パターン認識中間レポート 1 澤 祐里

(21-1-037-0801)

2021年6月18日

目次

| 1 | 手順 | | 1 |
|---|---------------------|------------------------------------|---|
| | 1.1 課題 | [1 | 1 |
| | 1.1.1 | k 最近傍法 | 1 |
| | 1.1.2 | ホールドアウト法(最初の 80% を学習・残りの 20% をテスト) | 1 |
| | 1.1.3 | ホールドアウト法(最初の 20% をテスト・残りの 80% を学習) | 2 |
| | 1.1.4 | 5 分割の交差確認法 | 2 |
| | 1.2 課題 | 12 | 3 |
| | 1.2.1 | 最近傍法のハイパーパラメータの最適化 | 3 |
| ^ | 6 + F | | 4 |
| 2 | 結果 | | 4 |
| | 2.1 課題 | [1 | 4 |
| | 2.1.1 | k 最近傍法 | 4 |
| | 2.1.2 | ホールドアウト法(最初の 80% を学習・残りの 20% をテスト) | 5 |
| | 2.1.3 | ホールドアウト法(最初の 20% をテスト・残りの 80% を学習) | 5 |
| | 2.1.4 | 5 分割の交差確認法 | 6 |
| | 2.2 課題 | 12 | 7 |
| | 2.2.1 | 最近傍法のハイパーパラメータの最適化 | 7 |
| 3 | 考察 | | 8 |
| J | | | • |
| | 3.1 課題 | | 8 |
| | 3.2 課題 | $i \gamma$ | 8 |

1 手順

最初に、student_id の部分に「801」を入力してデータを生成しておく。

1.1 課題 1

1.1.1 k 最近傍法

パラメータを表1のように設定して実行する。

表 1 1.1.1 のパラメータ

| パラメータ | 値 |
|--------------|-------|
| train_start | 1 |
| train_end | 400 |
| train2_start | 0 |
| train2_end | 0 |
| test_start | 501 |
| test_end | 10000 |
| k_NN | 3 |

1.1.2 ホールドアウト法 (最初の80%を学習・残りの20%をテスト)

- 1. 全データ 500 の内、最初の 80% を学習データ、残りの 20% をテストデータにするため、400 個を学習データに、100 個をテストデータに分ける
- 2. 上記のことから、1 番目から 400 番目までのデータを学習データに、401 番目から 500 番目のデータをテストデータに する。
- 3. よって、パラメータを表2のように設定して実行する。

表 2 1.1.2 のパラメータ

| パラメータ | 値 |
|--------------|-----|
| train_start | 1 |
| train_end | 400 |
| train2_start | 0 |
| train2_end | 0 |
| test_start | 401 |
| test_end | 500 |
| k_NN | 3 |

1.1.3 ホールドアウト法 (最初の 20% をテスト・残りの 80% を学習)

- 1. 全データ 500 の内、最初の 20% をテストデータ、残りの 80% を学習データにするため、400 個を学習データに、100 個をテストデータに分ける
- 2. 上記のことから、1 番目から 100 番目までのデータをテストデータに、101 番目から 500 番目のデータを学習データに する。
- 3. よって、パラメータを表3のように設定して実行する。

表 3 1.1.3 のパラメータ

| パラメータ | 値 |
|--------------|-----|
| train_start | 101 |
| train_end | 500 |
| train2_start | 0 |
| train2_end | 0 |
| test_start | 1 |
| test_end | 100 |
| k_NN | 3 |

1.1.4 5分割の交差確認法

- 1. 全データ 500 を 5 分割して、それらを 1 つずつテストデータにする。
- 2. その時の学習データには、それぞれの残りのデータを使う。
- 3. よって、パラメータを表4~表8のように設定して、それぞれ実行する。

表 4 testdata:1~100 表 5 testdata:101~200 表 6 testdata:201~300 表 7 testdata:301~400 表 8 testdata:401~500

| パラメータ | 値 | パラメータ |
|--------------|-----|--------------|
| train_start | 101 | train_start |
| train_end | 500 | train_end |
| train2_start | 0 | train2_start |
| train2_end | 0 | train2_end |
| test_start | 1 | test_start |
| test_end | 100 | test_end |
| k_NN | 3 | k_NN |

| パラメータ | 値 |
|--------------|-----|
| train_start | 1 |
| train_end | 200 |
| train2_start | 301 |
| train2_end | 500 |
| test_start | 201 |
| test_end | 300 |
| k_NN | 3 |

