

Dự đoán giá trị cổ phiếu google trong thị trường bằng cách sử dụng mô hình model markov ẩn

Trịnh Quốc Huy - 20120013
Nguyễn Anh Khoa - 20120118
Võ Thị Phước Thảo - 20120191

Ngày 15 tháng 5 năm 2022

1 Bài toán và mục tiêu

Mục tiêu là dự đoán xem giá chứng khoán vào thời điểm đóng của các ngày trong tương lai như thế nào. Với bài toán này, đầu ra chính là giá của 1 cổ phiếu Google

Ý tưởng chính của bài toán là tính toán chỉ số MAP dựa vào dữ liệu trong quá khứ để dự đoán giá cổ phiếu. Để huấn luyện mô hình, ta dựa vào giá cổ phiếu vào thời điểm mở, đóng, cao nhất và thấp nhất của những ngày trước đó. Sau đó, mô hình sẽ tạo ra được chỉ số MAP cho giá cổ phiếu những ngày tiếp theo. Cuối cùng so sánh với giá trị thực tế để đánh giá mô hình

2 Mô hình:

Mô hình markov ẩn này gồm có 4 trạng thái ẩn(hidden states), trạng thái ta có thể thấy được là tỉ lệ thay đổi(được tính bằng sự chênh lệch giá lúc bắt đầu và kết thúc ngày, lúc cao nhất và lúc bắt đầu, và lúc bắt đầu và lúc thấp nhất), chúng được sử dụng để huấn luyện mô hình. Mô hình sẽ dự đoán ra được giá của cổ phiếu vào thời điểm đóng(close price).

2.1 Lí do lựa chọn:

Mô hình HMM khá trực quan, đơn giản, cũng được sử dụng khá nhiều trong các bài toán dự đoán. Bài toán này có các giá trị không thể quan sát, sự thay đổi giá cả phụ thuộc vào rất nhiều yếu tố: thị trường, chiến tranh, tình hình phát triển của công ty,..

Trong bài toán này, ta dựa vào độ chênh lệch giá trị giữa 2 thời điểm để huấn

luyện, trong đó giá trị cao nhất và thấp nhất giúp thể hiện được hướng thay đổi của giá trị cổ phiếu

2.2 Huấn luyện:

Ta trích xuất từ dataframe giá trị của cột close, high, open, low và lưu 4 numpy array. Sau đó tính các giá trị thay đổi. Ta áp dụng dựa trên thuật toán Baum-Welch bằng cách sử dụng tham số EM để đạt mức tối ưu cho HMM

$$O_t = (\frac{open - close}{open}, \frac{open - low}{open}, \frac{high - open}{open}) = (fracChange, fracHigh, fracLow)$$

Sau đó, ta thực hiện kiểm tra bằng ước lượng MAP. Dữ liệu ban đầu gồm giá cổ phiếu của d ngày và giá cổ phiếu lúc bắt đầu của ngày thứ $d+1$, chúng ta cần tính giá cổ phiếu vào thời điểm đóng cửa của ngày thứ $d+1$, cũng tương ứng với ước tính giá trị FracChange

$$\hat{O}_{d+1} = \arg \max_{O_{d+1}} P(O_1, O_2, \dots, O_d, O_{d+1} | \lambda)$$

Về số trạng thái ẩn, ta cũng chọn theo số chiều của dữ liệu là 4, ngoài ra các thông số khác ta đã xét để tìm ra những con số tối ưu nhất

2.3 Dự đoán:

Để ước tính O_{d+1} ta tìm giá trị xác suất tương ứng dựa trên bộ (fracChange, fracHigh, fracLow) và tìm giá trị lớn nhất. Giá trị có độ chính xác cao nhất sẽ sử dụng cho việc dự đoán giá cổ phiếu.

Trong đó khoảng giá trị ta thu được như sau:

- Vì open có thể thấp hơn hoặc cao hơn close, nên fracChange nhận giá trị từ âm sang dương, ta quy chuẩn về -0.1 đến +0.1
- Vì open có thể thấp hơn hoặc bằng hơn high, nên highChange nhận giá trị từ 0 đến dương vô cùng, ta quy chuẩn về 0 đến +0.1
- Vì open có thể cao hơn hoặc bằng hơn low, nên lowChange nhận giá trị từ âm vô cùng đến 0 vô cùng, ta quy chuẩn về 0 đến +0.1

2.4 Đánh giá:

MSE là giá trị sai số bình phương trung bình hoặc là lỗi bình phương trung bình. Ta đánh giá độ chính xác của mô hình dựa trên giá trị MSE, MSE càng thấp thì mô hình càng hiệu quả:

$$MSE = \frac{1}{n} \sum_{i=1}^n ((y_i - \tilde{y}_i)^2)$$

3 Dataset:

Dataset được sử dụng trong bài toán này được lấy từ kagge, nó là giá cổ phiếu của google từ năm 2012 đến 2017: <https://www.kaggle.com/vaibhavsn/google-stock-prices-training-and-test-data>

- Date: Ngày thực hiện
- Open: Giá cổ phiếu lúc bắt đầu ngày
- Close: Giá cổ phiếu lúc kết thúc ngày
- High: Giá cổ phiếu tại thời điểm cao nhất
- Low: Giá cổ phiếu tại thời điểm thấp nhất

Bộ data có chia sẵn 2 file csv tương ứng với train(Google_Stock_Price_Train.csv) và test(Google_Stock_Price_Test.csv), ta thực hiện đọc và lưu vào 2 data frame, sau đó xử lý dựa trên 2 data frame

4 Kết quả:

Sau khi thực hiện chạy mô hình nhiều lần, ta thu được kết quả 2 kết quả chính MSE như sau:

- MSE: 179.2439919244886
- MSE: 111.91221716264049
- MSE: 268.31056319879207

Trong đó kết quả tốt nhất thu được là:

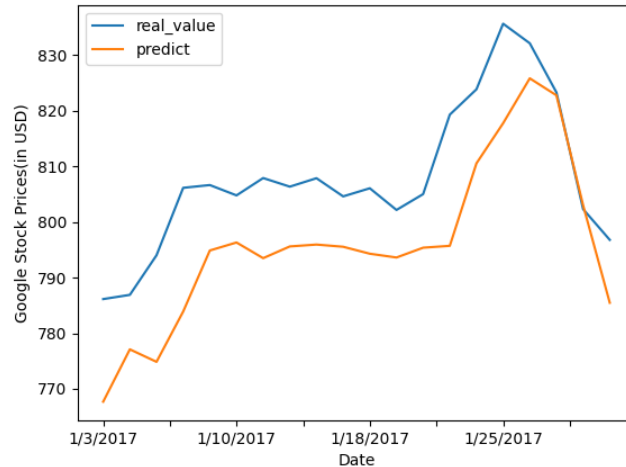
```

      real_value    predict
0      786.14    770.862959
1      786.90    780.315510
2      794.02    778.058776
3      806.15    787.145102
4      806.65    798.171429
5      804.79    799.616531
6      807.91    796.785714
7      806.36    798.903878
8      807.88    799.240408
9      804.61    798.844490
10     806.07    797.587449
11     802.17    796.904490
12     805.02    798.676224
13     819.31    799.012755
14     823.87    813.909184
15     835.67    821.154490
16     832.15    829.260918
17     823.31    826.192551
18     802.32    806.347143
19     796.79    788.728776
MSE value = 111.91221716264049

```

Hình 1: Hình ảnh kết quả

Nhìn chung giá trị dự đoán và giá trị thực tế không chênh lệch nhiều, MSE value khoảng 111. Mô hình thể hiện sự tương quan giá trị thực tế và giá trị dự đoán



Hình 2: Tương quan dữ liệu

Kết quả mô hình không quá xuất sắc nhưng nhìn chung ta có thể cải tiến thêm ở một số điểm.

5 Cải tiến

Để cải tiến mô hình, ta có thể sử dụng tập dataset cụ thể hơn (Ví dụ có thể thêm những thông tin về những công ty liên quan, điều kiện chiến tranh, dịch bệnh, kinh tế,...), nhiều thông tin hơn. Ta có thể sử dụng thêm một số thuật toán phân cụm để tiền xử lý data.

Ngoài ra, ta cũng có thể kết hợp HMM model với những mô hình học sâu, những mạng nơ ron khác để mang lại kết quả tốt hơn

6 Tài liệu tham khảo:

- [1] Viet Nguyen Trong Hoang, HiddenMarkovModelPredictingStock, <https://github.com/meocong/HiddenMarkovModelPredictingStock>
- [2] Aditya Gupta, Non-Student Member, IEEE and Bhuwan Dhingra, Non-Student member, IEEE, Stock Market Prediction Using Hidden Markov Models, https://users.cs.duke.edu/~bdhingra/papers/stock_hmm.pdf
- [3] Stock Market prediction using Hidden Markov Models, Jay-code-collection, https://github.com/Jays-code-collection/HMMs_Stock_Market