原油庫存量之增減方向預測

- 實驗報告 -

Alexis Lin

Partner: Alan Chen

Date Performed: March 10, 2019

Goal

- 取得美國政府提供的能源資料，以每週的原油輸入等各種數值為特徵值，隔週的原油存量為標注資料，以支持向量機等機器學習演算法，訓練一個預估模型。
- 若可以用能源歷史資料預估未來(隔週)的原油存量增減方向，即可藉此預估未來的油價漲跌方向，進而以此作為期貨交易的參考資訊。

Method

- 特徵值 (Features)
  - 取 原油輸入量(日千桶), 原油蒸餾能力(日千桶), 汽油庫存(千桶) 等欄位資料, 將 t-1日至t日之差額的正負作為特徵 值, 差額正 值為1, 負值為0。
- 標注值 (Label)
  - 取 不包括SPR的原油庫存(千桶)等欄位資料, 將 t日至t+1日之差額的正負作為標註 值, 差額正值為1, 負值為0。
- 訓練/測試資料集 (Training Set / Test Set)
  - 以前80%的資料集作為訓練集
  - 以後面20%的資料集作為測試集
- 演算法 (Algrithms)
  - 以 Decision Tree, Support-Vector Machine(SVM) 等多種演算法進行訓練
- 評量方式 (Evaluation)
  - 以正確率(Accurancy)做為評量模型效果優劣的方式

Experiment 1-1

- Datasource :
  - U.S. Energy Information Administration
- Dataset :
  - Weekly Supply Estimates
  - Aug 20,1982 ~ March 1, 2019.
  - Total is 1,899 data points.
- Features : (day t-1 -> day t)
  - D2-B : 2Weekly U.S. Refiner Net Input of Crude Oil (Thousand Barrels per Day)
  - D2-D : Weekly U. S. Operable Crude Oil Distillation Capacity (Thousand Barrels per Calendar Day)
  - D6-I : Weekly U.S. Ending Stocks of Total Gasoline (Thousand Barrels)
  - D6-J : Weekly U.S. Ending Stocks of Finished Motor Gasoline (Thousand Barrels)
  - D6-S : Weekly U.S. Ending Stocks of Gasoline Blending Components (Thousand Barrels)
  - D6-AB : Weekly U.S. Ending Stocks of Distillate Fuel Oil (Thousand Barrels)
  - D7-B : Weekly U.S. Days of Supply of Crude Oil excluding SPR (Number of Days)
- Label : (day t -> day t+1)
  - **D6-E : Weekly U.S. Ending Stocks excluding SPR of Crude Oil (Thousand Barrels)**

Result :

Training a GaussianNB using a training set size of 1519. . .
Accuracy for training set: 0.5313.
Accuracy for test set: 0.6053.

Training a DecisionTreeClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5885.
Accuracy for test set: 0.5158.

Training a SVC using a training set size of 1519. . .
Accuracy for training set: 0.5425.
Accuracy for test set: 0.5816.

Training a AdaBoostClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5550.
Accuracy for test set: 0.5342.

Training a RandomForestClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5885.
Accuracy for test set: 0.5289.

Experiment 1-2

- Datasource :
  - U.S. Energy Information Administration
- Dataset :
  - Weekly Supply Estimates
  - Jan 5, 1990 ~ March 1, 2019.
  - Total is 1,521 data points.
- Features : (day t-1 -> day t)
  - D2-B : 2Weekly U.S. Refiner Net Input of Crude Oil (Thousand Barrels per Day)
  - D2-D : Weekly U. S. Operable Crude Oil Distillation Capacity (Thousand Barrels per Calendar Day)
  - D6-I : Weekly U.S. Ending Stocks of Total Gasoline (Thousand Barrels)
  - D6-J : Weekly U.S. Ending Stocks of Finished Motor Gasoline (Thousand Barrels)
  - D6-S : Weekly U.S. Ending Stocks of Gasoline Blending Components (Thousand Barrels)
  - D6-AB : Weekly U.S. Ending Stocks of Distillate Fuel Oil (Thousand Barrels)
  - D7-B : Weekly U.S. Days of Supply of Crude Oil excluding SPR (Number of Days)
- Label : (day t -> day t+1)
  - **D6-E : Weekly U.S. Ending Stocks excluding SPR of Crude Oil (Thousand Barrels)**

Result :

Training a GaussianNB using a training set size of 1216. . .
Accuracy for training set: 0.5493.
Accuracy for test set: 0.6361.