| パラメータ | 値 |
|--------------|-----|
| train_start | 1 |
| train_end | 300 |
| train2_start | 401 |
| train2_end | 500 |
| test_start | 301 |
| test_end | 400 |
| k_NN | 3 |

| 値 |
|-----|
| 1 |
| 400 |
| 0 |
| 0 |
| 401 |
| 500 |
| 3 |
| |

1.2 課題 2

1.2.1 最近傍法のハイパーパラメータの最適化

- 1. 全データ 500 の内、401 番目から 500 番目をテストデータとする。
- 2. 残りのデータ 400 から、学習データと検証データの比が 3:1 となるように分けると、それぞれ 300 個と 100 個のデータ に分けられる。
- 3. 上記より、学習データを1番目から300番目とし、検証データを301番目から400番目のデータとする。
- 4. まずは、適当に選んだ5個以上の候補値にkの値を変えながら、検証データをテストに用いて、汎化能力を評価する。
- 5. 上記より、パラメータを表 9~表 17 のように設定して、それぞれ実行する。
- 6. 実行した中で、評価値が一番高くなる k の値を見つける。
- 7. ハイパーパラメータ最適化の結果の評価をするために、表 18 のように、テストに用いるデータを検証データからテストデータに変えた上で、学習に用いるデータに検証データも入れて、評価値が一番高い k の値を入れて再度実行する。

| 表 9 k=1 | | 表 10 k= | =6 | 表 11 k=10 | | 表 12 k=17 | | 表 13 k=24 | |
|--------------|-----|--------------|-----|--------------|-----|--------------|-----|--------------|-----|
| パラメータ | 値 |
| train_start | 1 |
| train_end | 300 |
| train2_start | 0 |
| train2_end | 0 |
| test_start | 301 |
| test_end | 400 |
| k_NN | 1 | k_NN | 6 | k_NN | 10 | k_NN | 17 | k_NN | 24 |

| 表 14 k=30 | | 表 15 k= | 38 | 表 16 k=45 | | 表 17 k=52 | | 表 18 test run | |
|--------------|-----|--------------|-----|--------------|-----|--------------|-----|---------------|-----|
| パラメータ | 値 | パラメータ | 値 | パラメータ | 値 | パラメータ | 値 | パラメータ | 値 |
| train_start | 1 | train_start | 1 | train_start | 1 | train_start | 1 | train_start | 1 |
| train_end | 300 | train_end | 300 | train_end | 300 | train_end | 300 | train_end | 400 |
| train2_start | 0 | train2_start | 0 | train2_start | 0 | train2_start | 0 | train2_start | 0 |
| train2_end | 0 | train2_end | 0 | train2_end | 0 | train2_end | 0 | train2_end | 0 |
| test_start | 301 | test_start | 301 | test_start | 301 | test_start | 301 | test_start | 401 |
| test_end | 400 | test_end | 400 | test_end | 400 | test_end | 400 | test_end | 500 |
| k_NN | 30 | k_NN | 38 | k_NN | 45 | k_NN | 52 | k_NN | 38 |

2 結果

2.1 課題1

2.1.1 k 最近傍法

• スクリプトの出力、精度は図1のようになった。

図1 1.1.1 の出力結果

正解クラスラベル 0 Icついての結果:
クラス 0 Ic識別した個数は 1494.
クラス 1 Ic識別した個数は 808.
クラス 2 Ic識別した個数は 943.
正解クラスラベル 1 Icついての結果:
クラス 0 Ic識別した個数は 825.
クラス 1 Ic識別した個数は 1705.
クラス 2 Ic識別した個数は 556.
正解クラスラベル 2 Icついての結果:
クラス 0 Ic識別した個数は 975.
クラス 1 Ic識別した個数は 753.
クラス 2 Ic識別した個数は 753.
クラス 2 Ic識別した個数は 1442.
TP+TN: 4841個, FP+FN: 4880個, 精度: 0.4884748973792232

• 混合行列は表 19 のようになった。

表 19 1.1.1 の混合行列

| | クラス 0 と推定 | クラス2と推定 | | | |
|----------|-----------|-----------|-----------|--|--|
| クラス 0(P) | 1494 (TP) | 808 (FN) | 943 (FN) | | |
| クラス 1(N) | 825 (FP) | 1705 (TN) | 556 (FN) | | |
| クラス 2(N) | 975 (FP) | 753 (FN) | 1442 (TN) | | |

2.1.2 ホールドアウト法 (最初の80%を学習・残りの20%をテスト)

