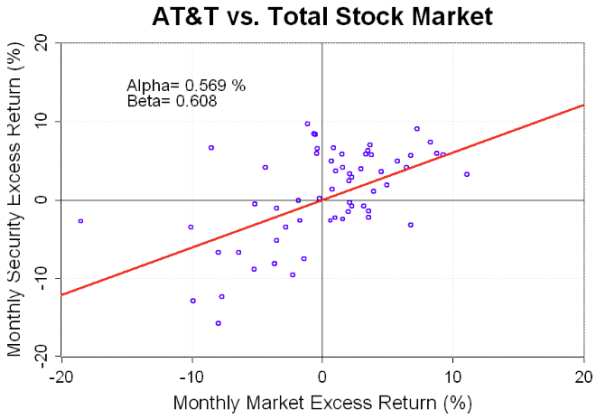
Machine Learning Engineer Nanodegree
Capstone Proposal

Jaehyoung Yoo 03/04/2017

Domain Background

Efficient market hypothesis says that it is almost impossible to beat the market, where market means market index such as SP500 index. CAPM(capital asset pricing model) provides a groundwork for this hypothesis. It is controversial whether Efficient Market Hypothesis is true or not. Hedge funds and active portfolio managers can only prove this hypothesis is wrong by delivering higher return that beats market performance to their investors.



Equation for Capital Asset Pricing Model

$$Expected\ Return = r_f + \beta(r_m - r_f)$$

$r_f = risk\ free\ rate$
$\beta = Beta$
$r_m = return\ on\ the\ market$

Figure 1: Captial Asset Pricing Model

Problem Statement

To prove we can actually beat the market, we will use machine learning models to predict outputs such as market trends and optimized allocation for portfolios. We will use time series data such as price, volume, interest rate, and others as input data. With those input data, we will do various tasks such as clustering into two labels which are bull market[1] and bear market[2]. After clustering, we will teach classification model that predicts whether it is bull or bear market.

Datasets and Inputs

There are several open sources for historical stock price data which are free to use:
Yahoo! Finance: Can directly query stock through the web API, or download .csv files
Bloomberg API: Multiple APIs available, including Python.
Quandl API: Also multiple APIs, including Python.
I will use adjusted close price, because this compensates for any price anomaly resulting from dividends or stock split. Also we will use volume, interest rate(3,10year T-Bond), and others as

[1] A bull market is a period of generally rising prices
[2] A bear market is a general decline in the stock market over a period of time

well as input feature for each stock. Period for historical data is from 2010 to 2016, and currently I plan to include every stock that was included in SP500(based on 2010) including SP500 index itself too.

Solution Statement

We will use traditional CAPM with ML in order to refute Efficient Market Hypothesis.

1) First, use historical data(i.e. daily return for stocks) to cluster series into two classes using unsupervised learning: bull market, bear market.

2) Then, with market status(bull, bear) as label, we will train our classification model which outputs market status from input features. Our solution is to select high beta stock in bull market, and low beta stock in bear market. In rise of stock price, high beta stocks will bring higher return than market. On the other hand, our loss will be lesser than the market in declining period if we have low beta stocks.

3) As mentioned above, depending on the market status, we will select appropriate stocks for our portfolio. When selecting stocks, we will stick to principles from CAPM. We are not thinking of reinforcement learning here but will find simple threshold approach that guarantees our overall return better than the market.

4) Finally, we will refine allocation of selected stocks in our portfolio to maximize portfolio's return. Here, we are looking for refining optimal portfolio allocation on monthly basis. We will approach this in two manners and compare the results from each other. First, we will use previous monthly or yearly data to get optimal allocation(weight) by maximizing cost function. Then, we will use this as a ground for next month portfolio settings. Second, rather than just conveying optimal allocation to next month, we will use it as a monthly training label. We will construct a regression model which predicts optimal weight for each stocks based on inputs. Details are explained in Project Design part.

Benchmark Model

Our benchmark return will be SP500 index daily return. We will see how our learned portfolio can beat the SP500 index.

Evaluation Metrics

Dealing with time series data, we will use roll forward cross validation which guarantees training data always being before testing data. This is to avoid data leakage or looking into future which can lead to unrealistically optimistic result. For performance evaluation metrics in regression, we will use Root Mean Square Error (RMSE) to tell how concentrated the data is around the line of best fit. For classification models, we will use accuracy for our evaluation metrics since our data is not skewed.

Project Design

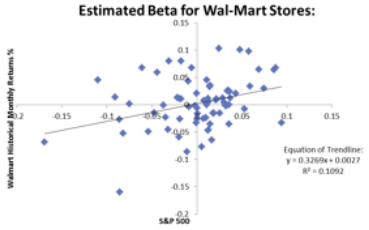
1) Use historical data to cluster into two classes: bull, bear market

2) Train classification model to predict market status from input features

	T-Bond 3Y	T-Bond 10Y	S&P500	MARKET
2007-11-12	1.2	2.4	32.1	BULL
2007-11-13	1.3	2.3	32.3	BULL
2007-11-14	1.2	2.4	32.2	BULL
2007-11-15	1.2	2.5	32.5	BULL
2007-11-16	1.2	2.6	32.6	BEAR
2007-11-17	1.3	2.7	32.7	BEAR

3) Depending on market status, select appropriate stocks in portfolio

:We will calculate beta and alpha for every stocks based on CAPM.



4) Find optimized allocation for portfolio that maximize return.

: We are looking for optimal allocation of portfolio on monthly basis, not just one time optimization. And we will try two approaches` here.

(1) Optimize based on maximizing cost function from previous month.

(2) With optimized allocation for every month as train label, we use machine learning to predict allocation from coming inputs.

(1) Cost function for deriving optimal allocation in each time period

$X = \{x_1, x_2, \cdots, x_n\}$,

x_i means allocation for each stock

Constraint on X

1) $0 \le x_i \le 1$

2) $\sum_0^n x_i = 1$ 0

Cost function (price of stock i at day t)

$\underset{X}{\text{argmax}} f(X) = \sum_0^n x_i \cdot \left(\frac{p_t}{p_0} - 1\right)$, ($p_0$: price of stock i at day 0)

Step

1) Start with inital X

2) Run optimizer to to find X that maximizes $f(X)$

(2) Derive optimal allocation for each time period based on cost function, then use it as training the model