Machine Learning Engineer Nanodegree

Capstone Proposal

George Fedorneac
September 10, 2018

Proposal

Trading Strategy based on PCA

Domain Background

I have opted to apply my knowledge in machine learning to the investment and trading domain. This project focuses on practical application of PCA to stock portfolio optimization.

The idea is based on Markowitz (1952)'s mean-variance theory that makes the foundation of modern portfolio theory. Harry Markowitz, in 1952, published a paper on "Modern Portfolio Theory" for which he also received the Nobel Price in Economics. The main concepts are:

1. The investor's goal is to maximize return for any level of risk
2. Risk can be reduced by creating a diversified portfolio of unrelated assets

Lowenfeld (1909) also discussed the benefits of diversification and at times Lowenfeld's work is looked at as the first rigorous academic discussion of diversification.

While there are other more sophisticated modeling methods being explored in stock market pricing I strongly believe as a new machine learning practitioner I need to understand the fundamentals first.

Problem Statement

My problem statement is simple and direct. To build a stock portfolio script as a simple trading strategy. An abundance of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process. By building a mathematical model for portfolio construction based on principal component analysis (PCA), I will expect the results to be quantifiable, measurable and replicable.

Datasets and Inputs

This project will use stocks price historical information using S&P 500 Index stock data. It contains 2768 rows of daily closing price of 419 US stocks including the S&P 500 Index spanning 10 years from 2003 to 2013.

In addition, I will use one additional dataset for benchmarking. It is composed of stock closing prices of two indices VTI (Vanguard Total Stock Market ETF) and AGG(iShares Core U.S. Aggregate Bond ETF).

Solution Statement

Here we have a number of assets that represent stock prices of corporate America and I want to find some way to get the optimal mix of stocks. In other words, I should have percentages (weights) for each asset in my portfolio and I do not want to do this qualitatively I want to do this mathematically. A mathematical solution is quantifiable, measurable, and replicable.

The portfolio weights are determined using principal component analysis (PCA) and are based on Absorption Ratio Delta, which equals the fraction of the total variance of a set of asset returns explained by a fixed number of eigenvectors. Absorption Ratio is a measure of implied systemic risk as presented in a paper from MIT "Principal Components as a Measure of Systemic Risk" in 2010 by Mark Kritzman (http://web.mit.edu/finlunch/Fall10/PCASystemicRisk.pdf)

Benchmark Model

A common benchmarking method of stock portfolios performance is to compare a against a portfolio made of 50/50 Equity/Fixed Income ETF using the same performance metrics. I will be using VTI and AGG. VTI represents the equity market index and AGG represents the fixed income (bonds) market index.

Evaluation Metrics

I propose the metrics Absorption Ratio Delta, sharpe ratio, annualized return, annualized volatility.

Absorption ratio predicts the systemic risk in the markets or particular investment portfolios by measuring the concentration of risk. It shows when markets are "fragile" and vulnerable to loss and when markets are resilient.

Using the ratio, we can anticipate shifts in the portfolio volatility and exposure to loss or opportunity for gain.

- Absorption ratio delta: $\Delta AR = (AR15\ day - AR1\ year)/\sigma$
    - AR15 day which is 15-day simple moving average of AR,
    - AR1 year which is one-year simple moving average of AR
    - σ which is the standard deviation of one-year AR.

The AR Delta trading strategy forms a portfolio of EQ and FI, following these simple rules:

- $-1\sigma<AR<+1\sigma-1\sigma<AR<+1\sigma$ 50 / 50 weights for EQ / FI
- $AR>+1\sigma AR>+1\sigma$ 0 / 100 weights for EQ / FI
- $AR<-1\sigma AR<-1\sigma$ 100 / 0 weights for EQ / FI

Next, I compute that portfolio weights and run the re-balancing strategy using time series of returns and compute the following financial metrics

- sharpe of the strategy
- strategy annualized return
- strategy annualized volatility

Then I compare the results with 50 / 50 Equity / Fixed Income ETF strategy performance using the same performance metrics. I use VTI as Equity and AGG as Fixed Income assets as benchmark.

Another important metric/parameter is the PCA variance explained threshold, which I set to 80%. In other words, the principal components I will select will have to explain 80 of variability of the original data.

Definitions:

- Sharpe ratio is defined as annualized return divided by annualized volatility.

- Annualized return is the average amount of money earned by an investment each year over a given time period. It is calculated as an average to show what an investor would earn over a period of time if the annual return was compounded.

- Annualizing volatility is volatility (standard deviation) volatility in annualized terms. We need to multiply our daily standard deviation by the square root of 252. This assumes there are 252 trading days in a given year.

Project Design

A summary of the project workflow would consists of these steps

- Data wrangling

- Use PCA to construct eigen-portfolios
- Use PCA to calculate a measure of market systemic risk
- Implement and analyze performance of portfolio strategy

An overview of execution steps:

- Load data
- EDA
- Calculate daily stock prices returns using daily log-returns. This is preferred over simple daily return because the log-returns are normally distributed, an assumption of modern portfolio theory.
- Normalize asset return data
- Compute Absorption Ratio (AR)
    - Implement exponential smoothing weighting for the time series of stock returns
    - Define moving look back window over which I calculate returns for computing PCA
    - Define the step size of the window to move though the historical data
    - Define the speed of exponential decay H
    - Calculate the sequence of exponentially decaying weights wj is defined as $\boxed{w_j = \frac{X_j}{\sum_{j=0}^{N} X_j}}$
    - Define sequence of Xj where j⊂[N,0], an integer taking all values in the interval from 0 to N. $X_j = e^{-\frac{log(2)}{H} \times j}$
- Compute PCA and absorption at each step using moving window of returns for both exponentially weighted returns and equally weighted return (for benchmarking)
    - Fit PCA to covariance matrix of assets returns
    - Pass the fitted pca model values to absorption_ratio()

    - Compute number of principal components it takes to explain at least the percentage threshold of variance. The result of this calculation goes into pca_components
- Implement the AR Delta Trading Strategy
    - Implement the benchmark rules in a function
    - Generate the portfolio weights
    - Run the re-balancing strategy using time series of returns and compute sharpe of the strategy, strategy annualized return, and strategy annualized volatility
    - Compare with 50 / 50 Equity / Fixed Income ETF strategy performance using the same performance metrics.
- Interpret the Strategy Performance Report
    - A set of trading rules can be applied to historical data to determine how it would have performed during the specified period. This is called backtesting and it is a valuable tool for testing a trading system before putting it in the market.
    - Backtesting lets me to simulate a trading strategy using historical data to generate results and analyze risk and profitability before risking any my actual capital. (https://www.investopedia.com/articles/fundamental-analysis/10/strategy-performance-reports.asp#ixzz5QiVcbn1K)