• スクリプトの出力、精度は図2のようになった。

図 2 1.1.2 の出力結果

正解クラスラベル 0 についての結果: クラス 0 に識別した個数は 25. クラス 1 に識別した個数は 5. クラス 2 に識別した個数は 8. 正解クラスラベル 1 についての結果: クラス 1 に識別した個数は 22. クラス 2 に識別した個数は 7. 正解クラス 0 に識別した個数は 7. 正解クラス 1 に識別した個数は 10. クラス 1 に識別した個数は 8. クラス 2 に識別した個数は 10. TP+TN: 57個, FP+FN: 43個, 精度: 0.57

2.1.3 ホールドアウト法 (最初の 20% をテスト・残りの 80% を学習)

スクリプトの出力、精度は図3のようになった。

図3 1.1.3 の出力結果

正解クラスラベル 0 についての結果:
クラス 0 に識別した個数は 21.
クラス 1 に識別した個数は 1.
クラス 2 に識別した個数は 9.
正解クラスラベル 1 についての結果:
クラス 0 に識別した個数は 10.
クラス 1 に識別した個数は 12.
クラス 2 に識別した個数は 7.
正解クラスラベル 2 についての結果:
クラス 1 に識別した個数は 7.
クラス 1 に識別した個数は 7.
クラス 2 に識別した個数は 7.
クラス 2 に識別した個数は 19.
TP+TN:52個, FP+FN:48個, 精度:0.52

2.1.4 5 分割の交差確認法

• スクリプトの出力、精度はそれぞれ図 4~8 のようになった。

 $\boxtimes 4$ testdata: $1 \sim 100$ $\boxtimes 5$ testdata: $101 \sim 200$ $\boxtimes 6$ testdata: $201 \sim 300$

正解クラスラベル 0 についての結果: 正解クラスラベル 0 についての結果: 正解クラスラベル 0 についての結果: クラス 0 Ic識別した個数は 17. クラス 1 Ic識別した個数は 11. クラス O Iこ識別した個数は 21. クラス 0 に識別した個数は 19. クラス 1 に識別した個数は 1. クラス 1 IC識別した個数は 10. クラス 2 に識別した個数は 9. クラス 2 に識別した個数は 9. クラス 2 Iこ識別した個数は 5. 正解クラスラベル 1 についての結果: 正解クラスラベル 1 についての結果: 正解クラスラベル 1 についての結果: クラス 0 に識別した個数は 10. クラス O IC識別した個数は 5. クラス O IC識別した個数は 15. クラス 1 IC識別した個数は 12. クラス 1 に識別した個数は 18. クラス 1 に識別した個数は 16. クラス 2 に識別した個数は クラス 2 に識別した個数は 7. クラス 2 に識別した個数は 正解クラスラベル 2 についての結果: 正解クラスラベル 2 についての結果: 正解クラスラベル 2 についての結果: クラス 0 に識別した個数は 14. クラス O Ic識別した個数は 16. クラス 0 に識別した個数は 12. クラス 1 に識別した個数は 7. クラス 1 に識別した個数は 4. クラス 1 に識別した個数は 6. クラス 2 に識別した個数は 19. クラス 2 に識別した個数は 15. クラス 2 に識別した個数は 10. TP+TN:52個, FP+FN:48個, 精度:0.52 TP+TN:50個, FP+FN:50個, 精度:0.5 TP+TN: 45個, FP+FN: 55個, 精度: 0.45

図 7 testdata:301~400

図 8 testdata:401~500

正解クラスラベル 0 についての結果:
クラス 0 に識別した個数は 15.
クラス 1 に識別した個数は 5.
クラス 2 に識別した個数は 6.
正解クラスラベル 1 についての結果:
クラス 0 に識別した個数は 18.
クラス 1 に識別した個数は 6.
正解クラス 2 に識別した個数は 6.
正解クラス 1 に識別した個数は 6.
正解クラス 1 に識別した個数は 4.
クラス 1 に識別した個数は 4.
クラス 2 に識別した個数は 4.
クラス 2 に識別した個数は 4.
アキTN: 47個, FP+FN: 53個, 精度: 0.47

正解クラスラベル 0 Icついでの結果:
クラス 0 Ic識別した個数は 25.
クラス 1 Ic識別した個数は 5.
クラス 2 Ic識別した個数は 8.
正解クラスラベル 1 Icついでの結果:
クラス 0 Ic識別した個数は 5.
クラス 1 Ic識別した個数は 22.
クラス 2 Ic識別した個数は 7.
正解クラスの Ic識別した個数は 7.
正解クラス 1 Ic識別した個数は 7.
正解カラス 1 Ic識別した個数は 10.
クラス 2 Ic識別した個数は 10.
TP+TN:57個, FP+FN:43個, 精度:0.57