Training a DecisionTreeClassifier using a training set size of 1216. . .
Accuracy for training set: 0.6028.
Accuracy for test set: 0.5443.

Training a SVC using a training set size of 1216. . .
Accuracy for training set: 0.5526.
Accuracy for test set: 0.5869.

Training a AdaBoostClassifier using a training set size of 1216. . .
Accuracy for training set: 0.5444.
Accuracy for test set: 0.5934.

Training a RandomForestClassifier using a training set size of 1216. . .
Accuracy for training set: 0.6028.
Accuracy for test set: 0.5410.

Experiment 1-3

- Datasource :
  - U.S. Energy Information Administration
- Dataset :
  - Weekly Supply Estimates
  - Jun 4, 2010 ~ March 1, 2019.
  - Total is 1,521 data points.
- Features : (day t-1 -> day t)
  - D2-B : 2Weekly U.S. Refiner Net Input of Crude Oil (Thousand Barrels per Day)
  - D2-D : Weekly U. S. Operable Crude Oil Distillation Capacity (Thousand Barrels per Calendar Day)
  - D6-I : Weekly U.S. Ending Stocks of Total Gasoline (Thousand Barrels)
  - D6-J : Weekly U.S. Ending Stocks of Finished Motor Gasoline (Thousand Barrels)
  - D6-S : Weekly U.S. Ending Stocks of Gasoline Blending Components (Thousand Barrels)
  - D6-AB : Weekly U.S. Ending Stocks of Distillate Fuel Oil (Thousand Barrels)
  - D7-B : Weekly U.S. Days of Supply of Crude Oil excluding SPR (Number of Days)
- Label : (day t -> day t+1)
  - **D6-E : Weekly U.S. Ending Stocks excluding SPR of Crude Oil (Thousand Barrels)**

Result :

Training a GaussianNB using a training set size of 364. . .
Accuracy for training set: 0.6538.
Accuracy for test set: 0.6630.

Training a DecisionTreeClassifier using a training set size of 364. . .
Accuracy for training set: 0.7060.
Accuracy for test set: 0.6304.

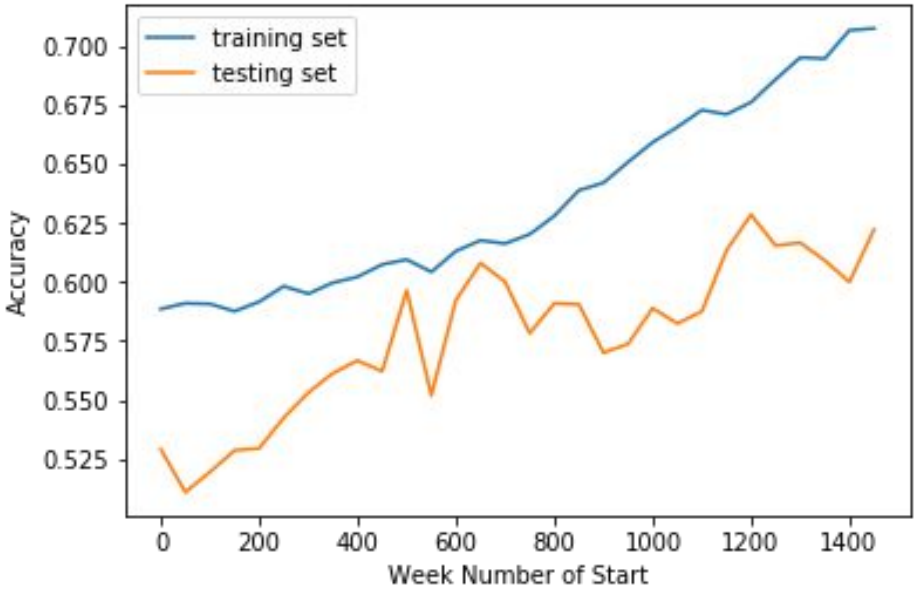
Training a SVC using a training set size of 364. . .
Accuracy for training set: 0.6484.
Accuracy for test set: 0.6739.

Training a AdaBoostClassifier using a training set size of 364. . .
Accuracy for training set: 0.6401.
Accuracy for test set: 0.6630.

Training a RandomForestClassifier using a training set size of 364. . .
Accuracy for training set: 0.7060.
Accuracy for test set: 0.6413.

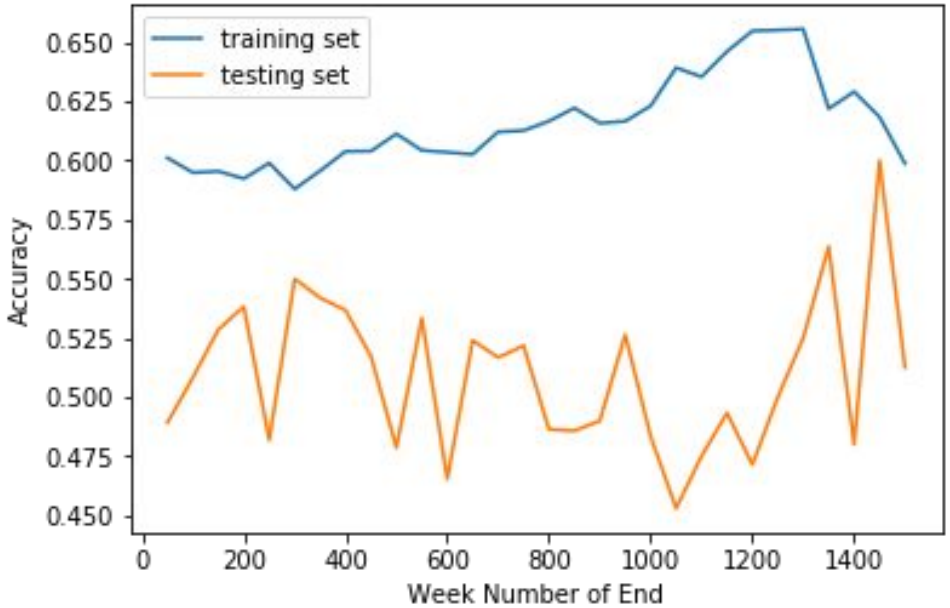
Experiment 1-4

- 方式
  - 全部資料集區間為 Aug 20,1982 ~ March 1, 2019。取第i*50筆往後至最後(最新)的資料為一個數據集。(i為1~30)
  - 以相同特徵值和標注值進行訓練
  - 以RandomForestClassifier訓練
  - 比較不同數據集所產生的正確率
  - 繪出折線圖
- 結論
  - **愈近期的資料(筆數愈少)，正確率愈高。**



Experiment 1-5

- 方式
  - 取第1筆至第i*50筆資料為一個數據集。(i為1~30)
  - 以相同特徵值和標注值進行訓練
  - 以RandomForestClassifier訓練
  - 比較不同數據集所 產生的正確率
  - 繪出折線圖
- 結論
  - **從最舊的資料遞增取得的數據集，在測試集上的正確率呈緩慢上升，但在測試集上波動較大，僅後端呈略為上升趨勢。**



Experiment 2-1

- Datasource :
  - U.S. Energy Information Administration
- Dataset :
  - Weekly Supply Estimates
  - Aug 20,1982 ~ March 1, 2019.
  - Total is 1,899 data points.
- Features : (day t-1 -> day t)
  - D2-B : 2Weekly U.S. Refiner Net Input of Crude Oil (Thousand Barrels per Day)
  - D2-D : Weekly U. S. Operable Crude Oil Distillation Capacity (Thousand Barrels per Calendar Day)
  - D6-I : Weekly U.S. Ending Stocks of Total Gasoline (Thousand Barrels)
  - D6-J : Weekly U.S. Ending Stocks of Finished Motor Gasoline (Thousand Barrels)
  - D6-AB : Weekly U.S. Ending Stocks of Distillate Fuel Oil (Thousand Barrels)
  - D7-B : Weekly U.S. Days of Supply of Crude Oil excluding SPR (Number of Days)
- Label : (day t -> day t+1)
  - **D6-S : Weekly U.S. Ending Stocks of Gasoline Blending Components (Thousand Barrels)**

Result :

Training a GaussianNB using a training set size of 1519. . .
Accuracy for training set: 0.5708.
Accuracy for test set: 0.5974.

Training a DecisionTreeClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5806.
Accuracy for test set: 0.6000.

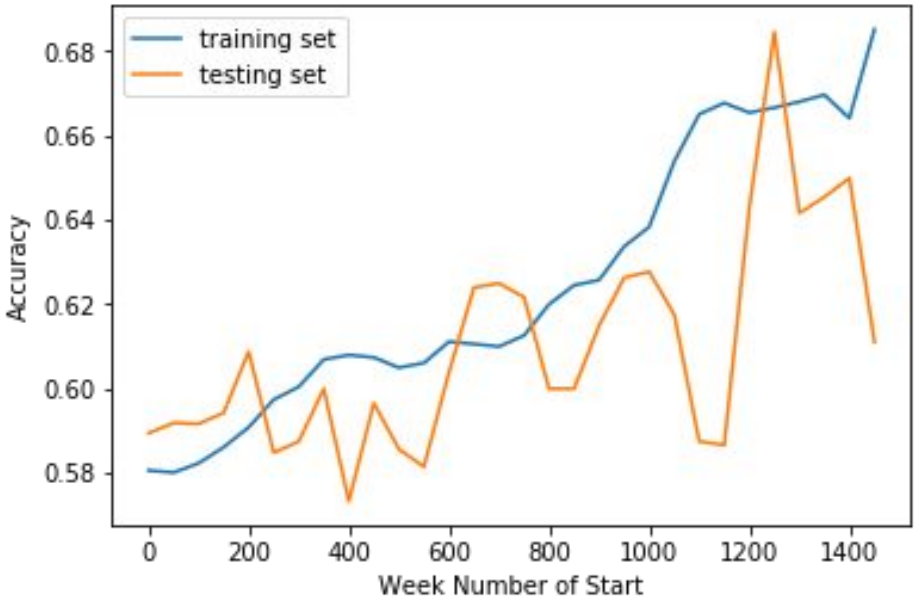
Training a SVC using a training set size of 1519. . .
Accuracy for training set: 0.5721.
Accuracy for test set: 0.6342.

Training a AdaBoostClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5655.
Accuracy for test set: 0.6026.

Training a RandomForestClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5806.
Accuracy for test set: 0.5895.

Experiment 2-2

- 方式
  - 全部資料集區間為 Aug 20,1982 ~ March 1, 2019。取第i*50筆往後至最後(最新)的資料為一個數據集。(i為1~30)
  - 以相同特徵值和標注值進行訓練
  - 以RandomForestClassifier訓練
  - 比較不同數據集所產生的正確率
  - 繪出折線圖
- 結論
  - **愈近期的資料(筆數愈少)，正確率愈高。在訓練集上，呈現較穩定的成長趨勢，但在測試集上，呈現出較大波動。**



Experiment 3-1

- Datasource :
  - U.S. Energy Information Administration
- Dataset :
  - Weekly Supply Estimates
  - Aug 20,1982 ~ March 1, 2019.
  - Total is 1,899 data points.
- Features : (day t-1 -> day t)
  - D2-B : 2Weekly U.S. Refiner Net Input of Crude Oil (Thousand Barrels per Day)
  - D2-D : Weekly U. S. Operable Crude Oil Distillation Capacity (Thousand Barrels per Calendar Day)
  - D6-I : Weekly U.S. Ending Stocks of Total Gasoline (Thousand Barrels)
  - D6-J : Weekly U.S. Ending Stocks of Finished Motor Gasoline (Thousand Barrels)
  - D6-S : Weekly U.S. Ending Stocks of Gasoline Blending Components (Thousand Barrels)
  - D7-B : Weekly U.S. Days of Supply of Crude Oil excluding SPR (Number of Days)
- Label : (day t -> day t+1)
  - **D6-AB : Weekly U.S. Ending Stocks of Distillate Fuel Oil (Thousand Barrels)**

Result :

Training a GaussianNB using a training set size of 1519. . .
Accuracy for training set: 0.5846.
Accuracy for test set: 0.6053.

