabla J(w)\right) \tag{10}$$

$$= prox_{\eta_t}\left(w^{(t)} - \frac{1}{L}2A(w-\mu)\right) \tag{11}$$

ここで、この問題では L1 regularization を考えているので、以下のように proximal operation は表せる。

$$prox_{q|\cdot|}(\mu) = ST_q(\mu) \tag{12}$$

$$ST_q(\mu) = \begin{cases} \mu - q & if\ \mu > q \\ 0 & if\ |\mu| \le q \\ \mu + q & if\ \mu < -q \end{cases} \tag{13}$$

上記を実装し、以下のグラフが得られた。

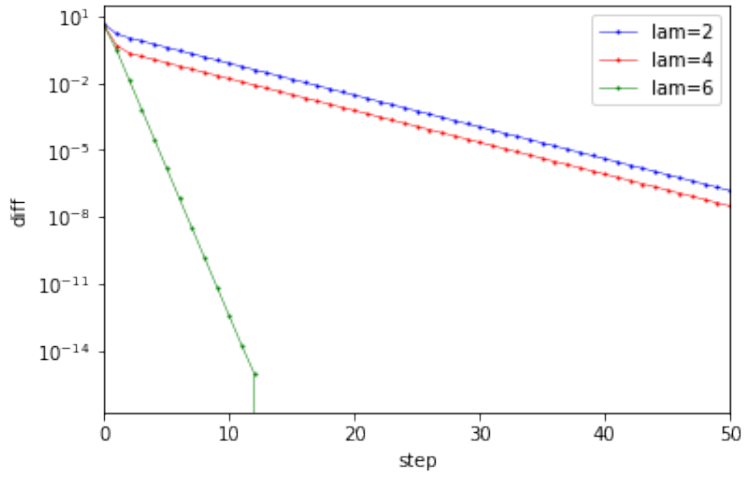
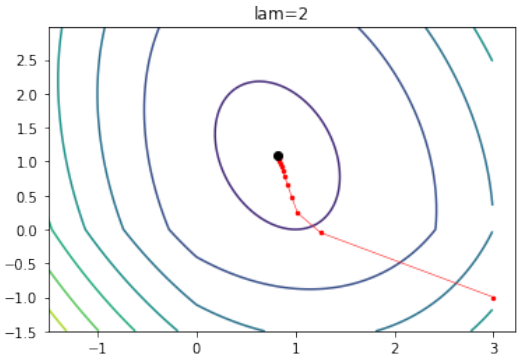
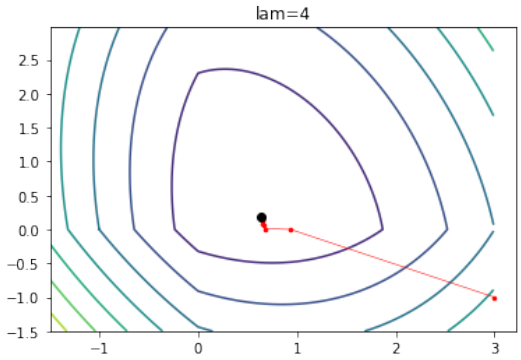
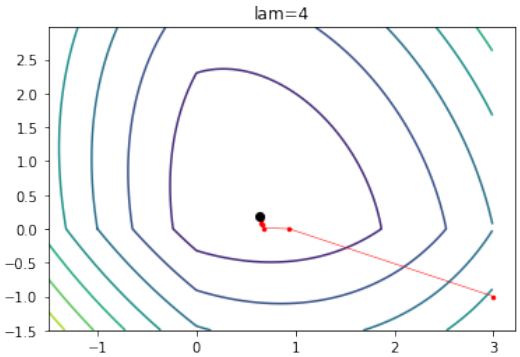


図 1: result







3 PROBLEM 3

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \max(0, 1 - y_i w^T x_i) + \lambda \|w\|_2^2 \right) \tag{14}$$

1, 2

式 (14) にスラック変数を導入して変形していくと、式 (14) の問題は以下のような問題に変形できる。

$$\begin{aligned} &\text{minimize} && \lambda w^T w + 1^T \xi \\ &\text{subject to} && \xi \geq 1 - y_i w^T x_i (, i = 1, ..., n) \\ & && \xi \geq 0 \end{aligned} \tag{15}$$

$\alpha \in \mathbb{R}^n$,$\beta \in \mathbb{R}^d$ としてラグランジュ方程式を定義する。ここで KKT 条件も考えると、以下のようになる。

$$L(w, \xi, \alpha, \beta) = \lambda w^T w + 1^T \xi + \sum_i \alpha (1 - y_i w^T x_i - \xi_i) - \beta^T \xi \tag{16}$$

