パターン情報学　プログラミング課題

03-230257 池嶋壮太

課題 1

a) 実装の簡単な説明

擬似逆行列を利用した学習方法のサンプルコードを参考にし、学習部分である fit メソッド内を変更した。具体的には、重みベクトル **W** に関して、**Wx**＋**b** の符号が教師信号と異なる場合にのみパーセプトロンの学習則にそって更新するように変更した。Epoch の回数 for 文を回し、全ての学習パターンを正しく分類できた場合に学習が終了するようにしている。

b) 考察

線形識別関数、2 次の識別関数、3 次の識別関数のパーセプトロンを用いた学習による決定境界の可視化はそれぞれ図 1~図 3 のようになった。正しく学習し分類できている。

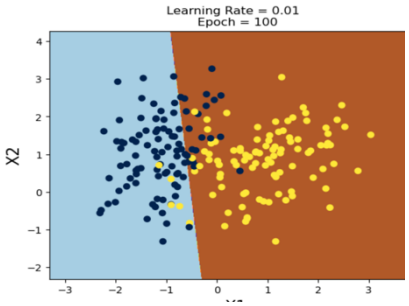

図 1

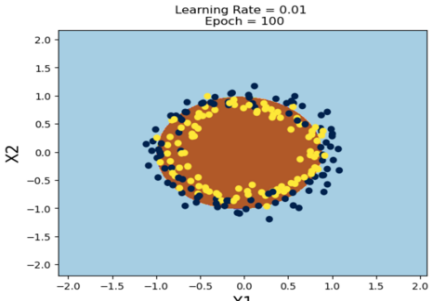

図 2

図 3

続いて、3 次の識別関数を使って、学習係数、エポック数、データノイズの影響について考察する。学習係数は **W** と **b** の更新の大きさを表し、エポック数は訓練データの学習させる回数である。学習係数が大きいと図 4 のように正しく学習できず、0.01 あたりから正しい分類結果が得られた。エポック数は小さいと図 5 のように正しい分類結果は得られず、大きくすると学習に時間を要し、200~500 あたりが適切であった。ノイズに関しては、生成される三日月状データのばらつき具合を示し、ノイズが大きいほど正確な分類が困難になった。noise=0.7 とした場合の結果を図 6 に示す。

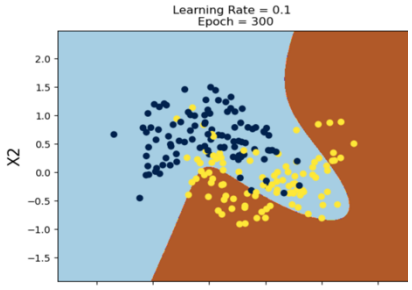

図 4

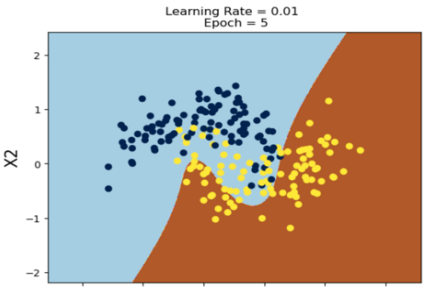

図 5

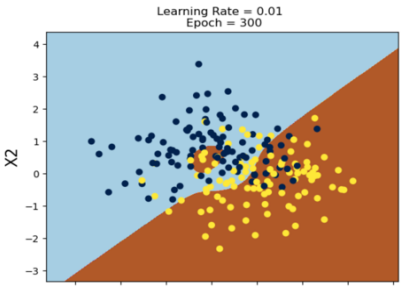

図 6

課題 2

a） 実装の簡単な説明

ソフトマックス関数を定義してソフトマックス回帰を実装した。train_data_split を用いて、入力データ digit_datasets の 80 パーセントを学習データとして学習させ、predict メソッド内で argmax () を用いて最も確率の高いクラスを予測として返している。モデルの性能の評価に関しては、クラスごとの True Positive、False Positive、True Negative、False Negative のテストデータ数を算出し、recall、precision、f-measure、accuracy の定義に代入することて計算している。コードの最後に比較として、sklearn.metrics の classification_report ライブラリを用いた評価の値も掲載した。

b） 考察

擬似逆行列を用いた多クラス分類とソフトマックス回帰を用いた多クラス分類の評価結果を以下の図 7、図 8 に示す。Accuracy に関しては、どちらの分類器を用いた場合でも近い値を取った。precision に関しては、8 個のクラスにおいてソフトマックス回帰による分類器のほうが高い値を示し、recall に関しては、7 個のクラスにおいて擬似逆行列を用いた分類器のほうが高い値を示した。このことから、擬似逆行列を用いた分類器は、recall が高い、すなわち再現性が高く、取りこぼしが少ないと考えられる。一方で、ソフトマックス回帰を用いた分類器は、precision が高い、すなわち正確性が高いと考えられる。precision と recall の定義式から、この二つはトレードオフの関係であるから、このような結果は合理的である。以上から、二つの分類器は僅かな特徴の違いはあるものの、優れた分類器であると結論づけられる。