• 精度を表にまとめると、表 20 のようになった。

表 20 5 分割の交差確認法の精度

| testdata | 精度 |
|----------|------|
| 1~100 | 0.52 |
| 101~100 | 0.45 |
| 201~100 | 0.50 |
| 301~100 | 0.47 |
| 401~100 | 0.57 |

• 各試行の結果を平均した精度は、「0.502」となった。

2.2 課題 2

2.2.1 最近傍法のハイパーパラメータの最適化

• 検証データを使った最近傍法のスクリプトの出力、精度はそれぞれ図 9~17 のようになった。

図 9 k=1

正解クラスラベル D についての結果: クラス O I に識別した個数は 13. クラス 1 に識別した個数は クラス 2 に識別した個数は 正解クラスラベル 1 についての結果: クラス 0 に識別した個数は 10. クラス 1 に識別した個数は 20. クラス 2 に識別した個数は 正解クラスラベル 2 Iこついての結果: クラス 0 Iこ識別した個数は 12. クラス 1 に識別した個数は 9. クラス 2 に識別した個数は 11. TP+TN: 44個, FP+FN: 56個,精度: 0.44

図 13 k=24

```
正解クラスラベル 11 についての結果:
クラス O Iこ識別した個数は 15.
クラス 1 に識別した個数は 8.
クラス 2 I に識別した個数は 3.
正解クラスラベル 1 I についての結果:
クラス 0 に識別した個数は 5.
クラス 1 に識別した個数は 19.
クラス 2 に識別した個数は 18.
正解クラスラベル 2 についての結果:
クラス O に識別した個数は 9.
クラス 1 に識別した個数は 2.
クラス 2 に識別した個数は 21.
TP+TN:55個, FP+FN:45個,精度:0.55
```

図 17 k=52

正解クラスラベル O についての結果: クラス O に識別した個数は 14. クラス 1 Iこ識別した個数は 9. クラス 2 Iこ識別した個数は 3. 正解クラスラベル 1 についての結果: クラス 0 に識別した個数は 6. クラス 1 に識別した個数は 22. クラス 2 に識別した個数は 14. 正解クラスラベル 2 についての結果: クラス O に識別した個数は 10. クラス 1 Iこ識別した個数は クラス 2 Iこ識別した個数は 19. TP+TN:55個, FP+FN:45個, 精度:0.55

図 10 k=6

| 正解クラスラベル 0 についての結果: |
|----------------------------------|
| クラス 0 に識別した個数は 13. |
| クラス 1 に識別した個数は 7. |
| クラス 2 に識別した個数は 6. |
| 正解クラスラベル 1 についての結果: |
| クラス 0 に識別した個数は 14. |
| クラス 1 に識別した個数は 18. |
| クラス 2 に識別した個数は 10. |
| 正解クラスラベル 2 についての結果: |
| クラス 0 に識別した個数は 17. |
| クラス 1 に識別した個数は 1. |
| クラス 2 に識別した個数は 14. |
| TP+TN: 45個, FP+FN: 55個, 精度: 0.45 |

図 14 k=30

```
正解クラスラベル 0 についての結果:
クラス 0 に識別した個数は 15.
クラス 1 に識別した個数は 8.
クラス 2 に識別した個数は 8.
クラス 2 に識別した個数は 3.
正解クラスラベル 1 についての結果:
クラス 0 に識別した個数は 5.
クラス 1 に識別した個数は 22.
クラス 2 に識別した個数は 15.
正解クラスラベル 2 についての結果:
エ肝フラスフ パル 2 12 ついての品
クラス 0 12識別した個数は 8.
クラス 1 12識別した個数は 2.
 クラス 2 に識別した個数は
TP+TN:59個, FP+FN:41個, 精度:0.59
```

| ラス 1 | 1こ識別 | した低 | 鼶対は | 7. | |
|---------|--------|--------|--------|--------------|----|
| ラス 2 | に識別 | した低 | 砂は | 6. | |
| 解クラ. | スラベル | 11 B | こついて | での結果: | : |
| ラス 0 | に識別 | した低 | 砂は | 14. | |
| ラス 1 | に識別 | した低 | 動物は | 18. | |
| ラス 2 | に識別 | した低 | 動物は | 10. | |
| 解クラ. | スラベル | ν2 B | こついて | での結果: | : |
| ラス 0 | に識別 | した低 | 砂は | 17. | |
| ラス 1 | に識別 | した低 | 動物は | 1. | |
| ラス 2 | に識別 | した低 | 砂は | 14. | |
| TH . 40 | c/main | DIEN - | CC/IEI | 维麻・ 0 | 4E |

