ART.T458 Advanced Machine Learning Midterm Assignment (Final report assigned by Shimosaka)

21M30821 山上理久

2021 年 7 月 30 日

Problem 1 (10pts)*

ここでは、線形ロジスティック回帰による二値分類を考える。$x \in \mathbb{R}^d$ を d 次元の入力ベクトル、$w \in \mathbb{R}^d$ をモデルのパラメータとする。分類器は，$f(x) = 2[\![w^\top x]\!] > 0 - 1$ で表され，c は，c が真であれば 1 を，そうでなければ 0 を返す指示関数を表す．この最適化問題は次のように書くことができる。教師付きデータセット $\boldsymbol{x}_i, y_i{}_{i=1}^n$ を用いて，ロジスティック回帰の最適化問題を考える．この最適化問題は次のように書くことができる。

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} J(\boldsymbol{w})$$
$$J(\boldsymbol{w}) := \sum_{i=1}^{n} (\ln(1 + \exp(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i))) + \lambda \boldsymbol{w}^\top \boldsymbol{w}.$$

ここでは，$d-1$ 次元の特徴空間でのオフセットに分類器を適応させるために，$\boldsymbol{x}_i$ に定数値 1 が含まれていると仮定する．いくつかの人工的なデータセット（Toy Dataset IV を使用）を用いて、以下のようにいくつかの最適化手法を実装することを考える。

1. バッチ式最急降下法の導入
   バッチ式最急降下法を実装し、二値分類を行った結果と誤分類は下図の通りである。以下、誤分類を表す図では赤が正常な分類、青が誤分類を意味する。
2. ニュートン法を導入
   ニュートン法を実装し、二値分類を行った結果と誤分類は下図の通りである。
3. 上記 2 つの最適化手法の性能を比較するために、$|J(\boldsymbol{w}^{(t)}) - J(\hat{\boldsymbol{w}})|$ w.r.t. t (ここで、$\boldsymbol{w}^{(t)}$ は、t 番目の繰り返しにおけるパラメータを表し、$\hat{\boldsymbol{w}}$ は、2 つの方法で得られた J の最小値に達する最適なパラメータを表す) を片対数プロットで図示する。
4. マルチクラス版ロジスティック回帰（Toy Dataset V を使用）にニュートン法と単純な最勾配法を導入し、上記の二値ロジスティック回帰と同じ実験を行う。
   バッチ式最急降下法を実装し、マルチクラス分類を行った結果と誤分類は下図の通りである。
   ニュートン法を実装し、マルチクラス分類を行った結果と誤分類は下図の通りである。