ラグランジュの導関数：

$$\frac{\partial L}{\partial w} = 2\lambda w - \sum_i \alpha_i y_i x_i \quad \Rightarrow \quad \hat{w} = \frac{1}{2\lambda} \sum_i \hat{\alpha}_i y_i x_i \tag{17}$$

$$\frac{\partial L}{\partial \xi} = 1 - \alpha - \beta \quad \Rightarrow \quad \hat{\alpha_i} + \hat{\beta_i} = 1 \quad \text{i=1,...,n} \tag{18}$$

ラグランジュ変数の不等式：$\hat{\alpha_i} \geq 0$,$\hat{\beta_i} \geq 0$
スラックの相補条件：

$$\hat{\alpha}(1 - \hat{\xi_i} - y_i \hat{w}^T x_i) = 0, \text{ i=1,...,n}$$
$$\hat{\beta_1}\hat{\xi_1} = 0, \text{ i=1,...,n}$$

ここでラグランジュの双対問題を考えると

$$\begin{aligned} &\text{maximize}_\alpha \quad -\frac{1}{4\lambda}\alpha^T K \alpha + 1^T \alpha \\ &\text{subject to} \qquad\qquad 0 \leq \alpha \leq 1 \end{aligned} \tag{19}$$

$$\{K\}_{i,j} = \sum_i y_i y_j x_i^T x_j$$

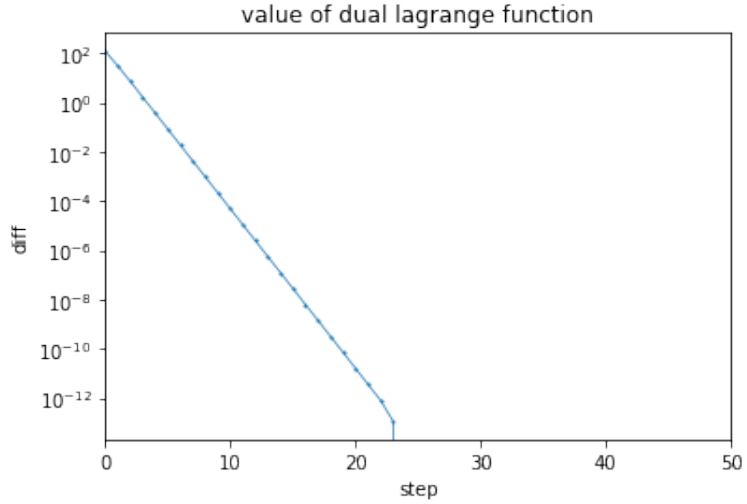
$$\hat{w} = \frac{1}{2\lambda} \sum_i \hat{\alpha}_i y_i x_i$$

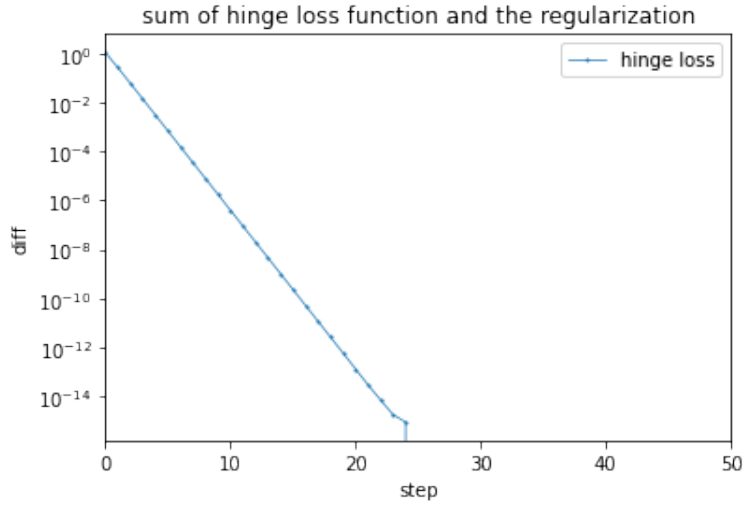
$$\hat{\alpha}_i + \hat{\beta}_i = C$$

以上より、1,2 の問題を示せた。

3

本課題で作成したコードは github(https://github.com/naoyaaan/AdvancedMachineLearning) に公開しています。





実装による計算結果として上図のようなグラフが得られた。最初のグラフは、$w = \hat{w}$ となるときのラグランジュ関数の値との差分の遷移を表したものである。2 つ目のグラフは、$w = \hat{w}$ となるときの hinge loss 関数の値との差分の遷移を表した物である。

4 PROBLEM 4

1

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \max(0, 1 - y_i w^T x_i) + \lambda \|w\|_1 \right) \tag{20}$$

