Wine Quality Assessment: Logistic Regression on Chemical Properties

Josuke MINAMIGUCHI

1 Introduction

1.1 Motivation

赤または白ワインの化学的特性 (特徴量) に基づいて，そのワインが品質基準—品質カテゴリの値が 6 以上—を満たすかどうかを予測したい.

1.2 Information of the data

データセットとして，UC Irvine Machine Learning Repository から取得したワインの品質データを利用した．そのデータセットは，赤ワインおよび白ワインの化学的特性 (アルコール含量、pH、密度など) からなる特徴量の列と品質 ($1 \sim 10$) を表す列とから構成される．赤ワインと白ワインのデータ数はそれぞれ，1599, 4898 であり，品質基準を満たすワインのデータ数はそれぞれ，855 (全赤ワインの 53.5 %), 3258 (全白ワインの 66.5 %) であった.

2 Method

特徴量データを標準化したのち，主成分分析を行ってデータ分布の確認を行った.

scikit-learn ライブラリの linear model モジュールの LogisticRegression クラスを用いてロジスティック回帰分析を実行した.

品質基準を満たすワインの品質ラベルを 1, 満たさないワインの品質ラベルを 0 と定義して，11 個のワインの特徴量から品質ラベルを推定するモデルを構築した.

その後，赤ワインと白ワインのそれぞれについて，100 回ずつ，訓練データ・テストデータの分割と学習を行い，評価指標 Accuracy, Precision, Recall, F1 score, Specificity AUC の 100 回の平均値とその誤差から予測を評価した.

また，データセットの性質の議論のために，ROC 曲線もプロットした.