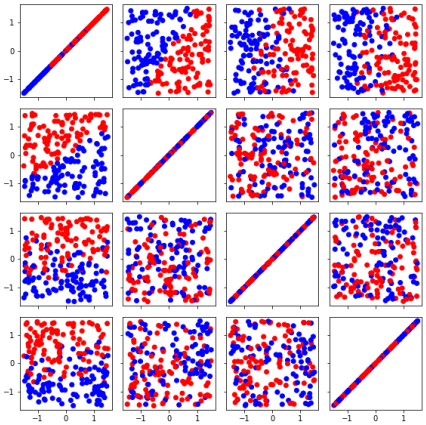


図 1　バッチ式最急降下法の二値分類の結果

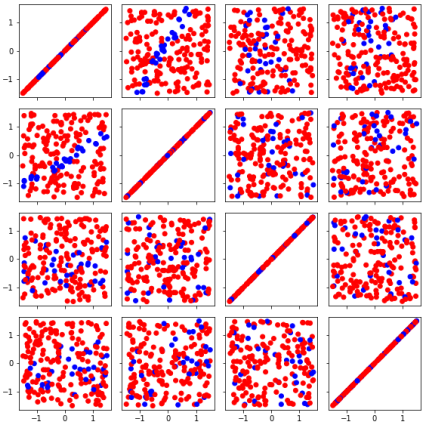


図 2　バッチ式最急降下法の二値分類の誤分類

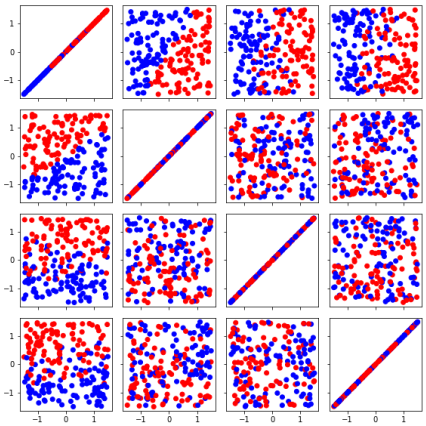


図 3　ニュートン法の二値分類の結果

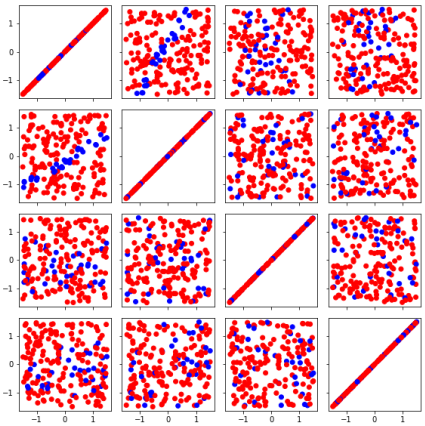


図 4　ニュートン法の二値分類の誤分類

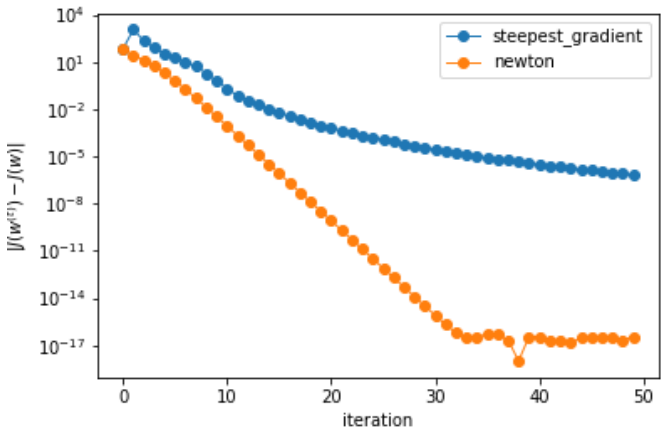


図 5　2 つの最適化手法の性能の比較

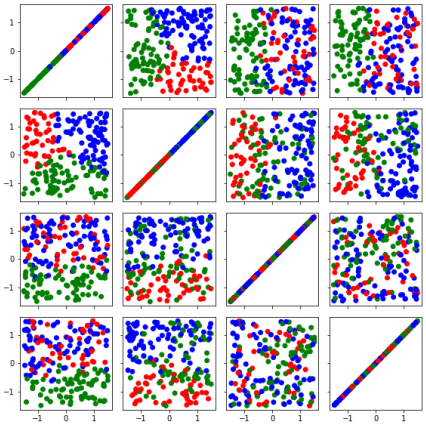


図 6　バッチ式最急降下法のマルチクラス分類の結果

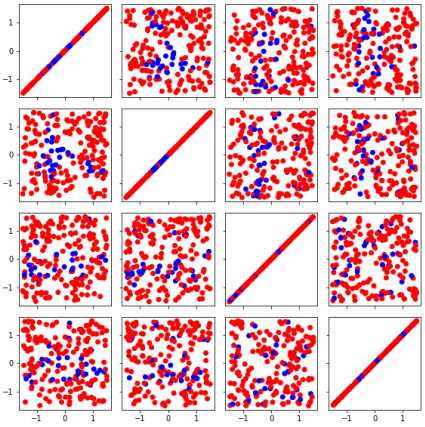


図 7　バッチ式最急降下法のマルチクラス分類の誤分類

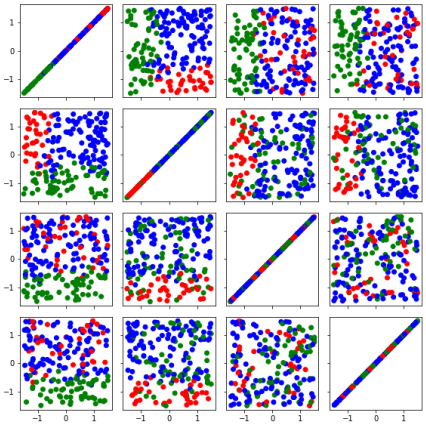


図 8　ニュートン法のマルチクラス分類の結果

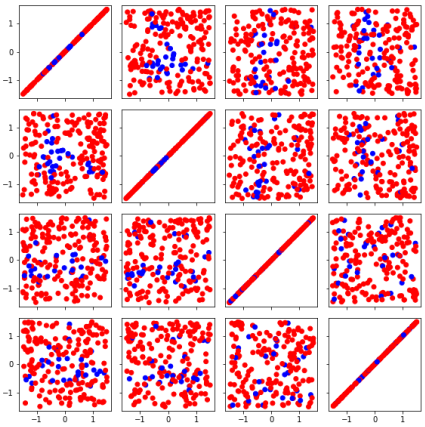


図 9　ニュートン法のマルチクラス分類の誤分類

Problem 2 (5 pts)*

線形回帰に平方損失と L1 正則化を採用した lasso を考える。この問題では、近似勾配法（PG）を採用する。議論を簡単にするために、以下の目的を使用する。

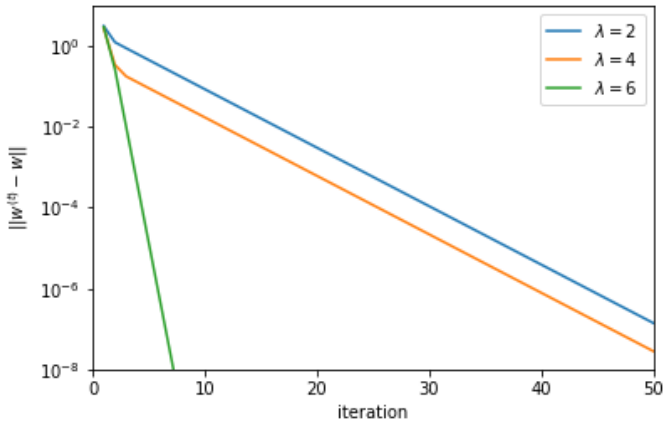
$$\hat{\boldsymbol{w}} = \underset{w}{\operatorname{argmin}} \left((\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{A} (\boldsymbol{w} - \boldsymbol{\mu}) + \lambda ||\boldsymbol{w}||_1 \right).$$

lasso の PG を実装し、いくつかの条件で結果を示す。この問題では、同じ学習率 $\eta_t = L_{-1}$ を使用する。ここで、L は目的の勾配のリプシッツ定数を表し、ヘシアン行列 $2\boldsymbol{A}$ から導かれる（つまり、学習率の逆数として $2\boldsymbol{A}$ の最大固有値を使用する：$\eta_t^{-1} - 1$）。

1. PG の結果を、繰り返し回数に応じて $||\boldsymbol{w}^{(t)} - \hat{\boldsymbol{w}}||$ を片対数プロットで示す。次の条件を使用します。

