Stock Market's Price Movement Prediction With LSTM Neural Networks (2017.05)

click original link

위 링크의 페이퍼를 번역하고 공부한 글입니다.

-저자 David M. Q. Nelson, Adriano C. M. Pereira, Renato A. de Oliveira

-LSTM을 이용한 주식시장 가격이동 예측

YBIGTA10 - TH Kim, JW Son

---

1. 요약

- 일반적인 방법론은 머신러닝 알고리즘으로 과거의 가격정보를 학습시켜서 미래의 가격을 예측하는 것.
- 이 연구에서는 RNN 중에서 LSTM 을 사용한다.
- 가까운 미래에 특정 주식이 상승할지 하락할지 예측 : 평균적으로 55.9%의 accuracy를 보였음.

2. 도입

- LSTM(Long-Short term memory)요약

> LSTM은 RNN의 type 중 한 가지이다. 초기의 자료와 최근의 자료를 구별하는 능력으로 여러 문제를 성공적으로 해결하였다. 다음 output을 예측하는데 무관하다고 생각되는 input을 잊어버리면서 가중치를 조절한다. 이러한 방식으로 short sequences를 예측하는데 적합한 일반적인 RNN 알고리즘보다 long sequences 예측에 적합하다.

- 데이터 수집

> 브라질 거래소의 각 주식별 과거 가격자료를 사용하였음. 또한 기술적 지표들을 생성하여 network의 feature로 활용하였음.
> 이 데이터셋으로 모델 학습, 평가, 예측. (**15분 이후** 특정 주식의 가격이 상승할지 하락할지 예측.)

- 연구의 기대효과

> (1) 딥러닝 기술을 이용한 주가 움직임 예측 모형
> (2) 브라질 거래소 실 데이터를 이용한 검증
> (3) 기존의 baseline과 비교 및 분석을 통한 평가

3. 배경지식

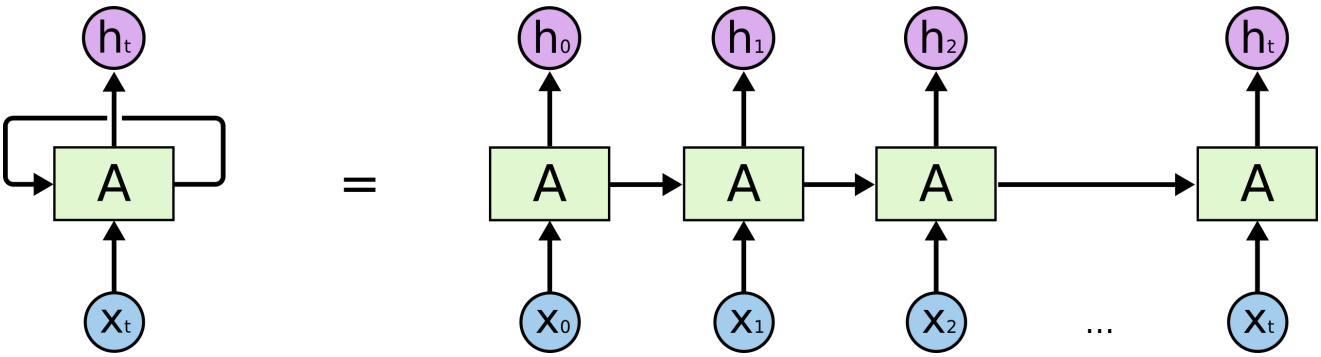
3.1 주가시장에 대한 기존 주장들

- 효율적 시장 가설(Efficient Market Hypothesis) : 시장은 이용가능한 모든 정보를 반영하고 있다. 따라서 시장을 초과하는 수익률 달성 불가능

- 랜덤워크 가설(Random-walk Hypothesis) : 주가는 과거의 가격과 독립적이다. 내일의 주가는 내일의 정보에 의존하는 것이지 오늘의 가격의 영향을 받는 것이 아니다.
- random 전략이 가장 유명한 기술적 트레이딩 방식(like MACD, RSI)보다 수익률이 좋았다.
- 반면에 어느정도 예측이 가능하다는 주장 (경제학, 통계학, 물리학, 컴퓨터과학의 연구주제), 2012년 미국주식시장 거래의 85%는 알고리즘트레이딩

