Task1: Prediction

- Two-stage Model: **Day model** and **Intraday model**
- Day model

Learn & predict on one sample each day

- Intra-day model

Learn & predict on one sample each minute

- A Keypoint

We **DO NOT** fit on **price Y_t** directly instead fit on its Price difference(差分) **$D_t = Y_{t+1} - Y_t$** ,which can significantly improve model robustness and decrease prediction error in practice.

e.g. if $Y_0 = 1$, our predictions are $D_0 = 0.1$, $D_1 = -0.1$, $D_2 = 0.2$, then predictions on Y are $Y_1 = 1.1$, $Y_2 = 1.0$, $Y_3 = 1.2$

Two-stage model

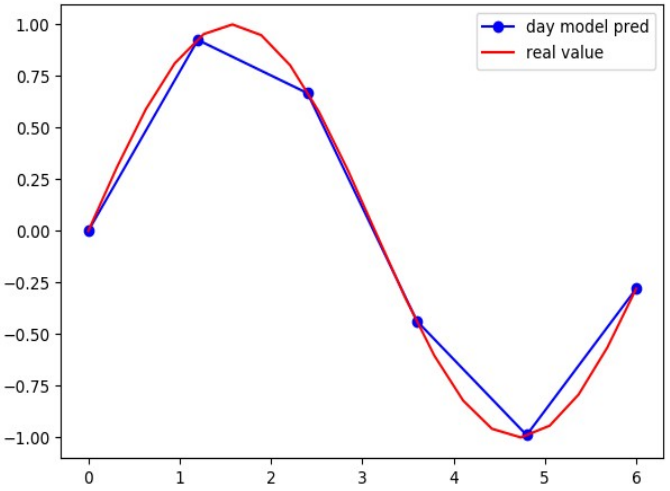
- Day model

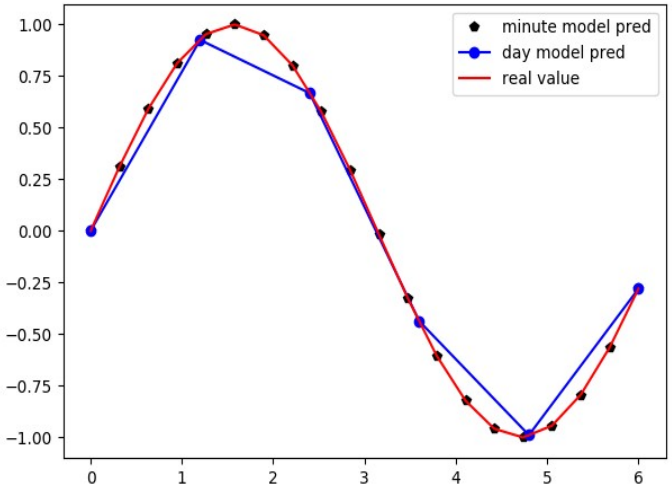
predict opening price each day.

- Minute model

predict minute price in day, filling the intra-day price predictions.

Like doing interpolation (插值)





Day Model: Seasonal ARIMA

- Name: Autoregressive integrated moving average (ARIMA)

An ARMA(p,q) process is given by

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

or equivalently, where the *L* is named lag operator.

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) = \left(1 - \sum_{i=1}^{p'-d} \phi_i L^i\right) (1 - L)^d.$$

An ARIMA(p,d,q) process is

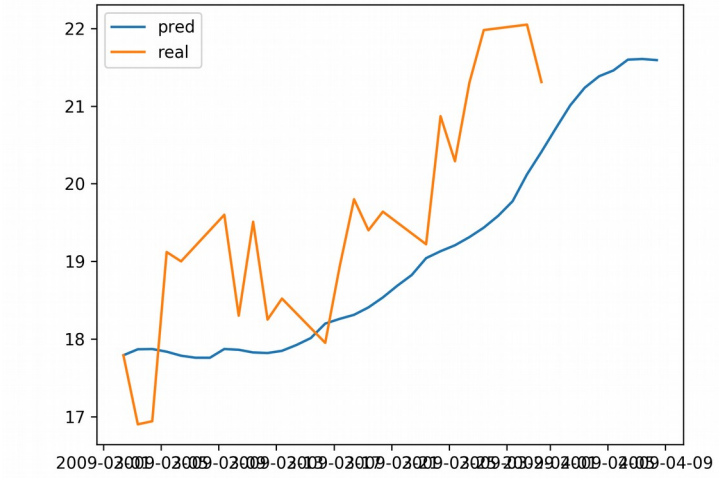
$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

- For more details please refer to ***Econometrics*** (计量经济学) !