$$A = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \mu^\top = (12).$$

L1 正則化の特性を確認するために、$\lambda = 2, 4, 6$ で実験を行う。$\lambda = 2, 4, 6$ で実験を行った結果は下図の通りである。



Problem 3 (15 pts)*

サポート・ベクトル・マシンの双対（L2 正則化されたヒンジ・ロスに基づくバイナリ分類器）を考える。この分類の元々の最適化問題は，次のように表される。

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

ここで $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{w} \in \mathbb{R}^d$, $y_i \in \pm 1$, $\lambda > 0$ はそれぞれ i 番目の入力変数、パラメータベクトル、i 番目の入力データのラベル、正則化項の係数を表す。

1. この最適化の双対ラグランジュ関数が次のように書けることを検証する。

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & -\frac{1}{4\lambda}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{1}, \\ \text{subject to} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1} \end{aligned}$$

ここで、$\mathbf{1}, \mathbf{0}$ は、それぞれ 1, 0 を要素とする n 次元のベクトルを表し、$K \in \mathbb{R}^{n \times n}$ は対称正方行列を表し、その i 番目の行と j 番目の列の要素は $y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$ である。
主問題を $\boldsymbol{\xi}$ を用いて再定義する。

$$\begin{aligned} \text{maximize}_{\boldsymbol{w},\boldsymbol{\xi}} \quad & \xi^\top \mathbf{1} + \lambda \boldsymbol{w}^\top \boldsymbol{w} \\ \text{subject to} \quad & \xi_i \ge 1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i \quad i = 1, \cdots, n \\ & \boldsymbol{\xi} \ge \mathbf{0} \end{aligned}$$

