Prediction of Stock Price Movement from Options Data

Charmaine Chia  (cchia@stanford.edu)

Background

An option is a contract that gives the buyer the right to buy or sell an underlying stock at an agreed upon strike price, during a certain period of time.

Types of options contracts

- CALL: Right to buy
- PUT: Right to sell

Underlying stock at price $S_t$

Call holders want the stock price to go above the strike price.

Put holders want the stock price to go below the strike price.

Q: Can we predict how the underlying stock price will move from time series data on the options market?

Examples of options variables:

- Volume of puts & calls traded
- Put-call parity (PCP) deviance
- Implied volatility (IV)
- Perceived future volatility of stock
- Used to calculate options price
- IV spread & skew
- Exponential Moving Average (EMA)
- Gaussian moving filter
- Simple Moving Average (SMA)
- Cross region of overlap of data points labelled +1 and -1
- Put volume
- Call volume
- Volume of puts & calls traded
- Put-call parity
- Raw returns
- Smoothed returns
- Returns (30-day SMA)
- Put volume (Gaussian averaged)

The Data

- Time series data of 57 healthcare companies from 1/3/2007 to 12/4/2014
  - Data from successive time points is not i.i.d
  - Not necessarily normally distributed
- From the stock price data, we can obtain daily returns:
  \[ \text{Return} = \frac{\text{Close price today} - \text{Close price yesterday}}{\text{Close price yesterday}} \]
  - From the options data, we get 39 options market-related variables, often highly correlated with each other.
  - Time series data can be smoothed by applying either
    - Simple Moving Average (SMA)
    - Exponential Moving Average (EMA)
    - Gaussian moving filter

Initial Approach

Q: Are there any obvious correlations between the returns data and individual options variables?

- Scatter plots of (Options variable(t), Returns(t))
- Cross-correlation of time series of
  - Raw Returns vs Options variable
  - Smoothed Returns vs Options variable
  - Returns rolling variance vs Options variable

Decision stumps boosting

- Decision stump: \( \phi_j(x) = \text{sign}(x_j - s) = \begin{cases} 1 & \text{if } x_j \geq s \\ -1 & \text{if otherwise} \end{cases} \)
- \( x_j \) consist of past returns, PCR and options variables
- 10-fold CV: training data drawn from S2 companies, testing data from remaining 5 companies
- Effect of omitting certain variables tested

Classification problem

- Can we even predict the direction of stock movement (i.e. + or - return)?
- Does a linear model accurately capture the relationship between the signal and effect?
- What variables contain should we use?

Try:

- Put-to-call ratio \( \rightarrow \) indication of trader sentiment about market direction
- 26-day EMA of Returns and PCR \( \rightarrow \) capture trends with less random jumps
- Second plot shows the PCR data processed to get daily fractional change in PCR.
- Possible (negative) correlation between position of thresholded spikes in PCR and returns?

Summary of average training error for different combinations of X variables used in decision stumps

- Raw returns are too noisy to make predictions on / with
- EMA return [-1] is the most predictive variable
- PCR, IV and PCPdev contain some signal
- Raw return can be extracted from EMA return prediction

Other classification schemes

- One problem with decision stumps boosting is that the way hypotheses are picked is somewhat arbitrary and difficult to intuitively explain
- Nonetheless, it informed us on which variables were more significant. We can plot the data along the top 3 variables from each class \( A, B, C \):

Future work

- Model to predict the magnitude of returns, given prediction of whether it will be + or -
  - MSE of 7.18e+06 on full dataset
  - MSE of elastic net regression ranged from 6.66e-06 to 1.82e-05 (as the lambda is decreased)
- Also try: Regression trees, SVM regression