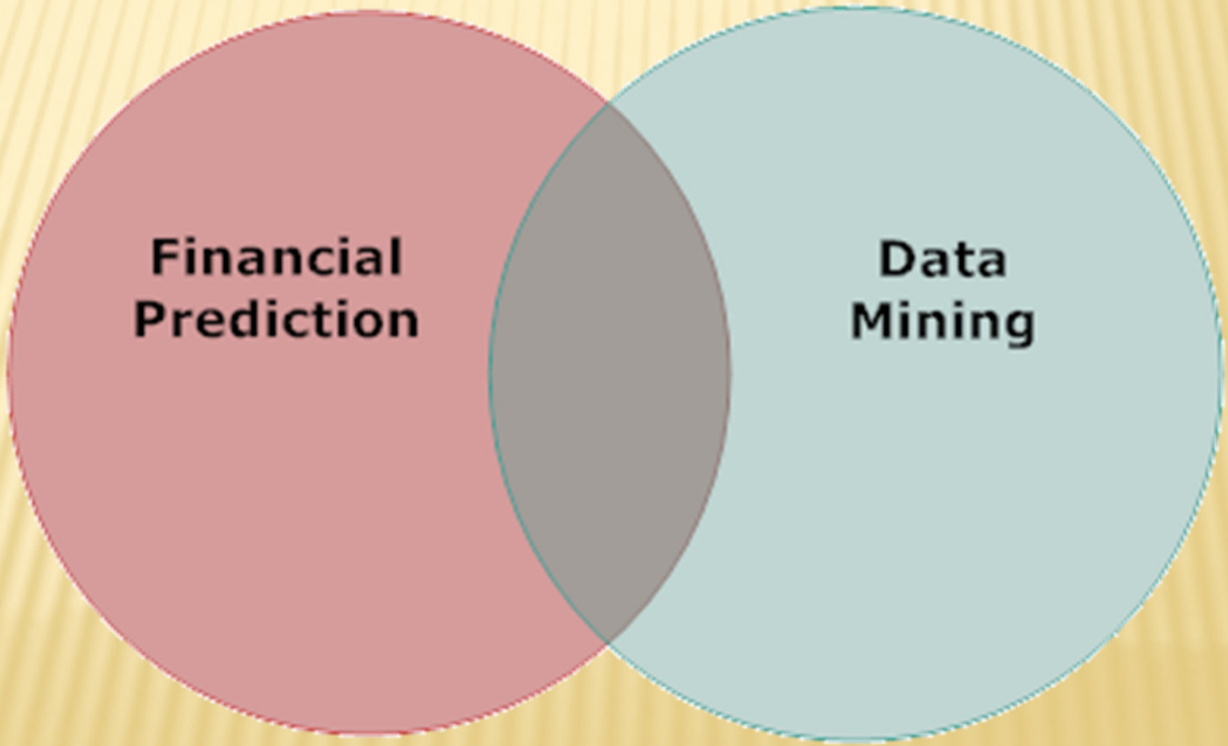
Price movement prediction in Hong Kong equity market

FINANCIAL DATA MINING III

INTRODUCTION

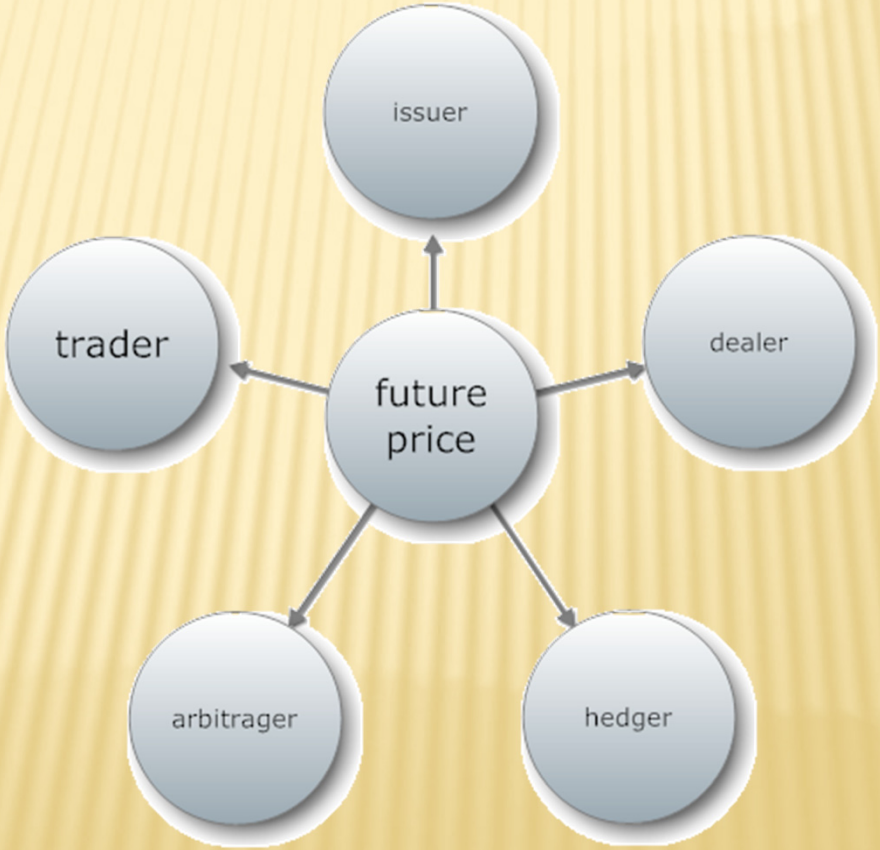
PROJECT SCOPE

- Price movement prediction in HK equity market using data mining techniques



PROJECT SCOPE

- Different roles in finance industry rely on future price information



PROJECT SCOPE

- Target user: Traders
- Task: Forecast price movement

- given any time *t*
- given all available data at *t*
- forecast price movement
  for time > *t*
- for some tradable assets in HK equity market

