

第一回レポート課題（回帰）

- Wine quality prediction dataset

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

- 物理化学的特徴からワインの質（10段階）を予測（回帰）

- 説明変数

- 0 - red or white
- 1 - fixed acidity
- 2 - volatile acidity
- 3 - citric acid
- 4 - residual sugar
- 5 - chlorides
- 6 - free sulfur dioxide
- 7 - total sulfur dioxide
- 8 - density
- 9 - pH
- 10 - sulphates
- 11 - alcohol

- 目的変数

- 12 - quality (score between 0 and 10)



コンペティション形式

- 課題用に一部のデータを利用（ランダムに抽出）

- 学習データ：赤200、白200
- テストデータ：赤500、白500
- 講義ページからダウンロード

- 評価指標

- Mean absolute error
$$\text{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |y_i - \hat{y}_i|$$

- ランキング

- スコアリングサーバ: <http://www.nlab.ci.i.u-tokyo.ac.jp/~nakayama/ds13/report1/index.php>
 - テストサンプルの推定結果を一行ずつ記載したテキストファイルを提出
 - ユーザ名を忘れずに（区別できればなんでもよい。ただし同じ名前をずっと使用すること）
 - 提出した結果は上書きされる。最後のものだけ保存されるので注意。
- 現在は、提出されたテストデータ1000の推定結果のうち、所定の200サンプルでスコアリングしている
- 最終的なスコアは、締め切り後に残りの800サンプルで算出

サンプルプログラム (線形回帰、sample.py)

```
import numpy as np
import pandas as pd
import regression as reg
```

```
nsample = 300
```

```
train_data = pd.read_csv("train.csv")
test_data = pd.read_csv("test.csv")
```

```
train_data = train_data.drop(['red or white'], axis=1) #remove 'red or white' information.
test_data = test_data.drop(['red or white'], axis=1)
```

```
X = np.matrix(train_data.drop(['quality'], axis=1))
y = np.matrix(train_data['quality'])
```

```
XTrain = X[:nsample,:] #use the first 300 samples for training
yTrain = y[:, :nsample]
XVal = X[nsample:,:] #use the rests for validation
yVal = y[:, nsample:]
```

```
w = reg.standRegres(XTrain,yTrain) #linear regression
```

```
yHatTrain = np.dot(XTrain,w)
yHatVal = np.dot(XVal,w)
```

```
print "Training error ", np.mean(np.abs(yTrain - yHatTrain.T))
print "Validation error ", np.mean(np.abs(yVal - yHatVal.T))
```

```
yHatTest = np.dot(np.matrix(test_data),w)
np.savetxt('result.txt', yHatTest)
```

(train.csv)

```
"red or white", "fixed acidity", "volatile acidity", "citric
acid", "residual sugar", "chlorides", "free sulfur dioxide", "total
sulfur dioxide", "density", "pH", "sulphates", "alcohol", "quality"
R, 7.7, 0.705, 0.1, 2.6, 0.084, 9, 26, 0.9976, 3.39, 0.49, 9.7, 5
R, 10.3, 0.27, 0.24, 2.1, 0.072, 15, 33, 0.9956, 3.22, 0.66, 12.8, 6
R, 8.5, 0.37, 0.32, 1.8, 0.066, 26, 51, 0.99456, 3.38, 0.72, 11.8, 6
R, 6.3, 0.3, 0.48, 1.8, 0.069, 18, 61, 0.9959, 3.44, 0.78, 10.3, 6
```

```
5.791953101912569402e+00
4.706598774763293136e+00
5.376977275096278319e+00
6.060673895165684222e+00
6.068585089649218389e+00
5.963326573799613506e+00
5.388982109344969018e+00
5.561331325216619881e+00
5.597390900380342593e+00
5.813484727415301201e+00
5.245247663082410305e+00
5.488444266101371483e+00
5.488444266101371483e+00
5.594112325354758219e+00
```

test.csv と同じ順番で、
各テストサンプルの予
測値が行ずつ入っ
ている。

(result.txt)

課題詳細

- レポート内容
 - Wine quality predictionの問題に取り組み、以下の点を中心にA4用紙1~2枚程度にまとめよ。講義で扱っていない技術を用いても構わない。
- 1. 予測性能を向上させるための自分なりの工夫点と結果、考察（必須）
 - コードを載せる必要はない（もちろん、実装がポイントの場合は載せてかまわない）
- 2. その他、自由にデータを分析した結果（+α）
 - 学習データ数と性能、正則化パラメータの関係
 - 各特徴の寄与の分析
 - 赤ワインと白ワインを判別分析してみる …など
- 3. ここまでの講義の感想、要望など（必須）
 - フィードバックをください！

課題詳細

- どうしても実装が難しい場合…（プログラミング未経験者の方など）
 - 1・2の代わりに、元論文を読み内容をまとめることでOKとします。
 - できれば挑戦して欲しいですが…
- 情報理工の人はだめですよ

課題詳細

- 評価の方針
 - アイデアや試行錯誤の過程を重視
 - スコアが悪くても、しっかり考察してくればOK
(もちろん良くなればプラスに評価しますが)
 - 面白い分析を期待します
- 提出先
 - 以下のアドレスへメールで提出すること（質問もこちらへ）
 - ds2013@nlab.ci.i.u-tokyo.ac.jp
 - 件名は「データサイエンスレポート課題1」
 - 氏名、学籍番号、所属を忘れずに
- 締め切り
 - 12月16日（月）
 - ただし、スコアリングサーバは13日に締め切り、最終結果発表

工夫できそうなところ

- 回帰のオフセット
- カテゴリ変数の扱い（赤 or 白）
 - ダミー変数？
 - 分けてモデルを作る？
- 手法・パラメータチューニング
 - 正則化項の入れ方（リッジ回帰）
 - クロスバリデーション
- 誤差の評価
 - L2? L1?
- 原論文を読んでみる