3 Results

3.1 Distributions of the data

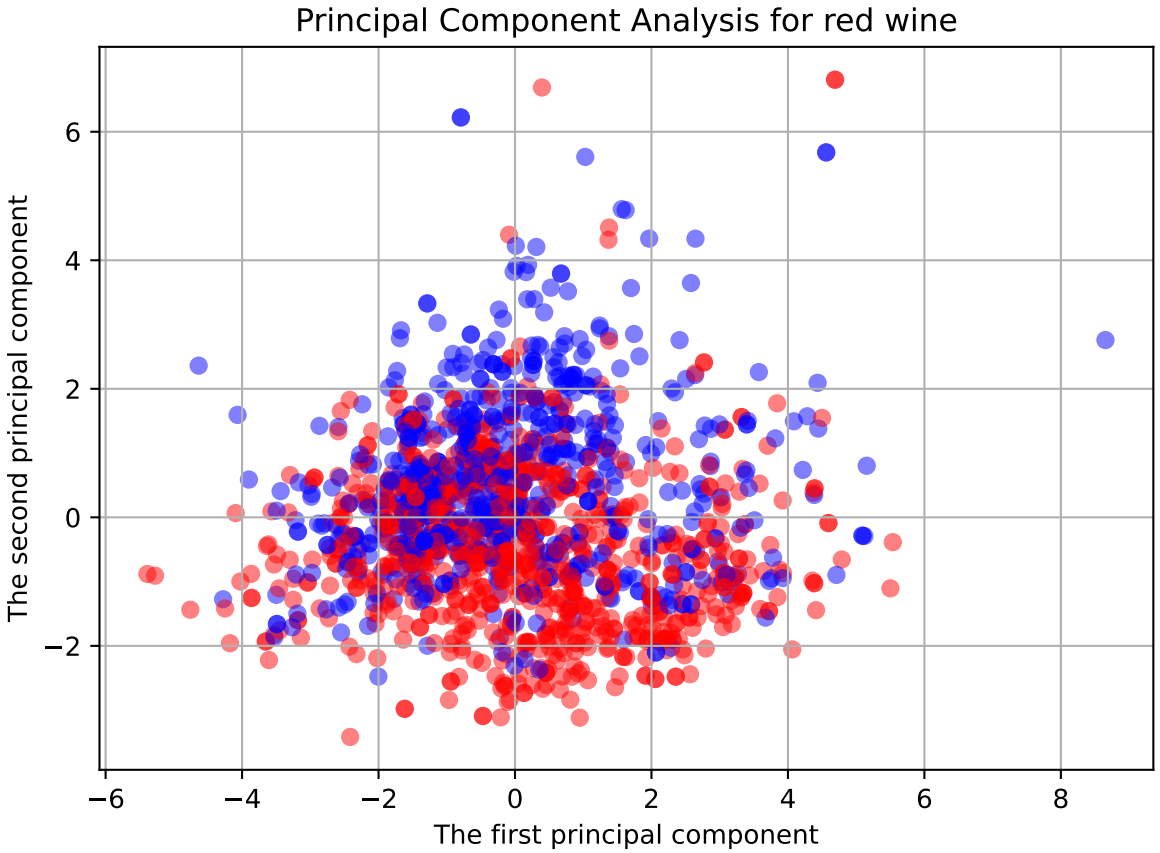

図 1: 標準化した赤ワインのデータの主成分分析

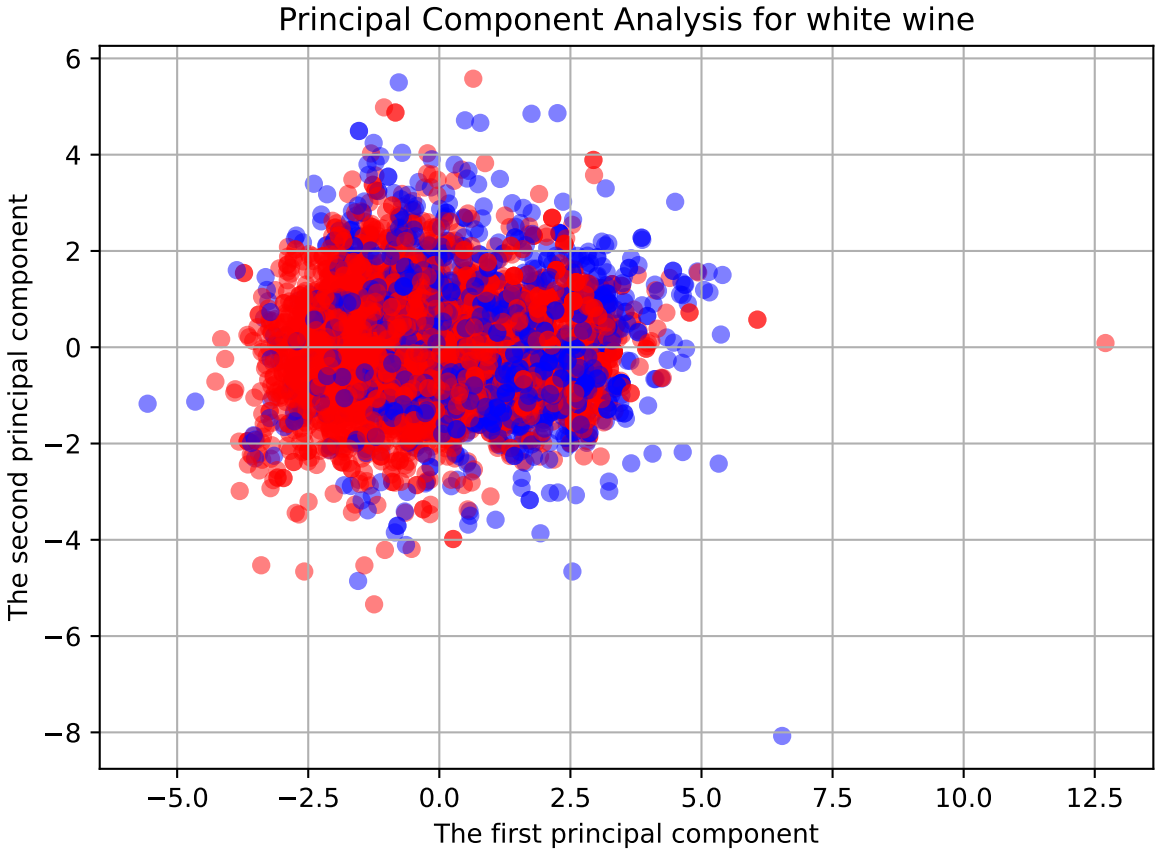

図 2: 標準化した白ワインのデータの主成分分析

3.2 Scores

表 1: 赤および白ワインのスコア

	Accuracy [%]	Precision [%]	Recall [%]	F1 [%]	Specificity [%]	AUC [%]
Red	73.9 ± 0.2	76.2 ± 0.3	74.5 ± 0.3	75.3 ± 0.2	73.2 ± 0.4	81.4 ± 0.2
White	75.2 ± 0.12	77.9 ± 0.2	87.7 ± 0.15	82.5 ± 0.10	50.3 ± 0.3	80.2 ± 0.12