クラス 0 に識別した個数は 14. クラス 1 に識別した個数は 7. クラス 2 に識別した個数は 5. 正解クラスラベル 1 についての結果: エディンスノール F E 2010 C 9 mil ス クラス 0 I に識別した個数は 10. クラス 1 I に識別した個数は 18. クラス 2 に識別した個数は 正解クラスラベル 2 についての結果: クラス 0 に識別した個数は 13. クラス 1 に識別した個数は 1. クラス 2 に識別した個数は 18. TP+TN:50個, FP+FN:50個, 精度:0.5

図 11 k=10

正解クラスラベル 0 についての結果:

図 15 k=38

正解クラスラベル D についての結果: クラス 0 に識別した個数は 14. クラス 1 に識別した個数は 9. クラス 2 に識別した個数は 正解クラスラベル 1 についての結果: エ解フラスラ (NATIC DATE COME クラス 0 I C識別した個数は 5. クラス 1 IC識別した個数は 23. クラス 2 に識別した個数は 14. 正解クラスラベル 2 についての結果: クラス 0 に識別した個数は 7. クラス 1 に識別した個数は 3. クラス 2 に識別した個数は TP+TN:59個, FP+FN:41個, 精度:0.59

図 12 k=17

正解クラスラベル 0 についての結果: クラス 0 I C識別した個数は 13. クラス 1 I C識別した個数は 8. クラス 2 I C識別した個数は 5. 正解クラスラ ベル1に ついての結果: クラス 0 に識別した個数は 9. クラス 1 に識別した個数は 18. クラス 2 に識別した個数は 正解クラスラベル 2 についての結果: クラス 0 に識別した個数は 8. クラス 1 に識別した個数は 5. クラス 2 に識別した個数は 19. TP+TN:50個, FP+FN:50個, 精度:0.5

図 16 k=45

正解クラスラベル O についての結果: クラス O に識別した個数は 14. クラス 1 に識別した個数は 9. クラス 2 に識別した個数は 3. 正解クラスラベル 1 についての結果: クラス 0 に識別した個数は 5. クラス 1 に識別した個数は 23. クラス 2 に識別した個数は 14. 正解クラスラベル 2 についての結果: クラス 0 I に識別した個数は クラス 1 I に識別した個数は クラス 2 I に識別した個数は TP+TN:59個, FP+FN:41個, 精度:0.59

図 18 k=38 テスト結果

正解クラスラベル 0 についての結果:

クラス 0 に識別した個数は、

クラス 1 に識別した個数は、

クラス 2 に識別した個数は

正解クラスラベル 1 についての結果:

クラス 0 に識別した個数は

クラス 1 に識別した個数は

クラス 2 IC識別した個数は 5.

正解クラスラベル 2 についての結果:

クラス 0 に識別した個数は、

クラス 1 に識別した個数は

クラス 2 に識別した個数は 12.

TP+TN:61個, FP+FN:39個, 精度:0.61

- k=30.k=38.k=45 の時の精度「0.59」が最も高くなった。
- 精度が最も高くなった k=38 で、テストデータを使った結果、図 18 のように、精度は「0.61」となった。

3 考察

3.1 課題1

- ここで計算した値は何に近づくべきなのか 精度は、0 から 1 まで存在し、1 に近づけば近づくほど正しくクラスを識別できているということであるため、1 に近づいた方が良い。
- 1.1.2~5分割の交差確認法で得られた精度に見られる差異とその原因1.1.2~5分割の交差確認法で、入力したパラメータの違いは、学習に用いたデータとテストに用いたデータだけであり、それらの違いで見られた精度の差異については、モデルが学習に用いたデータの誤った挙動なども学習してしまっ

り、それらの違いで見られた精度の差異については、モデルが字習に用いたデータの誤った挙動なども字習してしまったために、未知のデータであるテストデータで識別する際にも、誤った識別をしてしまっていて、それが学習に用いた データとテストに用いたデータが違うだけで精度が変わってしまう原因になったと考えられる。

3.2 課題 2

ハイパーパラメータを変えて、精度を出していった結果、k が 1 に近いときは精度が低く、そこから k を増やすほど精度が上がっていったが、k がある値になると精度が変わらなくなり、やがて、精度が低くなることもあった。これらのことを考察すると、k が 1 に近いときは、誤った挙動をしたデータなどに引っかかりやすくなったため、精度が低くなっていると考えられ、k がある値を超えた時に精度が低くなることもあったのは、データの偏りなどで、上手く分類できなかったと考えられる。