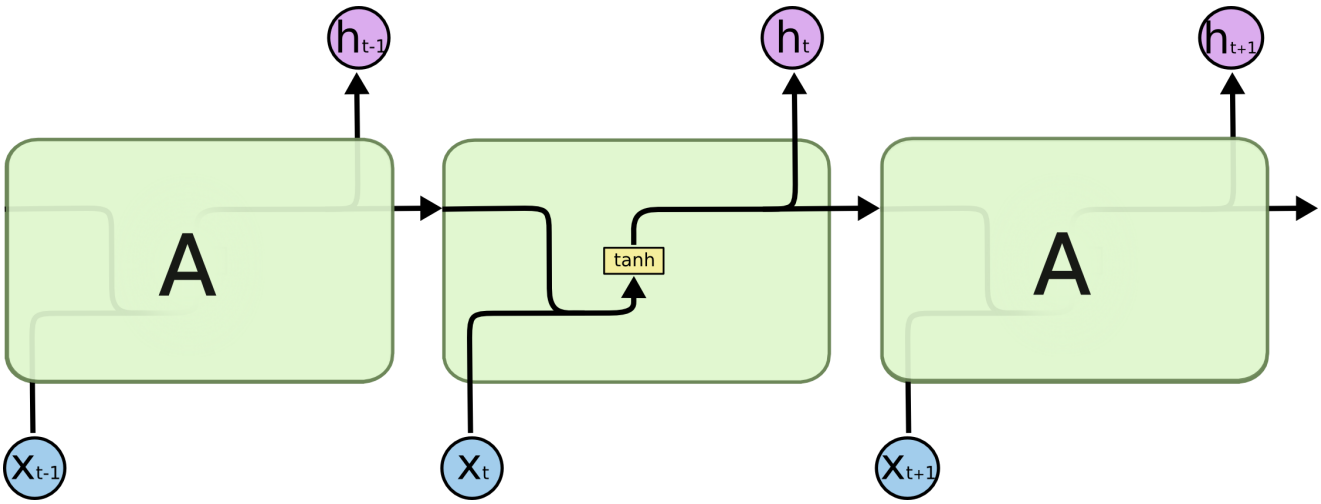
3.2. LSTM

recurrent network

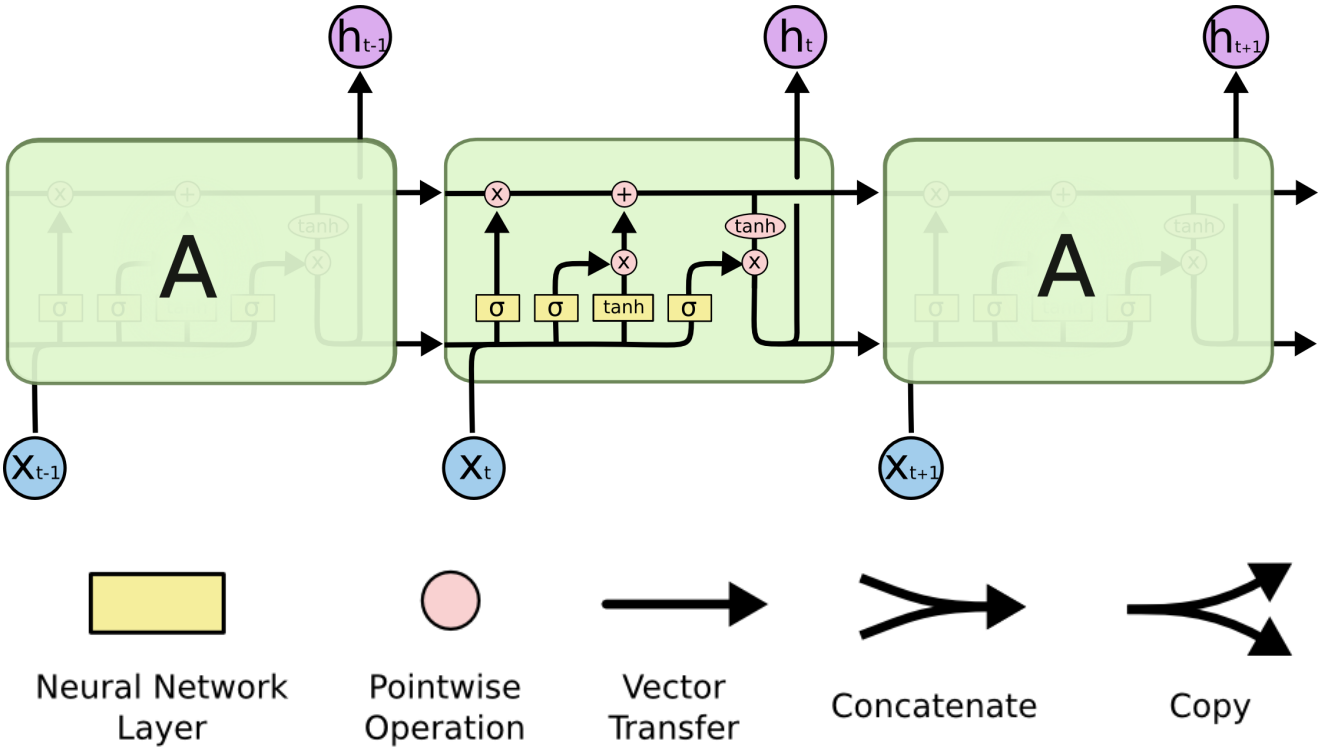


- sequence가 있는 자료 (시계열 자료)분석에 적합하다.
- 음성인식, 언어 모델링, 번역, 이미지 캡셔닝 등 많은 문제를 해결

RNN



LSTM



이미지출처:http://colah.github.io/posts/2015-08-Understanding-LSTMs/