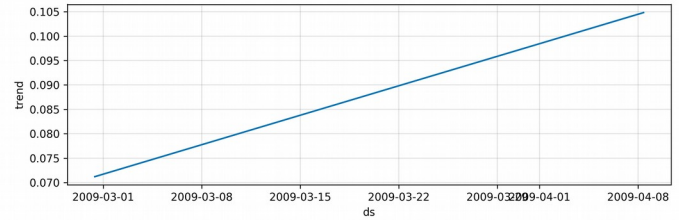
Day Model: Test Result

- Train: 2008/1/1 to 2009/3/1
- Test: 2009/3/2 to 2009/4/1
- Stock: 600000

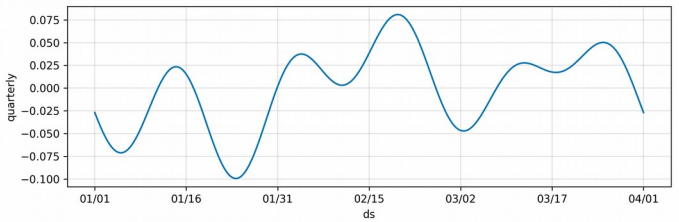
Test Result, MSE: 1.3



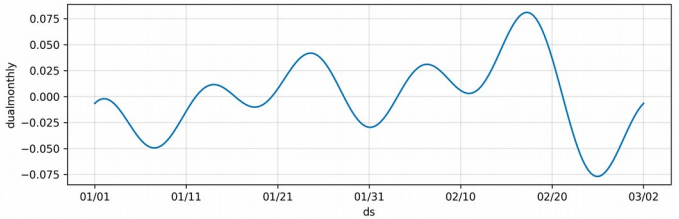
4 Seasonal component & Trend



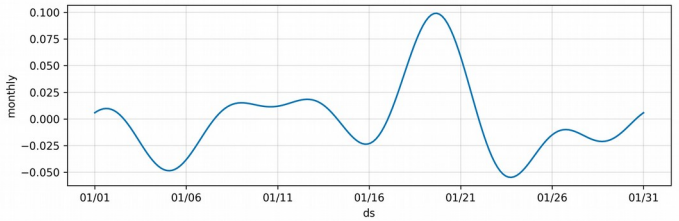
Trend



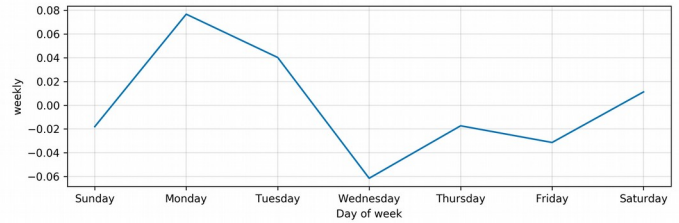
Quartely Seasonality (周期性)