3.3 ROC and AUC

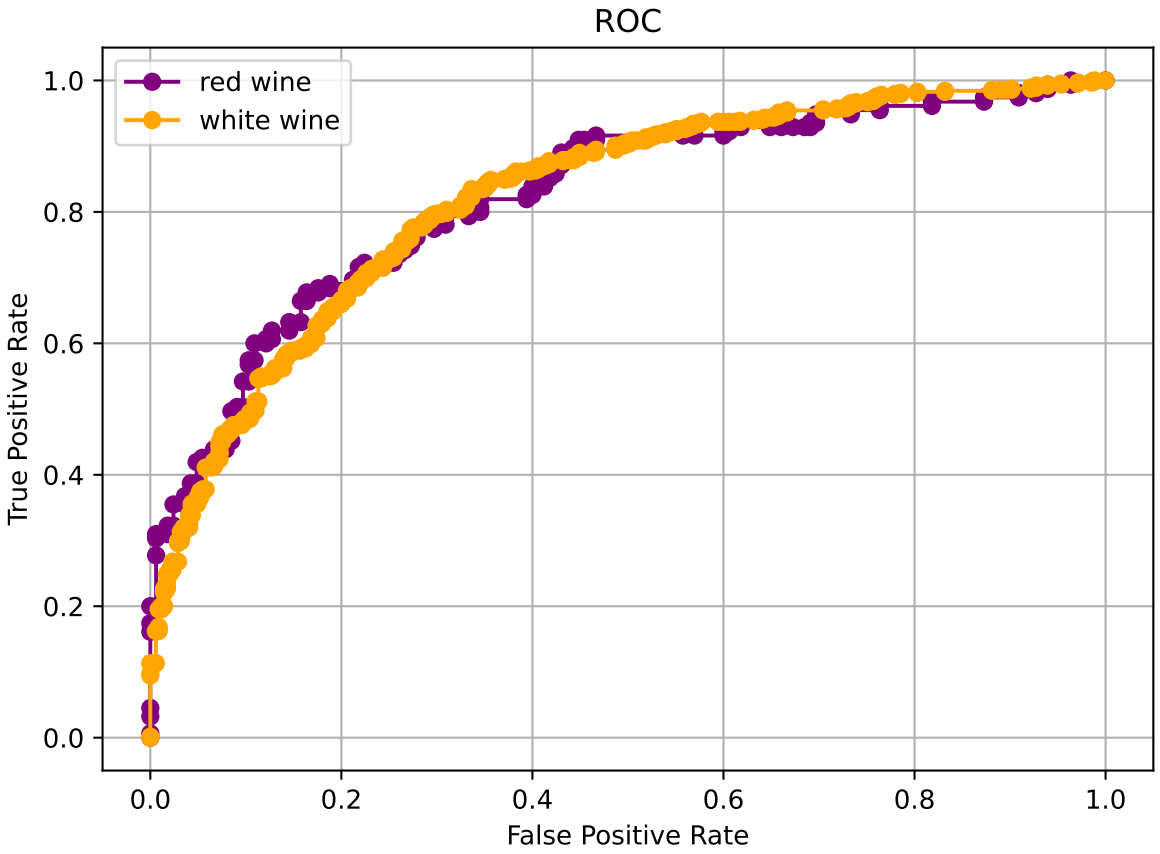

図 3: 赤および白ワインのモデルの ROC.

4 Discussion

赤および白ワインについてロジスティック回帰分析により，その品質を予測した．いずれの評価指標を用いても十分なスコア (90 % 以上) とは言えなかった (表 1).

赤ワインと白ワインの主成分分析の結果 (図 1，2) を比較すると，赤ワインの品質分類の方が，赤点 (品質ラベル 1) と青点 (品質ラベル 0) の区別がはっきりしているため，より分類精度が高まると考えられた.

Recall と Specificity とについては，Accuracy と Precision とに比べて, 赤と白の場合とでスコアが大きく異なっていた (10 % 以上の差)(表 1)．白ワインでは，赤ワインより Recall が有意に高いが，Specificity の値が有意に小さかった．原因としてはもともとのデータにおいて品質基準—品質カテゴリが 6 以上—を満たす白ワインのデータ数が，満たさない白ワインのデータ数に比べて多かったことが挙げられる．実際，全白ワインんのデータのうち 66.5 % が品質基準を満たしている．そのようなデータ数の偏りによって，品質基準を満たさないデータの学習が赤ワインに比べて不十分になり，白ワインの Specificity の値が赤ワインのものより悪くなったと考えた.

F1 スコアは，PresisiontoRecall の調和平均であり,Specificity の値を参照しないことから，白ワインの方が

スコアが良くなっていた (表 1).

AUC の赤ワインと白ワインの違いは，赤ワインのデータ数の少なさも影響していると考えられた (表 1). 実際，ROC も赤ワインの場合はデータ数がすくないために，白ワインの場合と比べて，曲線の揺らぎが大きいことが確認できた (図 3).

以上のことから，本プロジェクトで用いたロジスティック回帰分析と同じ手法でより高精度の予測を実現するためには，(i) 赤ワインについてはデータ数を増やすことと，(ii) 白ワインについては品質基準を満たさないデータ数を増やすこととが必要だ—と考えた.