ラグランジェ乗数 $\boldsymbol{\alpha}, \boldsymbol{\beta}(\ge \mathbf{0})$ を用いて, 以下のラグランジェ関数を考える。

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\xi}^\top \mathbf{1} + \lambda \boldsymbol{w}^\top \boldsymbol{w} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i - \xi_i\right) + \sum_{j=1}^{n} -\beta_j \xi_j$$

このとき、

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = 2\lambda \boldsymbol{w} - \displaystyle\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \\ \dfrac{\partial L}{\partial \boldsymbol{\xi}} = \mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta} \end{cases}$$

$\frac{\partial L}{\partial \boldsymbol{w}}\Big|_{(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{w}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})} = \frac{\partial L}{\partial \boldsymbol{\xi}}\Big|_{(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{w}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})} = \mathbf{0}$ より、

$$\begin{cases} \hat{\boldsymbol{w}} = \dfrac{1}{2\lambda} \displaystyle\sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i \\ \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}} = \mathbf{1} \end{cases} \tag{1}$$

よって、

$$\begin{aligned} L(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) &= \hat{\boldsymbol{\xi}}^\top \mathbf{1} + \lambda \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{w}} + \sum_{i=1}^{n} \hat{\alpha}_i \left(1 - y_i \hat{\boldsymbol{w}}^\top \boldsymbol{x}_i - \hat{\xi}_i\right) + \sum_{j=1}^{n} -\hat{\beta}_j \hat{\xi}_j \\ &= \hat{\boldsymbol{\xi}}^\top (\mathbf{1} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}) + \lambda || \frac{1}{2\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i ||_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i - \sum_{i=1}^{n} \hat{\alpha}_i y_i \left(\frac{1}{2\lambda} \sum_{j=1}^{n} \hat{\alpha_j} y_j \boldsymbol{x}_j^\top \right) \boldsymbol{x}_i \\ &= \lambda || \frac{1}{2\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i ||_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i - \sum_{i=1}^{n} \hat{\alpha}_i y_i \left(\frac{1}{2\lambda} \sum_{j=1}^{n} \hat{\alpha_j} y_j \boldsymbol{x}_j^\top \right) \boldsymbol{x}_i \\ &= -\frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\boldsymbol{\alpha}}_i \hat{\boldsymbol{\alpha}}_j y_i y_j \boldsymbol{x}_i^\top + \hat{\boldsymbol{\alpha}}^\top \mathbf{1} \\ &= -\frac{1}{2\lambda} \hat{\boldsymbol{\alpha}}^\top \boldsymbol{K} \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\alpha}}^\top \mathbf{1} \quad \left(\{\boldsymbol{K}\}_{i,j} = y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \right) \end{aligned}$$

以上より、$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ を $\boldsymbol{\alpha}, \boldsymbol{\beta}$ とすることによって、

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & -\frac{1}{2\lambda} \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1} \quad (\because \boldsymbol{\alpha} = \mathbf{1} - \boldsymbol{\beta} \leq \mathbf{1}) \end{aligned}$$

が導出される。

2. KKT 条件から、α で与えられる最適な重みパラメータ $\boldsymbol{w}$ が次のように書けることを検証する。

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

式 (1) より、

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

が導出される。

3. 負の双対ラグランジュ関数の最小化を、投影勾配を用いて実施する。妥当性については、双対ラグランジュ関数のスコアと、ヒンジ損失関数と正則化の和を、それぞれ反復回数に対して示す。双対性ギャッ

プが 0 になると収束することを確認する。ここでは、投影勾配がちょうど計算されると仮定する。

$$\boldsymbol{\alpha}^{(t)} = P_{[0,1]^n}\left(\boldsymbol{\alpha}^{(t\ -\ 1)} - \eta_t\left(\frac{1}{2\lambda}\boldsymbol{K}\boldsymbol{\alpha}^{(t\ -\ 1)} - \mathbf{1}\right)\right),$$

