Predicting the Stock Market using Artificial Intelligence

Lawrence Stark

CS 687

Spring 2014

Topic

- Using historical data (3 days), predict whether tomorrow's stock market will close UP or DOWN
- Automated prediction based on model developed from individual stock market data.

Utility

- Get Rich the Quick and Easy Way!
-
- Personal Finance
  - e.g. Self-managed 401k
-
- Complex Signal Analysis (Data Mining):
  - Find patterns given unknown distribution
  - Predict future behavior for irrational agents

Method

- Candlestick Pattern
  - Munehisa Homma: Japanese Rich Trader from 1700's
  - Steve Nison: Applied Homma's candlesticks to contemporary investment (stocks)
- Model Market Behavior
  - Use 500 stocks to learn individual stock movement
  - Use model to predict market value for next day

Background

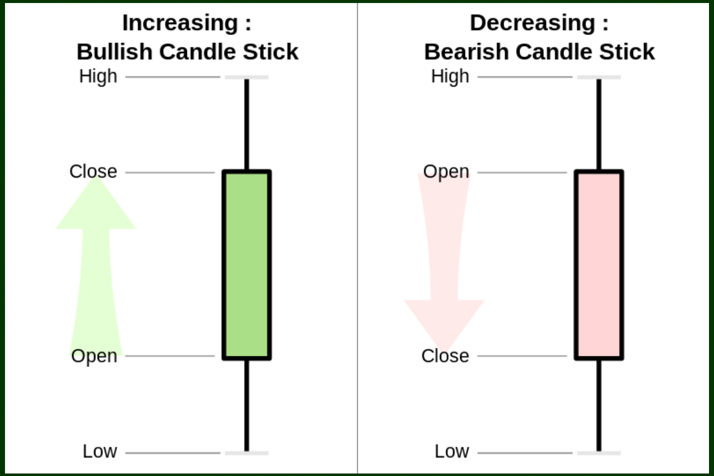
- JPM: Days of loss in 2013 = 0
- Virtu: Days of loss 2009-2013 = 1
- Support Vector Machines
- Neural Networks
- Twitter
- Autoregressive Integrated Moving Average (ARIMA)
- Echostate Networks

Data Source

- Tradestation: www.tradestation.com
- Stocks: S&P 500 + SPDR
- Timeframe:
  - January 1993 to October 2013
- 3 Day Sliding Window
  - Use day 4 for label
  - Train/Test : approximately 2.2 million samples
  - Validate: approximately 5,200 samples

Data

- Features:
  - Open, High, Low, Close
  - For each of Day 1 to 3
  - Delta Close Day1/2 and Day 2/3
  - Label: related to line slope: Up, Down, Peak, Trough



Example:
10.97,11.05,10.82,10.97
11.01,11.05,10.56,10.67
10.60,10.67,10.57,10.60
-0.30,-0.07,DOWN

Feature Extraction

- So Far: 3 Day candlestick patterns
  - Only 15 attributes
  - Manually reduced from 24
  - PCA suggests only 3: ΔC12, ΔC23, D_3Vol
- Possible Future:
  - 100 Day candlestick pattern
  - More than 500 attributes
  - PCA critical for dimensionality reduction

AI Methods

- Baseline: random buy and sell
- Bayesian Inference
  - Likelihood Ratio Test: $\Lambda(x) = \frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(x|\omega_2)}{P(x|\omega_1)}$
- Hidden Markov Models
  - Vector Quantization
  - P(X|S): $\Im_k(i,j) \equiv \Im_k(i,j,X|S) \equiv \frac{\Im_k(i,j,X|S)}{P(X|S)}$
  - EM: $\Theta(t+1): \frac{\partial Q(\Theta;\Theta(t))}{\partial \Theta}$
- RANSAC
  - Random subsets best model
  - K computed as: $k = \frac{\log(1-p)}{\log(1-w^n)}$

Software Platforms

- WEKA, Matlab and Java (preprocessing only)
- Naive Bayes (WEKA)
  - Implementation: John & Langley - Estimating Continuous Distributions in Bayesian Classifiers
- HMMWeka – Plugin for Hidden Markov Models
  - Implementation: Bishop - Pattern Recognition and Machine Learning
- RANSAC (Matlab)
  - Implementation: Fishler & Boles - Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography

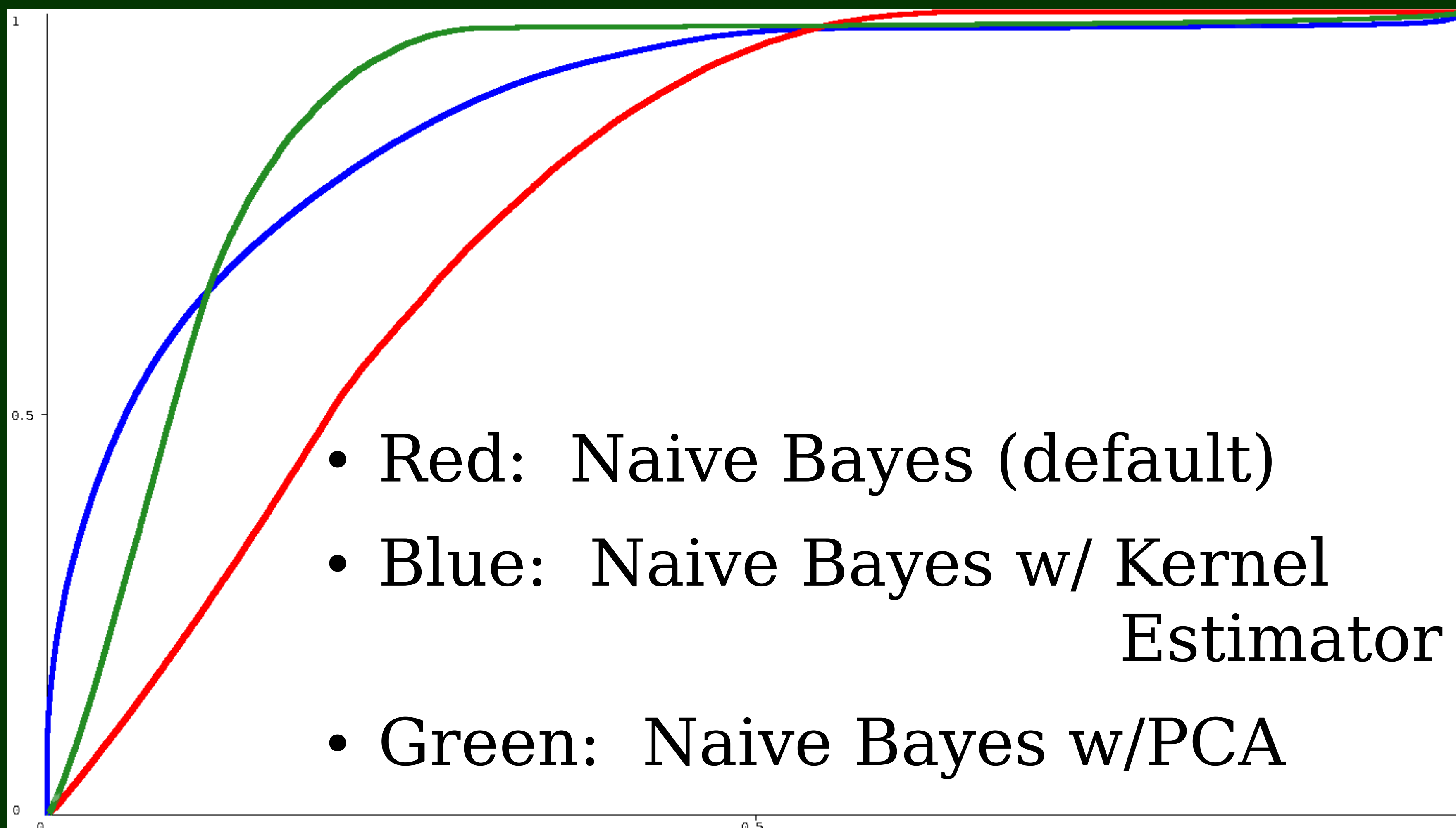
Performance Evaluation

- SPDR (spider)
  - Mimics entire S&P 500
  - Standard for performance evaluation
-
- Error: $$\sqrt{\left(Z\left(t+1\right)-SPDR\left(t+1\right)\right)^{2}}$$
- Metrics:
  - Accuracy: predicted market status vs. SPDR
  - ROI: the amount of money gained from trades
  - Market Days: days money is used for trading

Cross Validation

- Training Set
  - 50% of S&P 500 (1.1 million)
- Test Set
  - Remaining 50% of S&P 500 (1.1 million)
- Validation Set
  - 100% of SPDR (5235)
- Validation set deliberately not mixed with train/test sets to mimic real world.

Data Visualization



- Red: Naive Bayes (default)
- Blue: Naive Bayes w/ Kernel Estimator
- Green: Naive Bayes w/PCA

Preliminary Results

Trial	Accuracy	Market Days	ROI
Random	51%	2618	-31.69%
Naive Bayes Standard	44%	2396	100.32%
Naive Bayes Kernel Est.	52.47%	1201	185.98%
Naive Bayes3 w/ PCA	55.16%	1201	268.46%

Conclusion

- This is hard! Hence poor results so far
- 3 Day candlestick is standard but may not provide enough attributes
- Too much data. Looking forward to RANSAC for data reduction.
- May further employ K means and then use clusters for test and training.
- Influence of outside phenomenon like inflation