Double Monthly



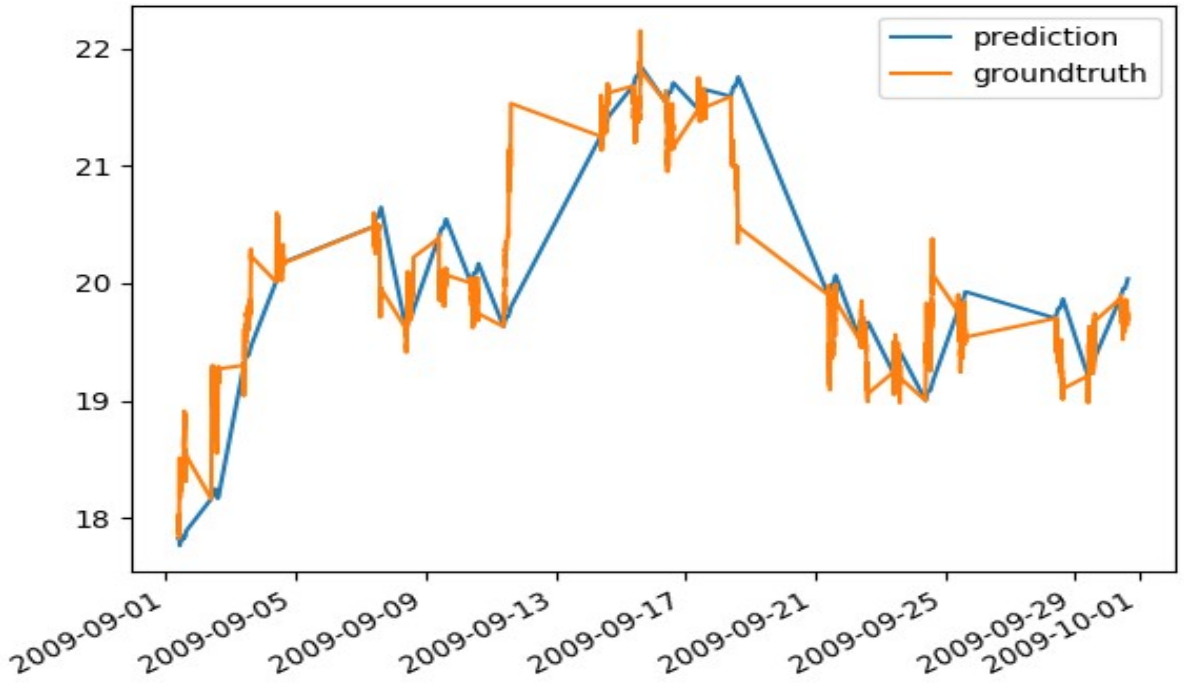
Monthly



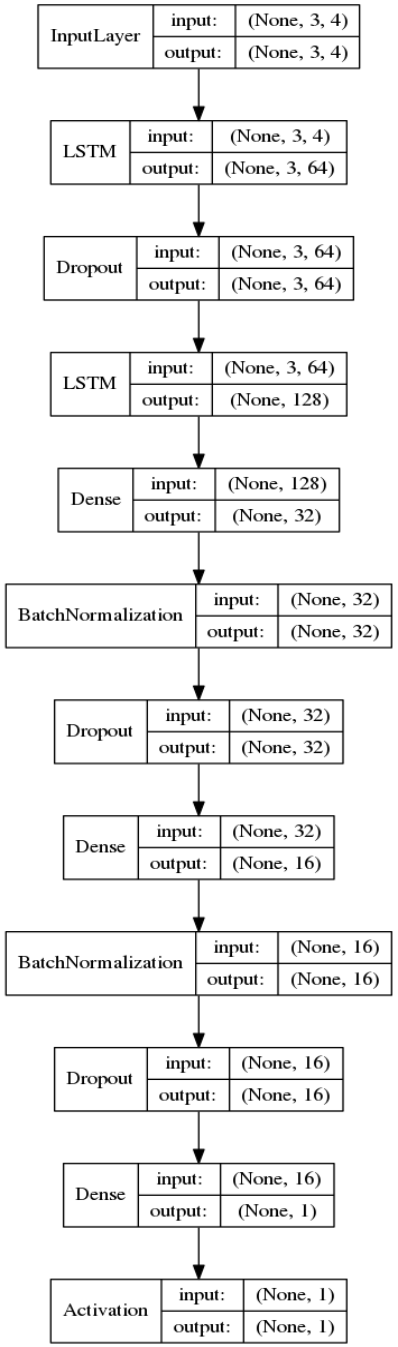
Weekly

Minute Model: LSTM

- Select 600000, train on 2009/6/1 to 2009/8/31, test on 2009/9/1 to 2009/9/30
- Test Result, MSE: 0.2, not bad, but reliant on precise day model prediction



NN Structrue



Task2: Trading Strategy

- Problem with the timeseries prediction model

The prediction highly depends on the last month modality i.e. if last month is downward, the prediction is most probably downward of flat. However, the future trend is uncertain, influenced by many factors, far more than historical trend!

- Possible solutions: **statistical arbitrage** (统计套利) strategy

One method: Mining the short-term stable relationship between stocks in the market thus building portfolio which can avoid market risk (**Risk Hedge**，风险对冲).

Task2: Trading Strategy

- Cointegration (**协整关系**)

The spread(差价) between two stocks (S1,S2) prices always fluctuates between a mean value, which means we can build a portfolio consists of S1 and S2 to earn profit.

- Transaction Rule

Step1. Let spread Y = S1 – S2, with a mean Y_m and a std Y_{std} , therefore we can build a up line $Y_{up} = Y_m + Y_{std}$ and a down line $Y_{down} = Y_m - Y_{std}$.

