ワインの種類による要素の違いの調査

201621639　山田純也

UCI Wine Quality Data Set[1]

ワインの品質評価のためのデータセット
ワインの品質評価 (専門家による評価) を
目的変数とし，回帰分析を行うための
データセット　全 **6497 件**
うち赤ワイン 1599 件，白ワイン 4898 件
データはそれぞれ以下の属性を持つ

酒石酸濃度	亜硫酸濃度
酢酸酸度	密度
クエン酸濃度	pH
残留糖濃度	硫酸塩濃度
塩化物濃度	アルコール度数
遊離亜硫酸濃度	品質のスコア

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.51	0.56	9.400000	5	red
1	7.8	0.880	0.00	2.60	0.098	25.0	67.0	0.99680	3.20	0.68	9.800000	5	red
2	7.8	0.760	0.04	2.30	0.092	15.0	54.0	0.99700	3.26	0.65	9.800000	5	red

図 1　データの例 (前処理 1 適用後)

前処理

分析前にデータに対して以下の 5 つの前処理を行った.

1. **赤ワイン・白ワインのデータ結合**
   赤ワインと白ワインのデータを結合した
   また，赤ワインと白ワインの区別をするために
   新たな属性として「ワインの種類」を付加した
2. **品質のスコアの削除**
   データの属性のうち，品質のスコアを削除した
3. **データの正規化**
   データの平均が 0，標準偏差が 1 となるよう正規化
   ワインの種類に関しては行っていない
4. **ワイン種類の数値化**
   ワインの種類を，赤ワインを 1，白ワインを -1 として数値化
5. **データの分割**
   データを訓練データとテストデータに分割
   テストデータは赤ワイン，白ワイン共に 100 件ずつ

分析

ワインの種類を目的変数として回帰分析を行った.
その後，得られたモデルを用いて
データの予測を行った.
データの予測は，訓練データとテストデータ
両方に対して行った.
予測値が 0 より大きければ赤ワイン,
0 以下なら白ワインと予測したとみなして
表 1-3 の正解率を算出した.
表 2,3 の値については，クロス集計後の値をそれぞれ
正解値のデータ数で割った値である.
正解率は高く，回帰分析は正しく行われていると言える.
訓練データ，テストデータ共に赤ワインを白ワインと
誤判断するケースが多少あるのは,
訓練データの白ワインの割合が高いためだと考えられる.

図 2 はこの回帰分析で得られたモデルの
回帰係数の絶対値である.
これが最も高いのが密度 (density) であることから,
ワインの種類によって密度が大きく異なることがわかる.

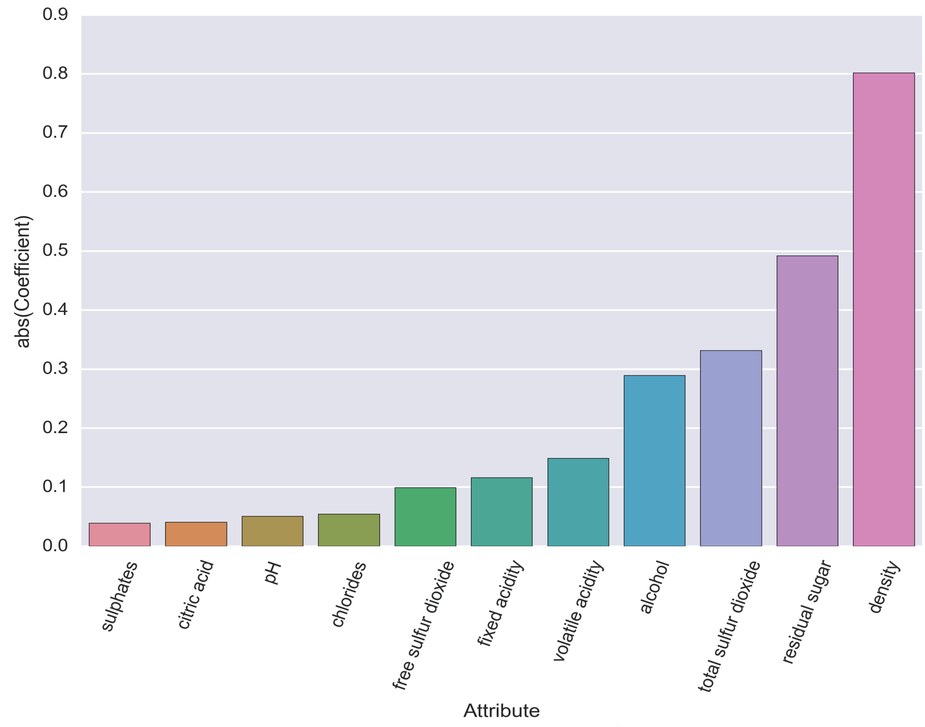

図 2　回帰係数の絶対値

表 1　回帰モデルの正解率

	正解率
訓練データ	0.99
テストデータ	0.96

表 2　訓練データの正解率詳細

		正解値	
		赤	**白**
予測値	**赤**	0.99	0.00
	白	0.01	1.00

表 3　テストデータの正解率詳細

		正解値	
		赤	**白**
予測値	**赤**	0.92	0.00
	白	0.08	1.00

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009).
Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.