ここで、η_t は、t 番目の反復における学習率を表し、$P_{[0,1]^n}$ は、入力された各キャストを $[0,1]$ に投影する投影演算子を表している。

Problem 5 (10 pts + optional 5 pts)**

以下の目的を持つ行列最適化問題を考える。

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{x\times n}}{\operatorname{argmin}}\left(\sum_{i,j\notin Q}|Ai,j\ -\ Zi,j|^2 + \lambda\ ||\boldsymbol{Z}||_*\right)$$

ここで $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ はデータ行列（推薦システムにおけるユーザー対映画の評価データ）を表し、Q にはヌル値も含まれる。また、$||\cdot||_*$ は行列の核ノルムを表し、$\lambda > 0$ は正則化のためのハイパーパラメータを表す。

1. 行列の核ノルムの定義を www から調べて記述しなさい。その定義に加えて、核ノルムを用いた近接写像を定義しなさい。
   核ノルムとは以下で定義され、トレースノルムとしても知られる。

   $$||\boldsymbol{A}||_* = \mathrm{tr}(\sqrt{\boldsymbol{A}^*\boldsymbol{A}}) = \sum_{i=1}^{min\{m,n\}}\sigma_i$$

   核ノルムを用いた近接写像は $\boldsymbol{Z}$ の特異値分解を $\boldsymbol{Z} = \boldsymbol{U}\mathrm{diag}(\sigma(\boldsymbol{Z}))\boldsymbol{V}^\top$ とすると以下で定義される。

   $$\mathrm{prox}_{\lambda||\cdot||_*}(\boldsymbol{Z}) = \boldsymbol{U}\mathrm{diag}(S_\lambda(\sigma(\boldsymbol{Z})))\boldsymbol{V}^\top \tag{2}$$

   $$S_\lambda(\sigma_i) = \begin{cases}\sigma_i - \lambda & \sigma_i \geq \lambda \\ 0 & -\lambda \leq \sigma_i \leq \lambda \\ \sigma_i + \lambda & \sigma_i \leq -\lambda\end{cases} \tag{3}$$

2. データセット（Toy Dataset III）を用いて、この機械学習問題に近接勾配法を実装し、復元されたデータ $\boldsymbol{Z}$ をで示す。データセットをカラーマップを用いて表したものが図であり、Q に含まれるヌル値は白で表されている。近接勾配法を実装し、復元した $\boldsymbol{Z}$ をカラーマップを用いて表したものが図である。

3. (Option: Additional 5 pts 獲得可能) $\boldsymbol{A}$ を復元するための代替アプローチとして非負行列因子分解を実装し、ハイパーパラメータを選択してパフォーマンスを比較する。ここで実装した 2 つの方法の長所と短所を説明してください。

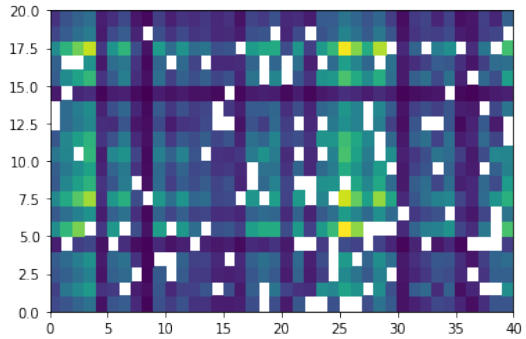


図 10　不完全なデータ行列 $\boldsymbol{A}$

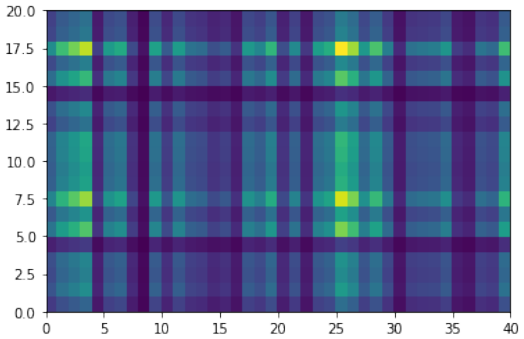


図 11　不完全なデータ行列 $\boldsymbol{A}$ から復元されたデータ $\boldsymbol{Z}$