PROJECT SCOPE

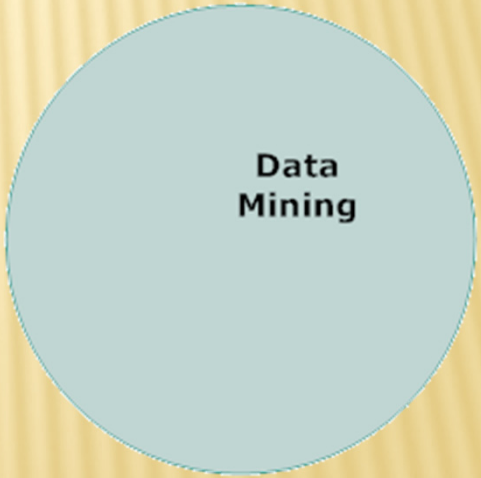
- Technique: Supervised learning

Training Set

Hairstyle	Wear
short	shirt
long	shirt
long	blouse

Class Label

Sex
boy
boy
girl



Input Value

short	blouse

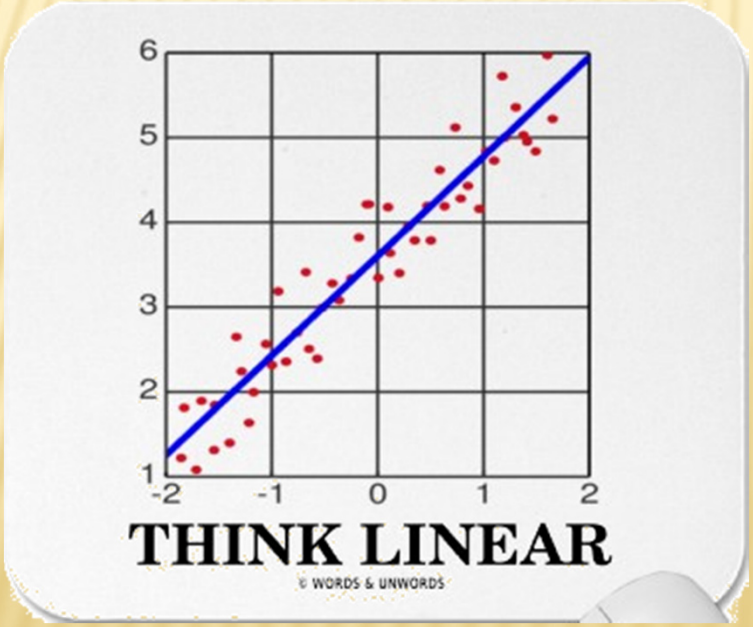
=>

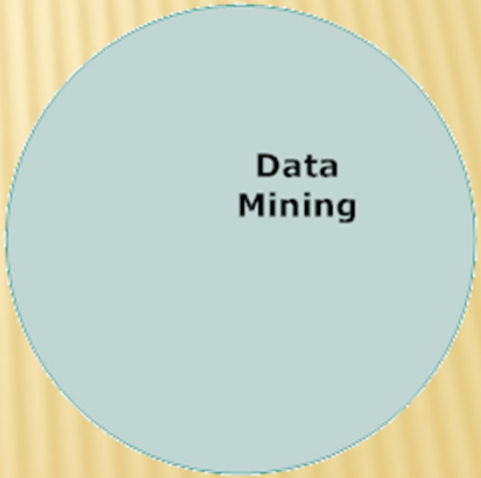
Predicted Label

?