Step2. The rule is :

When $Y > Y_{up}$, long S2, short S1; When $Y < Y_{down}$, long S1, short S2.

Task2: Our Method

Step1:

- Find several S2 which have large correlation coefficient with S1

Step2:

- Finding cointegration between pairs S1 and S2
- doing linear regression S2 = w*S1 + const with respect to each pair (S1,S2)
- then doing stationary test on Sp = S2 -w*S1, select pairs (S1,S2) with significant test result, and build pair transaction rule

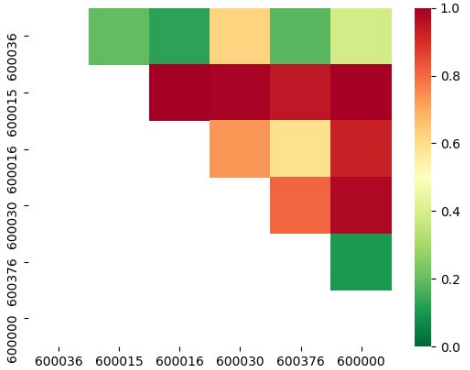
corr	s1	s2
0.72	600000	600036
0.69	600000	600015
0.61	600000	600016
0.61	600000	600030

S1	S2	pvalue	down	up	w
600156	600566	1.5E-07	0.6	1.1	1.5
600103	600012	6.5E-06	1.3	1.5	1.0
600130	600242	1.5E-05	-2.2	-1.3	2.1

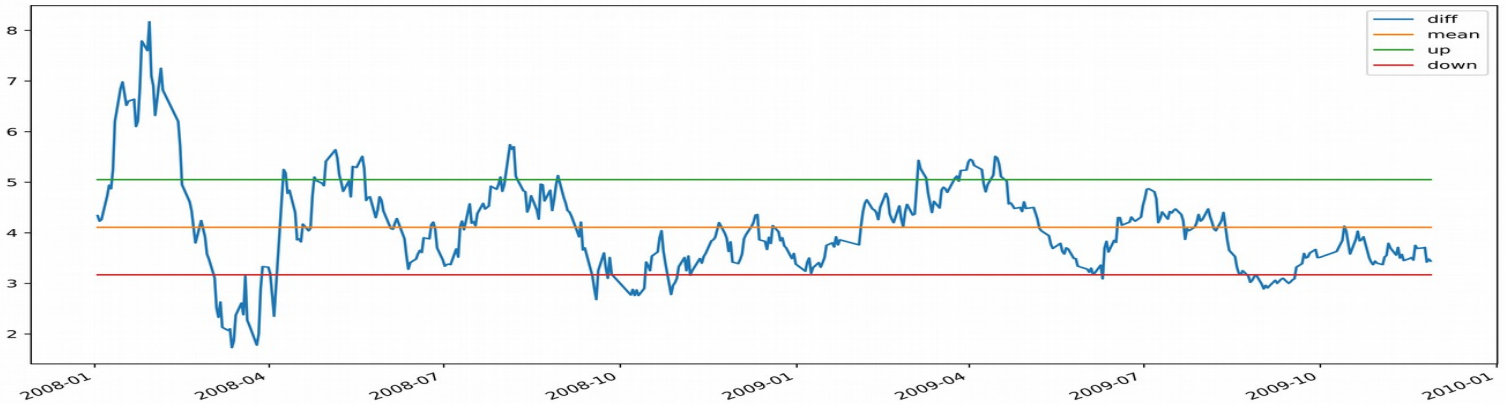
Task2: Cointegration

Example:
S1: 600015
S2: 600016
Spread price (Sp) = S2 - S1

Doing stationary test on Sp:
p-value: 9e-9 << 0.05



The pair (60015,60016) has minimum p-value on stationary test.



This spread series are stationary (平稳序列) proved by p-value << 0.05 via the test.