モデル : pseudo_inverse_multi_classifier

クラス名		precision	recall	f_measure	accuracy
クラス0	:	0.85	1.00	0.92	0.99
クラス1	:	0.95	0.95	0.95	0.99
クラス2	:	0.94	0.94	0.94	0.99
クラス3	:	1.00	1.00	1.00	1.00
クラス4	:	0.91	1.00	0.95	0.99
クラス5	:	0.95	1.00	0.98	0.99
クラス6	:	0.92	0.96	0.94	0.98
クラス7	:	0.95	0.95	0.95	0.99
クラス8	:	1.00	0.83	0.90	0.98
クラス9	:	0.91	0.88	0.89	0.97

図 7

モデル : Softmax_regression

クラス名		precision	recall	f_measure	accuracy
クラス0	:	1.00	1.00	1.00	1.00
クラス1	:	0.84	0.91	0.88	0.97
クラス2	:	0.97	0.94	0.96	0.99
クラス3	:	0.94	1.00	0.97	0.99
クラス4	:	0.94	0.97	0.95	0.99
クラス5	:	0.97	0.95	0.96	0.99
クラス6	:	0.96	0.98	0.97	0.99
クラス7	:	0.95	0.95	0.95	0.99
クラス8	:	1.00	0.87	0.93	0.99
クラス9	:	0.93	0.93	0.93	0.98

図 8

課題 3

a） 実装の簡単な説明

fit メソッドで訓練データと教師データを保持し、predict メソッドで訓練データとテ

ストデータの距離を計算し、dist 変数に格納している。dist 計算において、for 文を 2 回まわす実装では計算量が大きく低速であるため、行列演算によって高速化した。dist が小さい上位 k 個のなかで最も多く現れたクラスを、予測クラスとして返している。交差検証法では、digit_dataset を 2 分割し、2 回の分類結果の正確性の平均をとってプロットしている。

b) 考察

k 近傍法を用いた決定境界の可視化結果は以下の図 9~図 11 のようになった。ただし、すべての分類において k=4 とした。おおむね適切な決定境界を求められている。

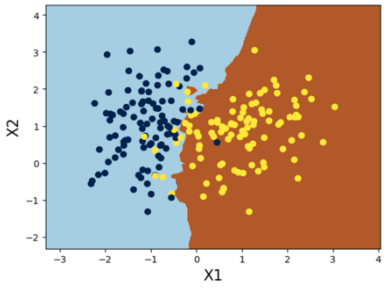

図 9

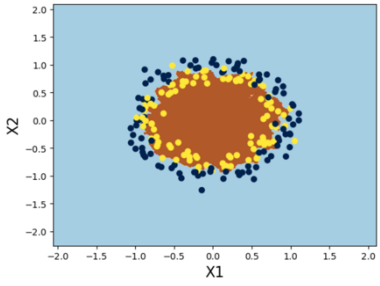

図 10

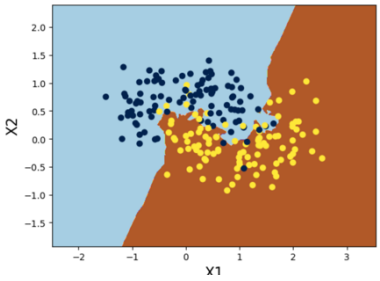

図 11

続いて、digit_datasets を用いた 2 分割交差検証法の結果を図 12 に示す。このグラフから k=1 のとき最も accuracy が高いといえる。また、k が増加するにつれて、accuracy は低下する傾向が認められる。したがって k 近傍法を用いて digit_datasets を多クラス分類する場合には、k=1 として適用するのが最適であると考えられる。このような結果が得られた理由については、データの特性が大きいと考えられる。その裏付けとして、skleran.datasets の load_breast_cancer データに今回の交差検証法を適用したグラフを図 13 に示す。コードは課題の最後に記している。この場合は k=6 で accuracy が最大となっており、データに依存していることがわかる。

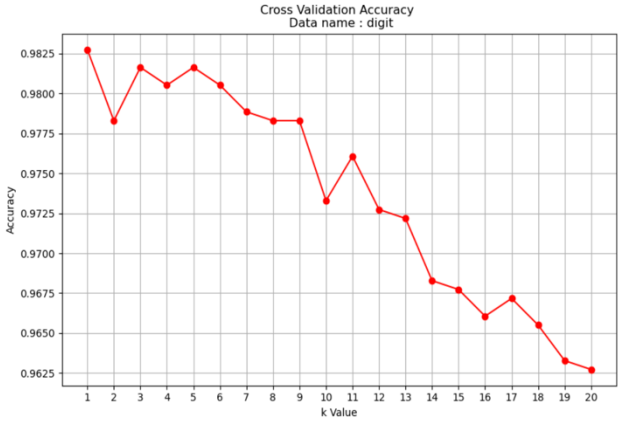

図 12

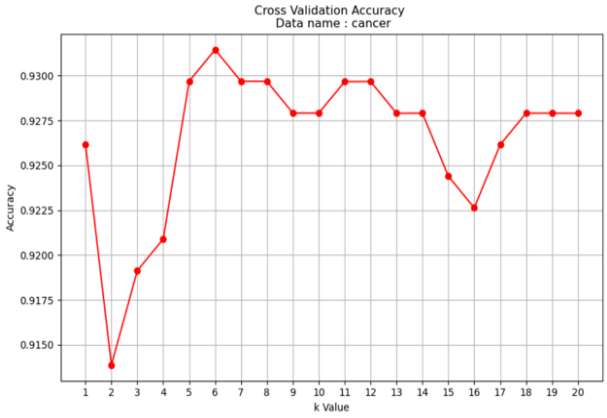

図 13