Tutorial_04_Tasks

August 16, 2018

QBUS6850 - Machine Learning for Business

1 Tutorial 4

1.1 Task 1 - OLS, Ridge and LASSO

- Download the Advertising.csv file from Canvas
- Load the data
- Split the data into training and test sets with 80/20 proportion (sales is target, spending in TV, Radio and Newspaper are features)
- Build models for OLS, Ridge and LASSO on the training set. Make sure to optimise the penalty value λ for Ridge and LASSO. You may use the LassoCV and RidgeCV functions to help.
- Determine which model performs best on the test set using the Mean Sqaure Error
- Can you tell me a possible reason for particular performance? (Hint: λ)

About dataset Advertising.csv contains monthly spending in dollars on advertising channels and the corresponding volume of sales for that period.

1.2 Task 2 - Feature Selection and Multiple Regression

- Download the winequality-white.csv file from Canvas
- Load the data
- Split the data into training testing sets (quality is target variable, everything else is a feature)
- Build LASSO and Elastic-Net models on training set
- Print the names of the columns for the selected features from both LASSO and Elastic-Net models
- Predict the quality attributes of a random sample of your choice from the test set

About dataset Some physical and chemical properties of wine was measured and an experts rating of wine quality was also taken. The goal is to model wine quality based on physicochemical tests.

https://archive.ics.uci.edu/ml/datasets/wine+quality

1.3 Task 3 - Logistic Regression

- Download the Smarket.csv file from Canvas
- Load the data, you will need to set parse_dates to True

```
In [1]: import pandas as pd
        smarket_df = pd.read_csv('Smarket.csv', usecols=range(1,10), index_col=0, parse_dates=
```

- Split the data into training and test sets using everything before 2004 as training and 2005 as test sets. We will predict whether the stock goes up or down (direction) based on previous days activity (Lag1 and Lag2).

```
In [2]: # Get everything before and including 2004
        X_train = smarket_df[:'2004'][['Lag1','Lag2']]
        y_train = smarket_df[:'2004']['Direction']

        # Get everything from 2005 onwards
        X_test = smarket_df['2005':][['Lag1','Lag2']]
        y_test = smarket_df['2005':]['Direction']
```

- Build a LogisticRegression model
- Display the confusion matrix and classification report

About dataset This dataset records the previous days stock movements over 5 days and the final direction that the stock finished.

1.3.1 Optional Extra

You should notice that Logistic Regression performs quite poorly. Instead try LDA or QDA.

http://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnaly
http://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantA