Ryan Chen

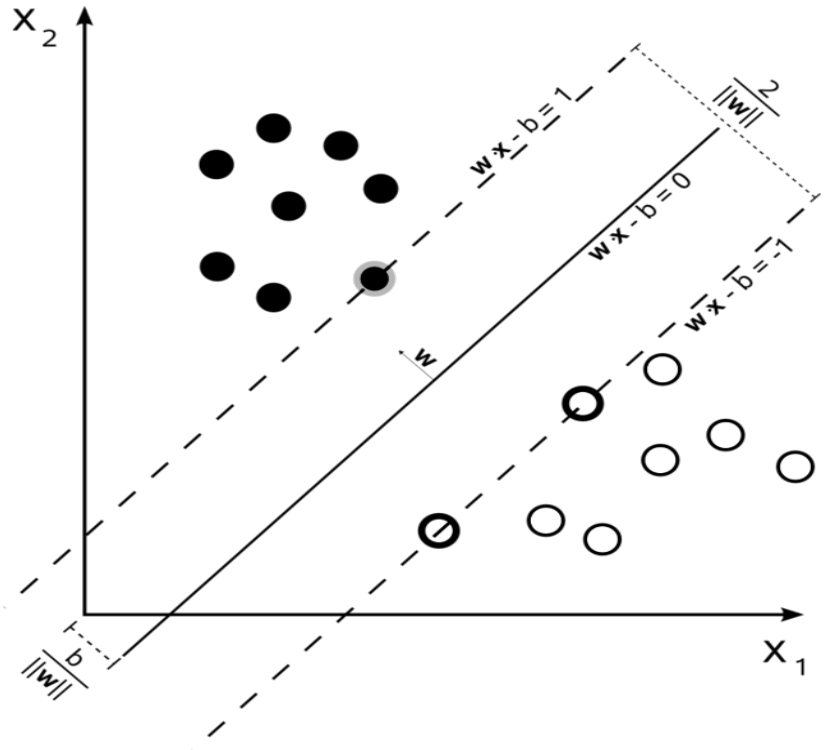
Predicting Stock Return Trend By Using SVM

Abstract

It is widely thought that Asian markets lead the US markets. The purpose of this case is to quantitatively analyze this phenomenon. In this report, I will also predict the US markets daily stock return trend by using support vector machine.

Support Vector Machine (SVM)

A support vector machine a machine-learning model that constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The graph below illustrates how it works. (From Wikipedia)



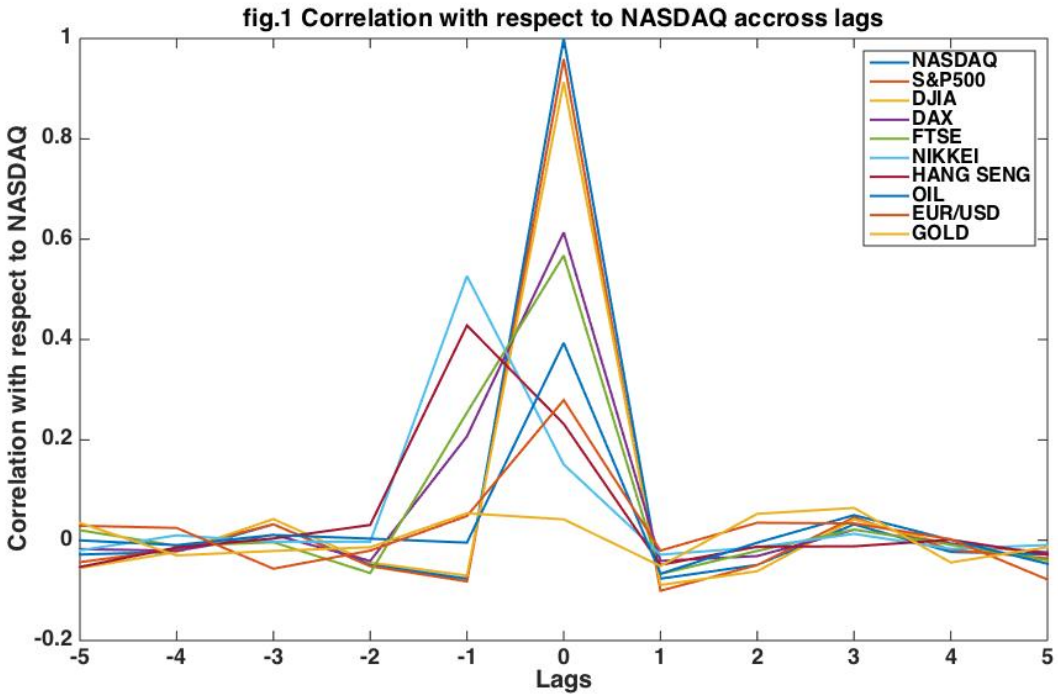
Data Collection

The data set used in this project is listed in table.1 and that covers daily price from 10-April-2006 to 10-April -2015:

Table.1 Data Source

Stock	DJIA, S&P 500, NASDAQ, DAX, FTSE, Nikkei, HANG SENG
Currency	USD/EUR
Commodity	Gold, Oil

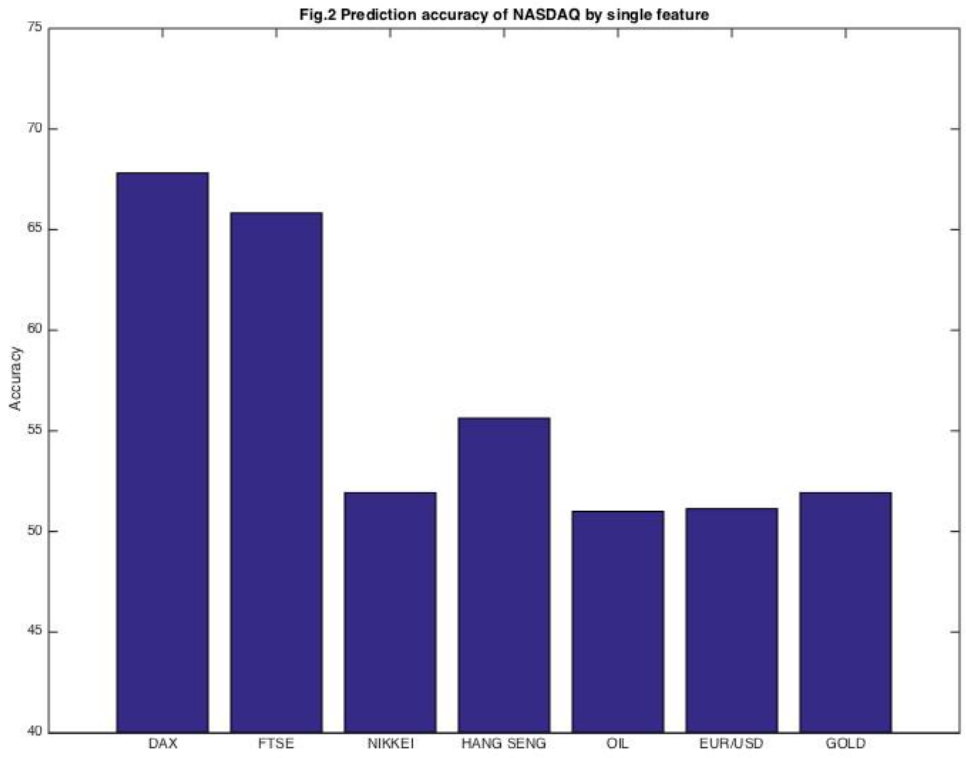
Since the markets are closed on holidays, which are different in different countries, I use NASDAQ stock close price as a basis for data alignment and missing data in other data sources is replaced by linear interpolation. Due to the difference in market value and basis of each market, I normalized the close price to daily returns.



The figure.1 shows the correlations between different market trends across lag range from -5 days to 5 days. Even though S&P 500 and DJIA have strong correlation with NASDAQ on the same day, they are not available to predict the NASDAQ trend. We can see from the graph that DAX, NIKKEI, FTSE, HANG SENG have relative strong correlation with NASDAQ, and since Asian and European markets open earlier than US markets due to the time difference, the information is available before or at the beginning of US markets trading time. Also from graph, we can see that DAX, NIKKEI, FTSE, HANG SENG have strong correlation with NASDAQ when lag is -1, so we can say that Asian and European markets lead US Markets.

SVM Algorithm

Next I will be using SVM to prove the statement quantitatively. Because we only care the trend of the stock market (sign of return), I used libsvm toolbox in Matlab to classify data into two groups (negative return and positive return). I used the first 6 years of data to train the algorithm then made forecasts of the NASDAQ stock market trend on daily basis during the remaining 3 years. Using twice the amount of data for training than for testing is a common practice when applying SVM classifier. The figure.2 shows the accuracy of the prediction based on every single feature. The top 4 features are DAX, FTSE, Hang Seng, NIKKEI with 67.8%, 65.8%, 55.6%, and 51.9% prediction accuracy respectively. Therefore, we can see that European markets have more impact on US markets than Asian markets.



I also predicted the trend of NASDAQ by considering multiple features, which is shown in Table.2

Table.2 Accuracy with multiple features

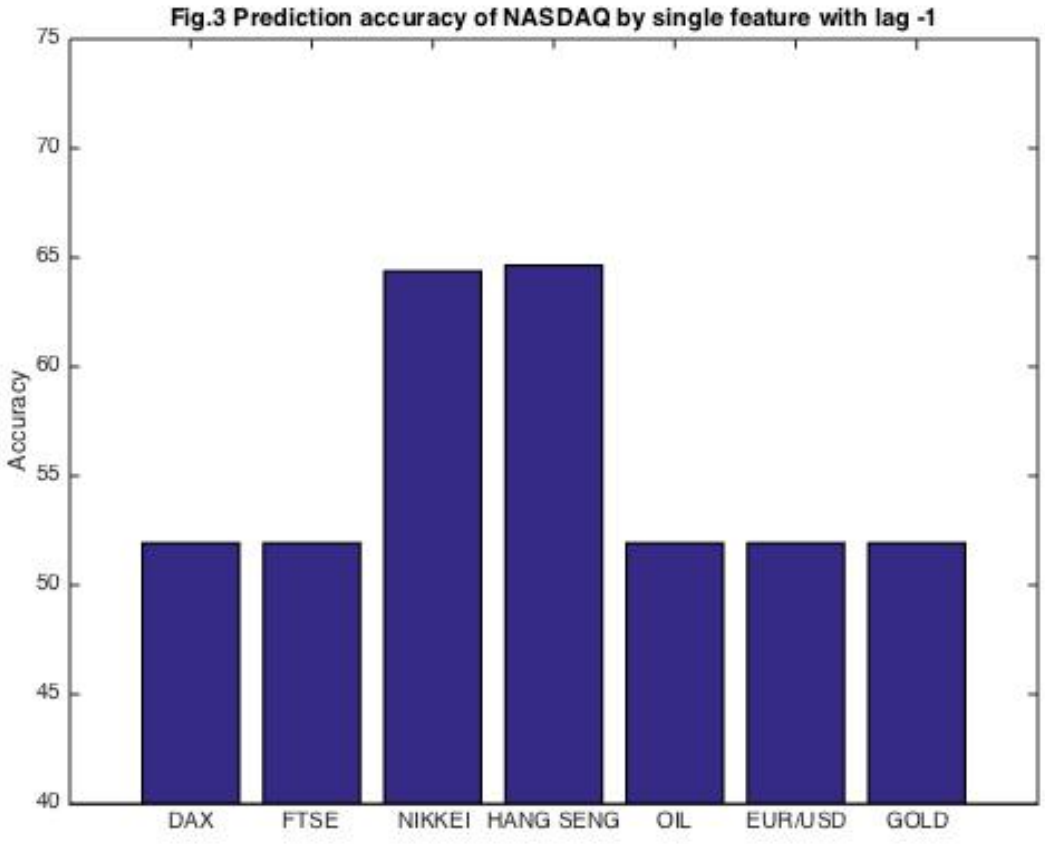
	Accuracy
Top 2 Features	68.08%
Top 4 Features	68.08%
All Features	66.09%

Using the same method, I calculated the accuracy of prediction for S&P 500 and DJIA. The highest accuracy for each index is shown in table.3.

Table.3 Highest Accuracy for US indices

	NASDAQ	S&P 500	DJIA
Accuracy	68.08%	69.8%	69.54%

Since the NIKKEI and HANG SENG have higher correlation with NASDAQ when lag is -1. I implement the SVM classier when the lag is -1. The graph is shown below, which tested that the accuracy of prediction is higher when we use previous days NIKKEI and HANG SENG returns.



Conclusion

In conclusion, Asian and European markets have relatively high correlation with US markets and can provide useful information to help predict the US markets trend on the daily basis. Based on the accuracy of prediction, we can roughly say that given the information of Asian and European markets, which open earlier than US markets, we can predict the trend of US markets 70% correctly. Therefore, the trading strategies based on the European and Asian Markets can be viable but they need to be more sophisticated.

A Few Thoughts On SVM

Machine learning is becoming more and more popular in the financial industry and applying SVM to stock market is promising. The most important thing to implement the SVM model is to pick the right features and the parameters of the model to improve the accuracy of the prediction, at the same time we need to balance between overfitting and underfitting problems. In the future studies, we can consider more inputs and features so that we can actually predict the numerical value of daily stock price and returns and build trading models and algorithms based on the prediction.

Also SVM is a good approach to classify and select stocks form the stock universe based on different features such as risk, growth rate, capitalization, so that we can diversify the portfolio by picking assets from each group and build models to determine weighs that minimize the risk and maximize the returns.

Ryan Chen

Reference

S. Shen, H. Jiang, T. Zhang, “Stock Market Forecasting Using Machine Learning Algorithms,” 2012
C. King, C.Vandrot, J. Wang,“A SVM Approach to Stock Trading”, 2001.

Matlab Source Code

```
format longG;
c =yahoo;
%download data from Yahoo
SP = fetch(c,'^GSPC','close','4/10/2006','4/10/2015');
NAS = fetch(c,'^IXIC','close','4/10/2006','4/10/2015');
DAX = fetch(c,'^GDAXI','close','4/10/2006','4/10/2015');
FTSE = fetch(c,'^FTSE','close','4/10/2006','4/10/2015');
NIkkei= fetch(c,'^N225','close','4/10/2006','4/10/2015');
HS= fetch(c,'^HSI','close','4/10/2006','4/10/2015');
DJIA = fetch(c,'DIA','close','4/10/2006','4/10/2015');
GLD= fetch(c,'GLD','close','4/10/2006','4/10/2015');
USO= fetch(c,'USO','close','4/10/2006','4/10/2015');
EUR=csvread('exchange.csv');

close(c)

%Interplate daily price based on NASDAQ
DAX=interp1(DAX(:,1),DAX(:,2),NAS(:,1));
FTSE=interp1(FTSE(:,1),FTSE(:,2),NAS(:,1));
NIkkei=interp1(NIkkei(:,1),NIkkei(:,2),NAS(:,1));
HS=interp1(HS(:,1),HS(:,2),NAS(:,1));
SP=interp1(SP(:,1),SP(:,2),NAS(:,1));
DJIA=interp1(DJIA(:,1),DJIA(:,2),NAS(:,1));
GLD=interp1(GLD(:,1),GLD(:,2),NAS(:,1));
USO=interp1(USO(:,1),USO(:,2),USO(:,1));
EUR=interp1(EUR(:,1),EUR(:,2),NAS(:,1));
NAS=interp1(NAS(:,1),NAS(:,2),NAS(:,1));

NAS2=ones(length(NAS)-1,1);
DAX2=ones(length(NAS)-1,1);
FTSE2=ones(length(NAS)-1,1);
NIkkei2=ones(length(NAS)-1,1);
HS2=ones(length(NAS)-1,1);
DJIA2=ones(length(NAS)-1,1);
SP2=ones(length(NAS)-1,1);
USO2=ones(length(NAS)-1,1);
EUR2=ones(length(NAS)-1,1);
GLD2=ones(length(NAS)-1,1);

%Calculate daily returns of indices
for i=1:length(NAS)-1;
   NAS2(i) = (NAS(i+1)-NAS(i))/NAS(i);
   DAX2(i) = (DAX(i+1)-DAX(i))/DAX(i);
   FTSE2(i) = (FTSE(i+1)-FTSE(i))/FTSE(i);
   NIkkei2(i) = (NIkkei(i+1)-NIkkei(i))/NIkkei(i);
   HS2(i) = (HS(i+1)-HS(i))/HS(i);
   SP2(i) = (SP(i+1)-SP(i))/SP(i);
   DJIA2(i) = (DJIA(i+1)-DJIA(i))/DJIA(i);
   USO2(i) = (USO(i+1)-USO(i))/USO(i);
   EUR2(i) = (EUR(i+1)-EUR(i))/EUR(i);
   GLD2(i) = (GLD(i+1)-GLD(i))/GLD(i);
end

index=[NAS2,SP2,DJIA2,DAX2,FTSE2,NIkkei2,HS2,USO2,EUR2,GLD2];
```

```
%Calculate correlation with repect to NASDAQ accross different lags
lag=ones(size(index,2),11);
for i=1:size(index,2);
   for j=-5:5;
      if j<=0;
      cor=corrcoef(index(1:length(index)+j,i),index(-j+1:length(index),1));
      lag(i,j+6)=cor(1,2);
      end
      if j>=1;
      cor=corrcoef(index(j+1:length(index),i),index(1:length(index)-j,1));
      lag(i,j+6)=cor(1,2);
      end
   end
end

x=-5:5;
plot(x,lag','LineWidth', 2)
legend('NASDAQ','S&P500','DJIA','DAX','FTSE','NIKKEI','HANG SENG','OIL','EUR/USD','GOLD')
set(gca,'FontSize',18,'fontWeight','bold')
xlabel('Lags') % x-axis label
ylabel('Correlation with respect to NASDAQ') % y-axis label
title('fig.1 Correlation with respect to NASDAQ accross lags')
%% Install libsvm tool box
mex -setup
make
%% Classiy NASDAQ daily returns into 2 groups, -1 and 1.
for i=1:length(index);
   for j=1:size(index,2)
     if index(i,j)<0;
        index(i,j)=-1;
     else index(i,j)=1;
     end
   end
end
% Using svm to predict the trend of NASDAQ index
% Train the data
ind=fix(2*length(index)/3);
trainlabel=index(1:ind,1);
testlabel=index(ind:end,1);
Accuracy=ones(1,size(index,2)-3);
for i=4:size(index,2)
   traindata=index(1:ind,i);
   % Use cross validation to find the best parameters
   [bestacc,bestc,bestg] = SVMcg(trainlabel,traindata,-2,2,-2,2,3,0.5,0.5,0.5);
   cmd = ['-c ',num2str(bestc),' -g ',num2str(bestg)];
   model = svmtrain(trainlabel,traindata,cmd);
   % Test and predict the trend of NASDAQ index
   testdata=index(ind:end,i);
   [predict_label, accuracy, prob_estimates] = svmpredict(testlabel, testdata, model);
   Accuracy(1,i-3)=accuracy(1);
end
bar(Accuracy);
set(gca,'XTickLabel',{'DAX','FTSE','NIKKEI','HANG SENG','OIL','EUR/USD','GOLD'})
ylabel('Accuracy') % y-axis label
ylim([40 75])
title('Fig.2 Prediction accuracy of NASDAQ by single feature')
%% Test multiple features
ind=fix(2*length(index)/3);
trainlabel=index(1:ind,3);
traindata=index(1:ind,4:5);
```

```
% Use cross validation to find the best parameters
[bestacc,bestc,bestg] = SVMcg(trainlabel,traindata,-2,2,-2,2,5,0.5,0.5,2);
cmd = ['-c ',num2str(bestc),' -g ',num2str(bestg)];
model = svmtrain(trainlabel,traindata,cmd);
% Test and predict the trend of NASDAQ index
testdata=index(ind+1:end,4:5);
testlabel=index(ind+1:end,3);
[predict_label, accuracy, prob_estimates] = svmpredict(testlabel, testdata, model);
%% Using svm to predict the trend of NASDAQ index with lag -1
% Train the data
ind=fix(2*length(index)/3);
trainlabel=index(2:ind,1);
testlabel=index(ind+1:end,1);
Accuracy=ones(1,size(index,2)-3);
for i=4:size(index,2)
   traindata=index(1:ind-1,i);
   % Use cross validation to find the best parameters
   [bestacc,bestc,bestg] = SVMcg(trainlabel,traindata,-2,2,-2,2,3,0.5,0.5,0.5);
   cmd = ['-c ',num2str(bestc),' -g ',num2str(bestg)];
   model = svmtrain(trainlabel,traindata,cmd);
   % Test and predict the trend of NASDAQ index
   testdata=index(ind:end-1,i);
   [predict_label, accuracy, prob_estimates] = svmpredict(testlabel, testdata, model);
   Accuracy(1,i-3)=accuracy(1);
end
bar(Accuracy);
set(gca,'XTickLabel',{'DAX','FTSE','NIKKEI','HANG SENG','OIL','EUR/USD','GOLD'})
ylabel('Accuracy') % y-axis label
ylim([40 75])
title('Fig.3 Prediction accuracy of NASDAQ by single feature with lag -1')

function [bestacc,bestc,bestg] = SVMcg(train_label,train,cmin,cmax,gmin,gmax,v,cstep,gstep,accstep)
%SVMcg cross validation by faruto
%% about the parameters of SVMcg
if nargin < 10
   accstep = 1.5;
end
if nargin < 8
   accstep = 1.5;
   cstep = 1;
   gstep = 1;
end
if nargin < 7
   accstep = 1.5;
   v = 3;
   cstep = 1;
   gstep = 1;
end
if nargin < 6
   accstep = 1.5;
   v = 3;
   cstep = 1;
   gstep = 1;
   gmax = 5;
end
if nargin < 5
   accstep = 1.5;
   v = 3;
   cstep = 1;
   gstep = 1;
```

```
    gmax = 5;
    gmin = -5;
end
if nargin < 4
    accstep = 1.5;
    v = 3;
    cstep = 1;
    gstep = 1;
    gmax = 5;
    gmin = -5;
    cmax = 5;
end
if nargin < 3
    accstep = 1.5;
    v = 3;
    cstep = 1;
    gstep = 1;
    gmax = 5;
    gmin = -5;
    cmax = 5;
    cmin = -5;
end
%% X:c Y:g cg:acc
[X,Y] = meshgrid(cmin:cstep:cmax,gmin:gstep:gmax);
[m,n] = size(X);
cg = zeros(m,n);
%% record acc with different c & g,and find the bestacc with the smallest c
bestc = 0;
bestg = 0;
bestacc = 0;
basenum = 2;
for i = 1:m
    for j = 1:n
        cmd = ['-v ',num2str(v),' -c ',num2str( basenum^X(i,j) ),' -g ',num2str( basenum^Y(i,j) )];
        cg(i,j) = svmtrain(train_label, train, cmd);

        if cg(i,j) > bestacc
            bestacc = cg(i,j);
            bestc = basenum^X(i,j);
            bestg = basenum^Y(i,j);
        end
        if ( cg(i,j) == bestacc && bestc > basenum^X(i,j) )
            bestacc = cg(i,j);
            bestc = basenum^X(i,j);
            bestg = basenum^Y(i,j);
        end

    end
end
%% to draw the acc with different c & g
[C,h] = contour(X,Y,cg,60:accstep:100);
clabel(C,h,'FontSize',10,'Color','r');
xlabel('log2c','FontSize',10);
ylabel('log2g','FontSize',10);
grid on;
```