PROJECT SCOPE

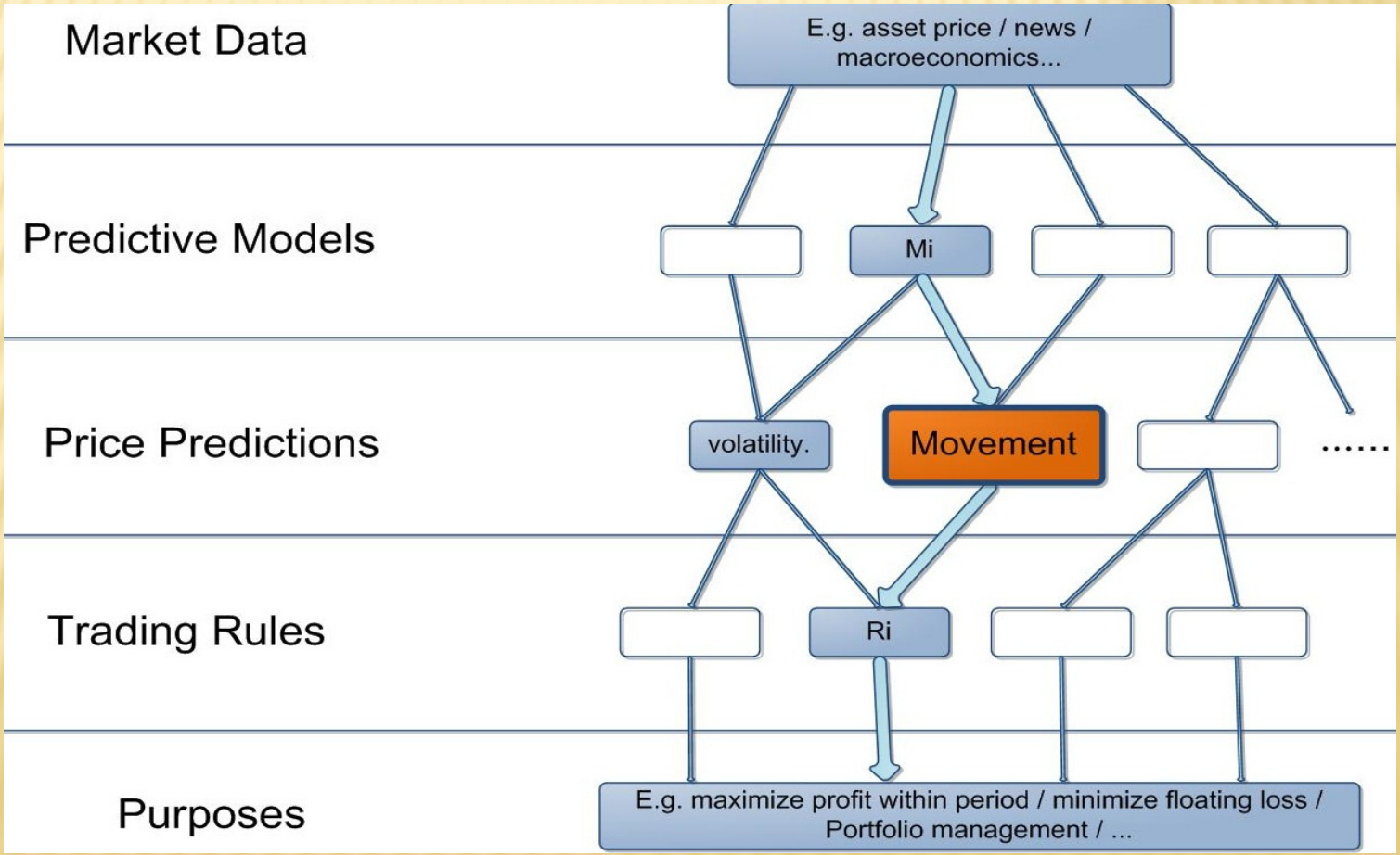
- Technique: Supervised learning
  - E.g. simple linear regression
  - Output model: $y = \alpha + \beta x$





PROBLEM FORMULATION

- Prediction problem framework:



PROBLEM FORMULATION

- "price movement"

  = magnitude + direction + timeframe

PROBLEM FORMULATION – DEFINITION

- Problem definition:

  - given any time *t*
  - given some attribute vectors at *t*
  - plus their associated observed labels
  - learn a predictive model
  - to forecast price movement
    for time > *t*
  - for some tradable assets
    in HK equity market

Attribute vectors

Trend	RSI(14)
x_1	y_1
x_2	y_2
x_3	y_3

Observed labels

Future Price movement
o_1
o_2
o_3

Attribute vector

x_t	y_t

=>

Predicted label

o_t

MOTIVATION

- Crystal ball => $
- To find out how hard this problem actually is…

FINANCIAL BACKGROUND

- Price Predictability
- Fundamental Analysis
- Technical Analysis
- Ensemble Approach

BACKGROUND – PRICE PREDICTABILITY

- Is financial asset price predictable?

- Lots of theoretical arguments…
- Lots of empirical studies…
- Lots of market wisdom…

BACKGROUND – PRICE PREDICTABILITY

- Is financial asset price predictable?

- Lots of theoretical arguments…
- Lots of empirical studies…
- Lots of market wisdom…

Just try it !

BACKGROUND – TRADITIONAL APPROACH

- We are not the first ones to try…

BACKGROUND – TRADITIONAL APPROACH

- Fundamental Analysis
  - Valuation of intrinsic value of a stock
    - - financial performance
    - - corporate management
    - - business environment
    - - …
  - Drawback:
    - Short-term price movement unexplained

BACKGROUND – TRADITIONAL APPROACH

- Technical Analysis
  - Calculation / charting on stock historical price, volume, O.I. information

BACKGROUND – TRADITIONAL APPROACH

- Technical Analysis
  - Calculation / charting on stock historical price, volume, O.I. information
  - For example, definition of RSI(n):

U = *closing price*$_{now}$ – *closing price*$_{previous}$ OR *0*
D = *0* OR *closing price*$_{previous}$ – *closing price*$_{now}$

$$RS = \frac{\mathrm{EMA}(U, n)}{\mathrm{EMA}(D, n)}$$

$$RSI = 100 - \frac{100}{1 + RS}$$

BACKGROUND – TRADITIONAL APPROACH

- Technical Analysis
  - Calculation / charting on stock historical price, volume, O.I. information
  - For example, charting of RSI(14):

BACKGROUND – TRADITIONAL APPROACH

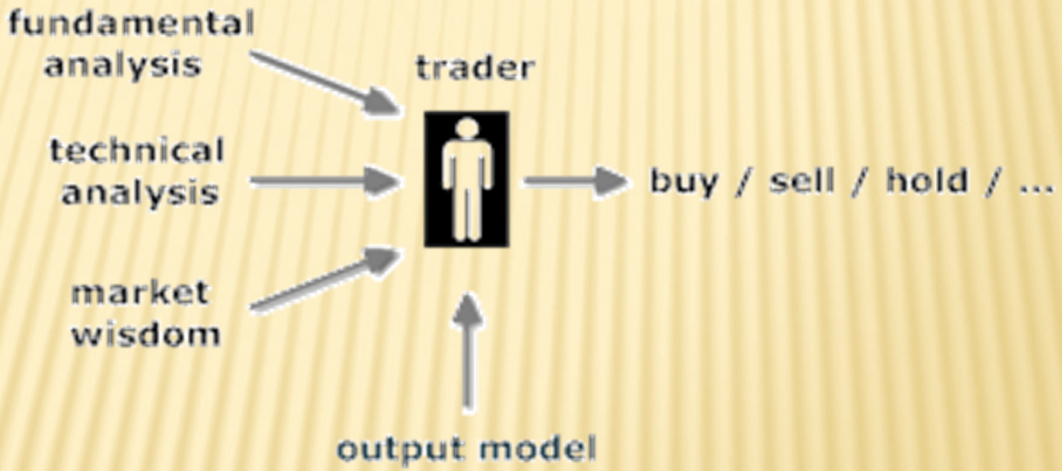
- Technical Analysis
  - Calculation / charting on stock historical price, volume, O.I. information
  - Drawback:
    - Decision making is subjective