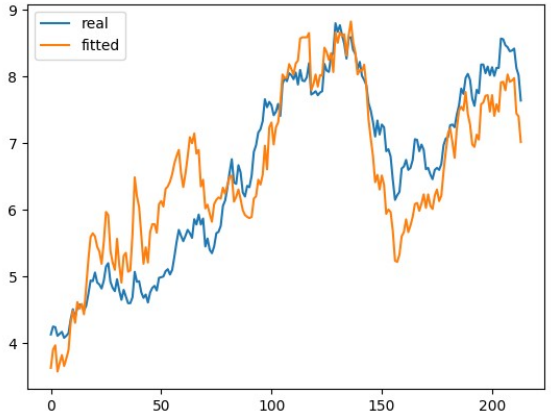
Task2: Example Strategy

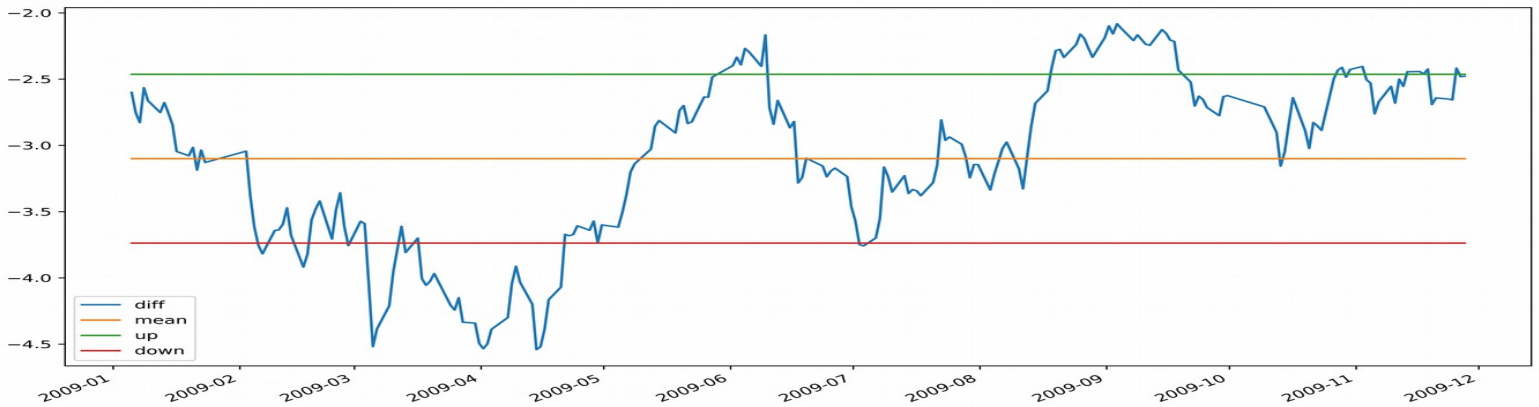
Doing linear regression via OLS:
- S2 = - 3.1 + 0.91 * S1
- Sp = S2 – 0.91 * S1

Compute up & down:
- Up = mean(Sp) + 1 * std(Sp) = - 2.46
- Down = mean(Sp) – 1 * std(Sp) = - 3.73

Strategy(S1,S2)
- If Sp > -2.46, buy S1, sell out S2;
- If Sp < -3.73, buy S2, sell out S1

Fitted S2 v.s. real S2





This **OLS spread** series (S1,S2) from 2009/1/1 to 2009/11/30.
S1: 600015
S2: 600016

Task2: Strategy List

- As there are more than 500 stocks in database, with C(500,2) = 500 * 499 / 2 = 124750 possible pairs
- We Use a pipeline to finding significant pairs automatically, after 12 hours mining finally dig out **137** significant pairs (S1,S2)
- Build rules with these pairs in mock trading during December 2009

Pairs list (part)

S1	S2	pvalue
600156	600566	0.00%
600103	600012	0.00%
600130	600242	0.00%
600103	600020	0.00%
600229	600219	0.00%
600500	600026	0.00%
600035	600033	0.01%
600400	600156	0.01%
600573	600077	0.01%
600116	600101	0.01%
600103	600156	0.03%
600308	600069	0.05%
600567	600083	0.07%
600074	600229	0.09%
600452	600116	0.09%
600567	600121	0.10%
600566	600077	0.16%
600512	600172	0.17%

Discussion

- Timeseries Prediction is pretty tough! **历史不一定会重演！**

All our method has assumption that the future is repeat of the history, such that the model learns the series distribution based on history data. But stock price is not predictable since it is influenced by numerous participants, the people who are involved in the game.

- Try to mining more sufficient statistics

The participants, the people are not predictable, however, a stock itself represented by a company has its own relatively stable characteristics, hence we can build arbitrage strategy via these robust statistical features.

- **All codes are on our Github, welcome your Star 0.0**

https://github.com/RyanWangZf/StockPricePrediction

Thank You