Excercise_6

October 16, 2019

1 ICAT3190, Module 6, Excercises

1.1 Wine quality determination (Regression)

A chemical analysis was carried out for 1599 red wine samples, after which the quality of each sample was analyzed by experts in scale 0..10.

The chemical anlaysis reveals 11 features for each wine sample, which are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol (concentration). Each feature is a floating point value, and the quality is an integer between 0 to 10.

1.2 Task 1

1.2.1 Read the data

- Read a CSV-file, called `winequality-red.csv`
- Separate 11 first columns and all rows to your design matrix X
- Use the last column, quality, as dependent variable (target)

1.2.2 Make training set and test set

Separate your data X and y to training set (X_train, y_train) which contains 75% of the data and to the test set X_test, y_test which contains 25% of the data.

```
In [2]: ##>>> Some code for bootstrap
        import matplotlib.pyplot as plt
        import pandas as pd
        import numpy as np
```

```
In [1]: ## >> Some tests, do not change
        assert(X.shape==(1599,11))
        assert(y.shape==(1599,))
```

```
        ---------------------------------------------------------------------------

        NameError                                 Traceback (most recent call last)

        <ipython-input-1-4906dee2cafa> in <module>()
```

```
      1 ## >> Some tests, do not change
----> 2 assert(X.shape==(1599,11))
      3 assert(y.shape==(1599,))


    NameError: name 'X' is not defined
```

1.3 Task 2

- Study the data, select a regression algorithm for predicting the quality of the wine, based on it's chemical features. You can assume that the quality is a floating point number.
- Train the regression algorithm using the training data.
- Use cross validatin to test the performance of the regressor and tune it's parameters as good as you can
- Finally test the regressor with the test set
- Report the score (= R^2 = coefficient of determination) of the regressor in the training set, cross validation and in the test set
- Plot the predicted quality against the known quality
- What does the R^2 score tells?
- What is your opinion of the performance? Is there signs of overfitting?

```
In [127]: ### >>>>> Write your code here
```

```
In [129]: ### >>>> Some testing
          # The coefficient of determination should be between 0.4 - 0.5 in the
          # test set depending on your regression algorithm and parameters
```

1.4 Task 3, Select the best features

Like it often is, some features are more important for regression than the others.

- Study which features are the most important for predicting the qulity. You can use LASSO or Elastic Nets to select variables (SelectFromModel), as shown in the lecture notes of Module 6, or ir you use random forests (Extratrees, or boosted trees), they already keep account on most often used features, see also lecture notes of Module 6.
- Plot the quality against the most important feature

```
In [108]: # >> Write your code here
```

```
In [121]: ### >>> Some testing
          # It seems that there are only a few bands which are sufficient for classification.
          # Threfore you should get pretty nice separation of clusters in the scatter plot
          # You can also try to plot a scatter plot with PCA
```