BACKGROUND – MIXED APPROACH

- In practice…

BACKGROUND – MIXED APPROACH

- Introducing financial data mining



FINANCIAL DATA MINING – ADVANTAGE

- Huge data set
- Multi-dimensional

FINANCIAL DATA MINING – METHODOLOGY

- Hypothesis testing

Scientific Method	Data Mining
1. Hypothesis formation	Feature extraction and selection
2. Theory formulation	Model selection
3. Experiment or observation	Model evaluation and comparison
4. Technology	Using discovered knowledge

PROJECT OBJECTIVE

- In 1st semester, our group seeks to
  - 1. understand the working domain
  - 2. develop a prototype data mining processs and get preliminary results with Hang Seng Index (HSI)

DATA MINING PROCESS

DATA COLLECTION

DATA PRE-PROCESSING

DATA PRE-PROCESSING

- Data cleansing

DATA PRE-PROCESSING

- Outlier treatment

DATA PRE-PROCESSING

- Data integration

K-NEAREST NEIGHBOR MODEL

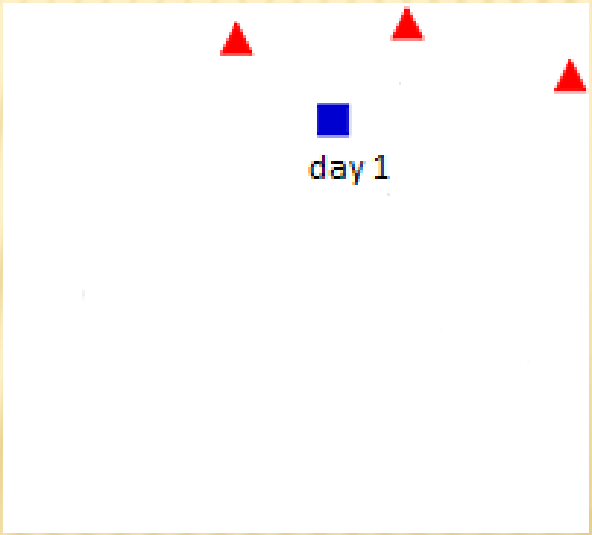
- Theory
- Input and Output Selection
- Experiment

K-NEAREST NEIGHBOURS MODEL

- Memory-based
- Applied with walk forward testing method, memory changes through time
- ‘Memory’ is defined to be the set of all vectors in the feature space

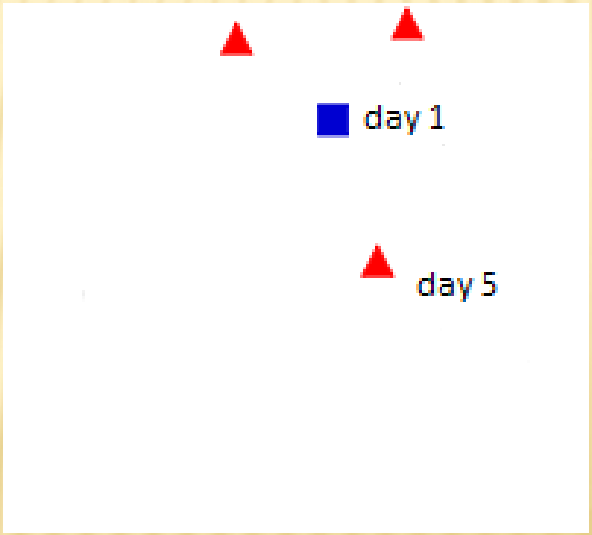
K-NEAREST NEIGHBOURS MODEL

- Changing memory
  - Consider test point at day 5,



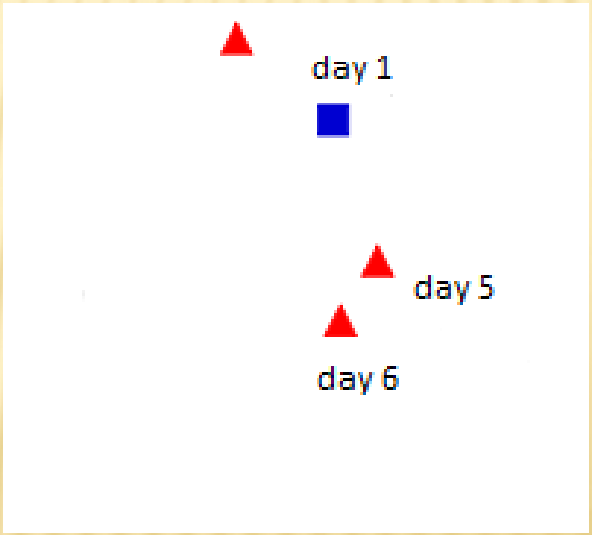
K-NEAREST NEIGHBOURS MODEL

- Changing memory
  - Consider test point at day 6,



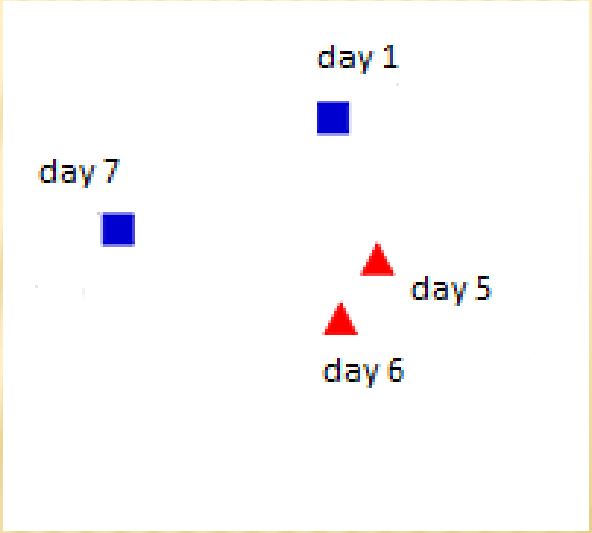
K-NEAREST NEIGHBOURS MODEL

- Changing memory
  - Consider test point at day 7,



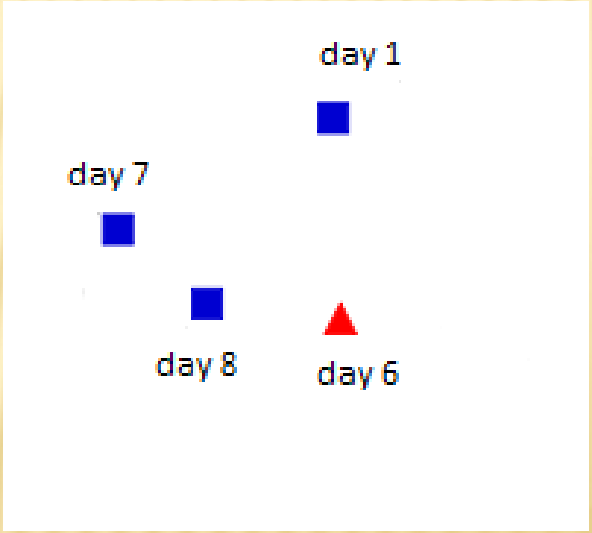
K-NEAREST NEIGHBOURS MODEL

- Changing memory
  - Consider test point at day 8,



K-NEAREST NEIGHBOURS MODEL

- Changing memory
  - Consider test point at day 9,



K-NEAREST NEIGHBOURS MODEL

- Memory = Model
- Memory changes = Model changes
  - Adaptive learning of dynamic market
- Avoid target function of the form

$$x(t+1) = F(Y_1(t), Y_1(t-1), \ldots, Y_2(t), Y_2(t-1), \ldots)$$

OUTPUT SELECTION

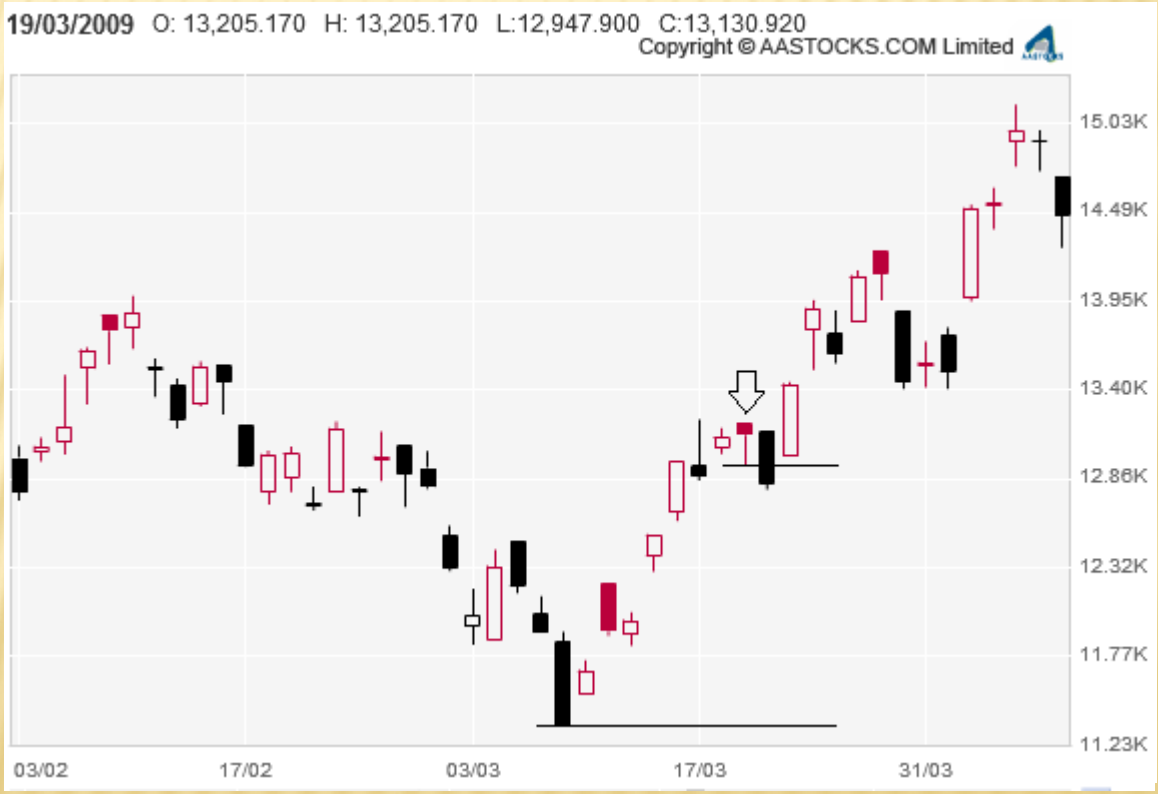
- "Future N-day bottom":
  - *For each day t, if daily closing price is lower than the minimum daily low price within N trading days in the future, then it is a (future) N-day bottom*

