Lawrence Stark CS 687 – Term Project Summary February 24th 2014

Problem:

Given a set of US Stock Market prices, Z(t) for t = 0 to n, predict the value of Z(t+1).

Data Source:

Tradestation: http://www.tradestation.com/
Stocks: S&P 500
Metrics: Open, Close, High, Low and Volume
Timeframe: January 1993, October 2013 (~20 years)
Total samples: ~3,650,000 (500 stocks * 365 days * 20 years)

Methods:

Statistical Learning: Bayesian Inference

1. Given: $X(t)=\frac{Z(t)-Z(t-1)}{Z(t-1)}*100$ where t spans from January 1st to December 31st of a given year. Compute the distribution from the observations for that year. At the end, there will be up to 20 different distributions.
2. Given the sample Y(t), use The Bayesian Classifier to classify which distribution (from step 1) most likely contains Y(t).
$$\frac{P(w_1)P(X|w_1)}{P(w_2)P(X|w_2)} < (>)\ 1$$
3. From sample Y(t) and observations X(t), compute the expected value for Y(t+1):
$\rho=\frac{E\{(X_t-\mu_x)(Y_{(t+1)}-\mu_y)\}}{\sigma_x \sigma_y}$ with ρ = 0.3 [or the correct value for the classified distribution].

Consensus: RANSAC

1. Given observations X(t) for t = 0 to n, select a random subset to fit a model.
2. Repeat step 1, k times until the best model is chosen. k can be computed as:
$$k=\frac{\log(1-p)}{\log(1-w^n)}$$
3. Using the consensus set, estimate the value for X(t+1).

Context-Dependent Learning: Hidden Markov Models (discrete observation)

1. Use Vector Quantization to categorize each observation into one of L possible distinct values in an *l*-dimensional space.
2. Use Baum-Welch Reestimation to compute the parameters of model S. Thus p(X|S) is the maximum likelihood estimator:
$$\Im_k(i,j) \equiv \Im_k(i,j,X|S) \equiv \frac{\Im_k(i,j,X|S)}{P(X|S)}$$
3. Using Expectation Maximization, compute X(t+1) such that the M (*maximization)* step is:
$$\Theta(t+1): \frac{\partial Q(\Theta;\Theta(t))}{\partial \Theta}$$

If time permits: attempt using the above algorithms in combination as an ensemble classifier to achieve even greater performance.

Validation:

The daily values extracted from Tradestation also include a 501st stock known as the SPDR (Spider). This ETF mimics the entire S&P. Thus the error can be computed as: $\sqrt{(Z(t+1)-SPDR(t+1))^2}$. Furthermore, with 20 years of data, the set can easily be partitioned for K folds cross validation where each fold is between 4 and 10 years for K = 2 to 5.