Training a DecisionTreeClassifier using a training set size of 1519. . .
Accuracy for training set: 0.6050.
Accuracy for test set: 0.5789.

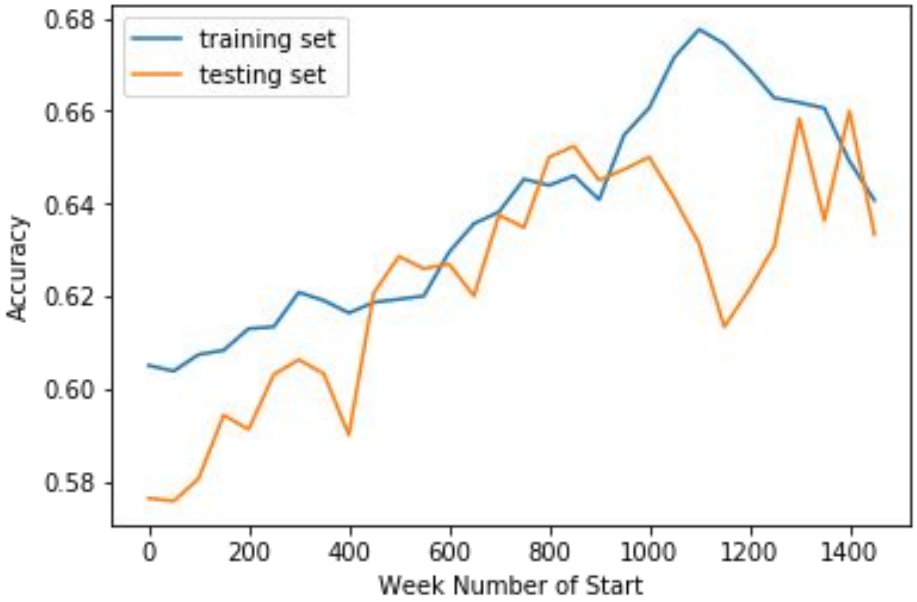
Training a SVC using a training set size of 1519. . .
Accuracy for training set: 0.5846.
Accuracy for test set: 0.6079.

Training a AdaBoostClassifier using a training set size of 1519. . .
Accuracy for training set: 0.5846.
Accuracy for test set: 0.6105.

Training a RandomForestClassifier using a training set size of 1519. . .
Accuracy for training set: 0.6050.
Accuracy for test set: 0.5763.

Experiment 3-2

- 方式
  - 全部資料集區間為 Aug 20,1982 ~ March 1, 2019。取第i*50筆往後至最後(最新)的資料為一個數據集 D。(i為1~30)
  - 以相同特徵值和標注值進行訓練
  - 以RandomForestClassifier訓練
  - 比較不同數據集所 產生的正確率
  - 繪出折線圖
- 結論
  - **愈近期的資料(筆數愈少)，正確率愈高。在訓練集上，呈現較穩定的成長趨勢，在測試集上，呈現略大波動。**



Conclusion

1. 以全部資料(Aug 20,1982 ~ March 1, 2019)進行訓練所得到的正確率，以Exp3的組合(Setting)表現最佳，以RandomForestClassifier訓練而得的模型，在測試集上得到57.63%的正確率。
2. 透過Exp1的分項實驗(1.1~1.3)中發現，以愈近期的資料來進行訓練，雖然資料集變少，但正確率愈高。
3. 第2點的現象，以Exp1的組合，訓練集上的上升趨勢最明顯，測試集上的波動也最小。