式 (20) にスラック変数 ($\xi_i \geq 1 - y_i w^T x_i \geq 0$,$e_i \geq |w_i| \geq 0$) を導入して変形していくと、式 (20) の問題は以下のような問題に変形できる。

$$\begin{aligned} \text{minimize} \quad & 1^T \xi + \lambda 1^T e \\ \text{subject to} \quad & \xi \geq 1 - y_i w^T x_i (, i = 1, ..., n) \\ & \xi \geq 0 \\ & e_i \geq |w_i| \end{aligned} \tag{21}$$

$$L(w, \xi, \alpha, \beta, \gamma) = 1^T \xi + \lambda 1^T e + \sum_i \alpha_i (1 - y_i w^T x_i - \xi_i) - \beta^T \xi + \gamma(w - e) - \delta e$$

$$\frac{\partial L}{\partial e} = \lambda - \gamma - 1$$

$$\frac{\partial L}{\partial \xi} = 1 - \alpha - \beta$$

$$\frac{\partial L}{\partial w} = \gamma - \sum_i \alpha_i y_i x_i$$

$$z := \begin{pmatrix} \xi & e \end{pmatrix}$$

$$c := \begin{pmatrix} 1 & \lambda \end{pmatrix}$$

$$a := \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$A := \begin{pmatrix} Y & 1 \end{pmatrix}$$

$$B := 1$$

PROBLEM 8

1

まず、$w\hat{_{LS}}$ について考える。

$$\begin{aligned} w\hat{_{LS}} &= \underset{w}{\operatorname{argmin}} \frac{1}{2}||y - Xw||_2^2 && (22) \\ &= \underset{w}{\operatorname{argmin}} \frac{1}{2}(y^T y - 2w^T X^T y + w^T X^T X w) && (23) \end{aligned}$$

ここで勾配を考える。

$$\begin{aligned} &\frac{\partial L(w)}{\partial w}|_{w=w\hat{_{LS}}} = 0 \quad \Rightarrow \quad \frac{\partial L}{\partial w}|_{w=w\hat{_{LS}}} = -X^T y + X^T X\hat{w} = 0 \\ &\hat{w} = (X^T X)^{-1} X^T y \end{aligned} \tag{24}$$

以上より、$w\hat{_{LS}} = (X^T X)^{-1} X^T y$ となる。

次に、$\hat{w}_{ridge}$ について考える。

$$\begin{aligned} \hat{w}_{ridge} &= \underset{w}{\operatorname{argmin}} \frac{1}{2}||y - Xw||_2^2 + \lambda||w||_2^2 && (25) \\ &= \underset{w}{\operatorname{argmin}} \frac{1}{2}(y^T y - 2w^T X^T y + w^T X^T X w) + \lambda w w^T && (26) \end{aligned}$$

ここで勾配を考える。

$$\frac{\partial L(w)}{\partial w}|_{w=\hat{w}_{ridge}} = 0 \quad \Rightarrow \quad \frac{\partial L}{\partial w}|_{w=\hat{w}_{ridge}} = -X^T y + (X^T X + \lambda I)\hat{w} = 0 \tag{27}$$

以上より、$\hat{w}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$ となる。

2

$X^T X$ は対象行列であるので直行行列 $P(P^T = P^{-1})$ で、

$$X^T X = P\Gamma P^T (\Gamma = diag(\gamma_0, ..., \gamma_D)) \tag{28}$$

と対角化が可能である。よって以下のようになる。

$$\begin{aligned} (X^T X + \lambda I)^{-1} &= (P\Gamma P^T + \lambda I)^{-1} \\ &= \left(P(\Gamma + \lambda I)P^T\right))^{-1} \\ &= P(\Gamma + \lambda I)^{-1} P^T \\ &= Pdiag(\frac{1}{\gamma_0 + \lambda}, \frac{1}{\gamma_1 + \lambda}, ..., \frac{1}{\gamma_D + \lambda})P^T \end{aligned}$$

最小の固有値がゼロのとき $\lambda > 0$ であるので、収束する。よって $X^T X + \lambda I$ は正則である。

講義資料の修正箇所について

1 つ目

本課題の midterm-assignment.pdf の資料の Problem 3 の第 2 問において、

$$\hat{w} = \frac{1}{2\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i$$

と α_i となっているが、ここでは

$$\hat{w} = \frac{1}{2\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i$$

のように $\hat{\alpha}_i$ とするべきである。

2 つ目

第 1 回の資料、m1-03.pdf に修正箇所がある。p23 において、「The optimal values $\hat{a}, \hat{b}, \hat{\sigma}$ are...」とあるが、ここでは、$\hat{\sigma}$ ではなく、$\hat{\sigma}^2$ とするべきである。