OUTPUT SELECTION

- "Future N-day bottom":
  - E.g. HSI Future 8-day bottom, w.r.t. day 9th MAR 09

OUTPUT SELECTION

- "Future N-day bottom":
  - E.g. HSI Future 8-day bottom, w.r.t. day 9th MAR 09

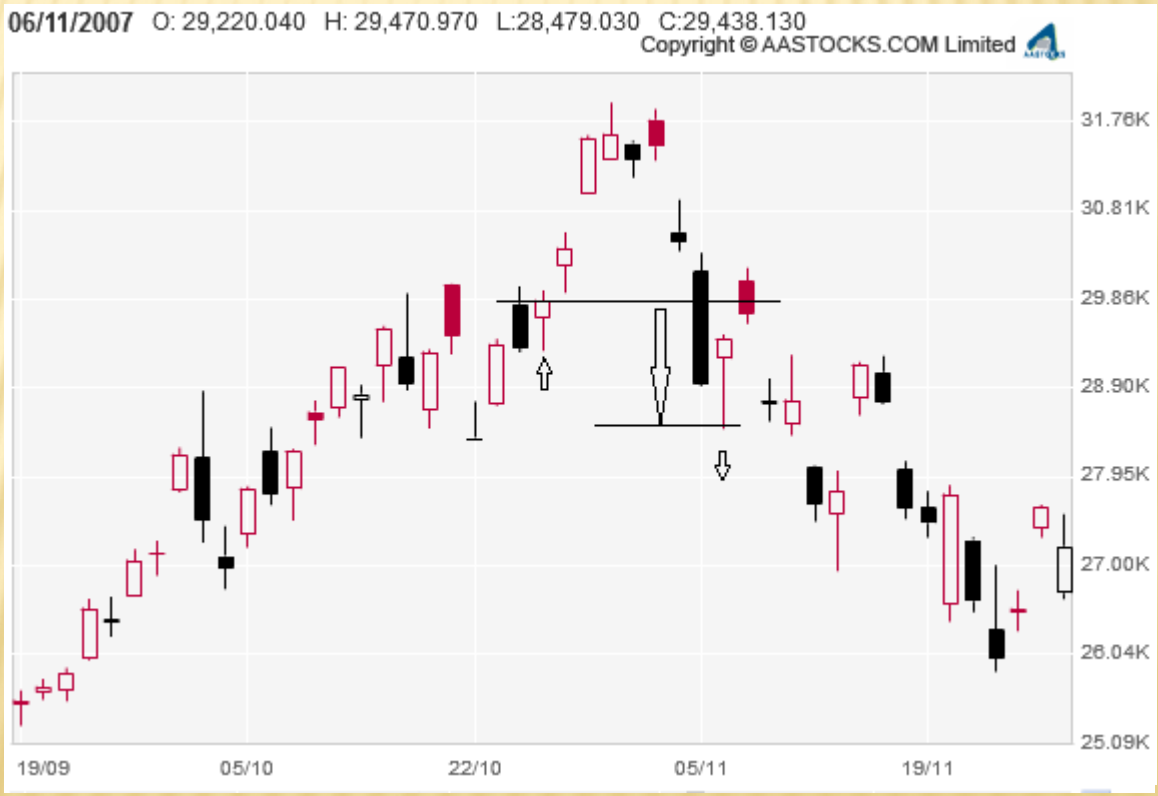


OUTPUT SELECTION

- "Future N-day bottom":
  - E.g. HSI Future 8-day bottom, w.r.t. day 25thOCT 07

OUTPUT SELECTION

- "Future N-day bottom":
  - E.g. HSI Future 8-day bottom, w.r.t. day 25thOCT 07

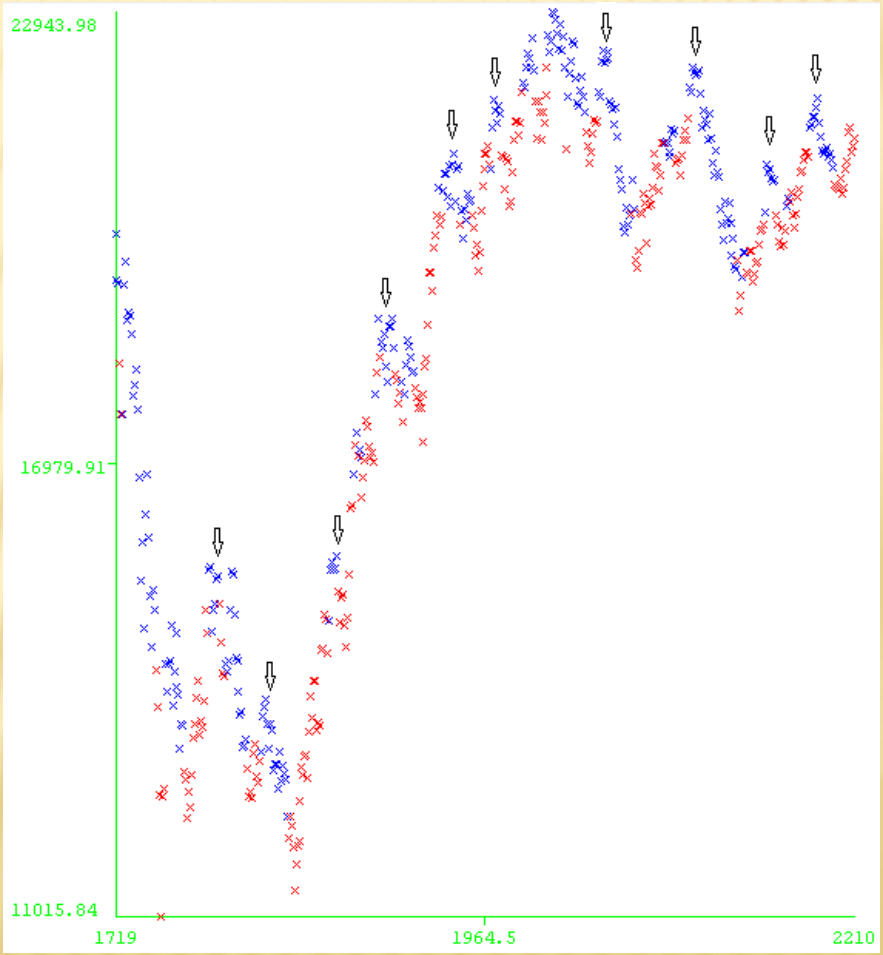


OUTPUT SELECTION

- "Future N-day bottom"
  - Describes direction and timeframe of the future price movement
  - Does not embed information about magnitude

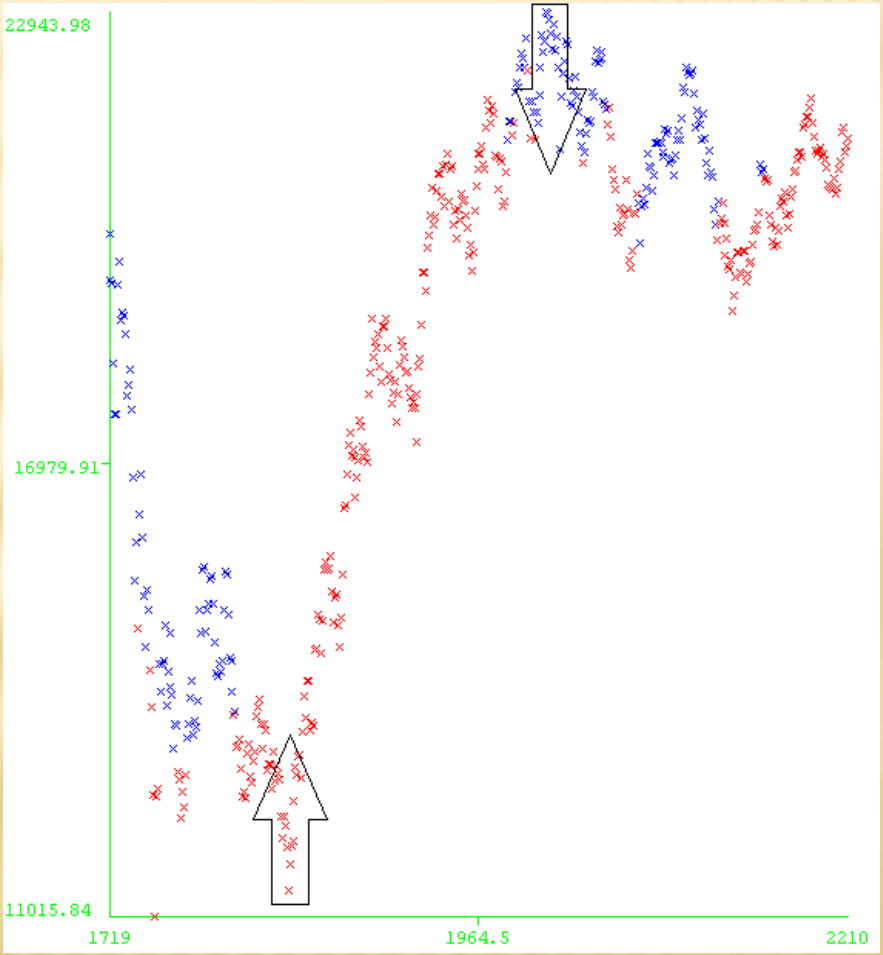
OUTPUT SELECTION

- HSI “future 8-day bottom”, red are positives



OUTPUT SELECTION

- HSI “future 50-day bottom”, red are positives



OUTPUT SELECTION

- Rationale of defining “future N-day bottom” as class label
  - Adjustable timeframe of prediction
  - Tuning N to fit the predicted financial asset

OUTPUT SELECTION

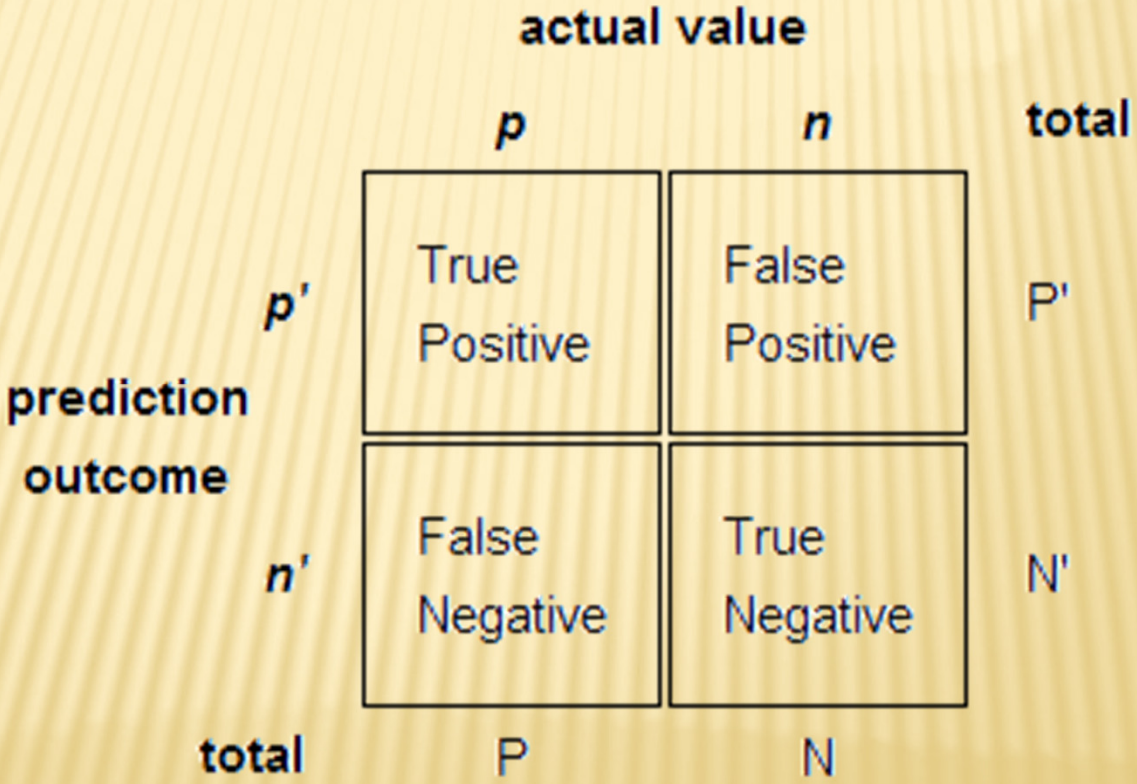
- Rationale of defining “future N-day bottom” as class label
  - In actual implementation, we allow users to set a ‘tolerance margin’ (in percentage of price)
    - If(curPrice[t] <= (futurePrice[t+timeframe] + curPrice[t] * stoplossmargin))
      bottoms[t] = TRUE;

OUTPUT SELECTION

- Rationale of defining “future N-day bottom” as class label
  - Definition is easily extended to “past N-day bottom”.
  - “N-day minimum”
    = “past N-day bottom” AND “future N-day bottom”
  - Definition could be mirrored to define “N-day maximum” as class label

EXPERIMENT – EVALUTATION METRICS

- Confusion matrix



EXPERIMENT – RESULTS

- HSI future N-day bottom against RSI(14)
- Walk forward testing method
- Look back 540 days (1.5 years)

EXPERIMENT – RESULTS

Class label	RSI(14)
Future 4 day bottom	$\begin{bmatrix} 577 & 451 \\ 372 & 316 \end{bmatrix}$
Future 8 day bottom	$\begin{bmatrix} 438 & 404 \\ 405 & 465 \end{bmatrix}$
Future 12 day bottom	$\begin{bmatrix} 451 & 431 \\ 379 & 447 \end{bmatrix}$
Future 16 day bottom	$\begin{bmatrix} 355 & 420 \\ 413 & 516 \end{bmatrix}$
Future 20 day bottom	$\begin{bmatrix} 314 & 379 \\ 423 & 564 \end{bmatrix}$

CONCLUSION

REFERENCES

REFERENCES

- Comma-separated values. (2010, November 25). In Wikipedia, The Free Encyclopedia. Retrieved 13:31, November 30, 2010, from http://en.wikipedia.org/w/index.php?title=Comma-separated_values&oldid=398858843.
- undefined. (September 1, 2010). Triennial Central Bank Survey of Foreign Exchange and OTC Derivatives Market Activity in April 2010 - Preliminary global results - Turnover. In Bank for International Settlements. Retrieved undefined, from http://www.bis.org/publ/rpfx10.pdf.
- Sullivan, Ryan & Timmermann, Allan & White, Halbert, 2001."Dangers of data mining: The case of calendar effects in stock returns," Journal of Econometrics, Elsevier, vol. 105(1), pages 249-286, November.
- Schwert, G. William, Anomalies and Market Efficiency (October 2002). Simon School of Business Working Paper No. FR 02-13. Available at SSRN: http://ssrn.com/abstract=338080 or doi:10.2139/ssrn.338080.

- Sullivan, Ryan & Timmermann, Allan & White, Halbert, 2001."Dangers of data mining: The case of calendar effects in stock returns," Journal of Econometrics, Elsevier, vol. 105(1), pages 249-286, November.

REFERENCES

- Pauly, R & Schell, A, 1989. "Calendar Effects in Structural Time Series Models with Trend and Season," Empirical Economics, Springer, vol. 14(3), pages 241-56.
- undefined. (undefined). Technical Analysis. In investopedia. Retrieved October, 2010, from http://www.investopedia.com/terms/t/technicalanalysis.asp.
- Vásquez, M.,González Osorio, F., & Hernández Losada, D.(2009). Diego Mining Candlesticks Patterns on Stock Series: A Fuzzy Logic Approach.
- Ting-Hao Huang, Zhang-Ye Li. (2008). Stock Prediction Using Web News Articles.
- Maimon, O. Z., & Rokach, L. (2005). Data mining and knowledge discovery handbook. New York: Springer Science+Business Media.
- undefined. (January 6, 2008). Thomson Financial NewsHKEx says late fluctuation in some stocks last Friday due to MSCI changes. In Forbes.com. Retrieved Noverber, 2008, from http://www.forbes.com/feeds/afx/2008/06/01/afx5067901.html.
- Leif E. Peterson (2009) K-nearest neighbor. Scholarpedia, 4(2):1883, revision #75484.

THANK YOU
THIS IS THE END OF THE PRESENTATION

Q & A