- LSTM은 recurrent network 중에서 long term dependency를 학습 가능한 모델
- RNN은 long term data 에서 vanishing gradient 문제가 발생하는데 LSTM이 이를 해소
- 어떻게 해소할지에 대해서는 앞으로 학습계획.

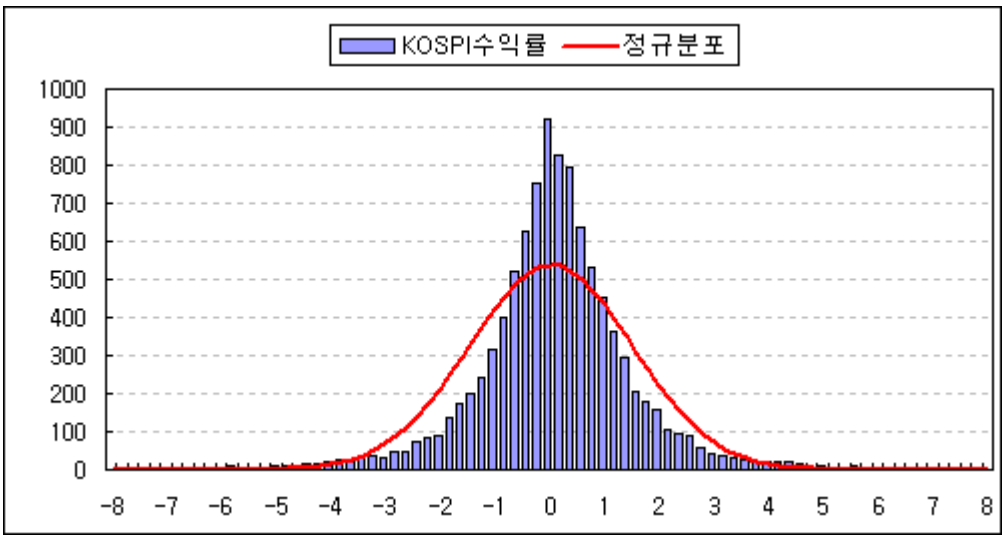
4. 방법론

> LSTM 네트워크를 활용한 모델
>
> 목표 : 15분 후 미래에 특정 주식가격의 상승/하락 예측
>
> 모델은 마지막 가격 데이터에 각 거래일을 추가로 학습하여 재생성된다.

4.1. 전처리

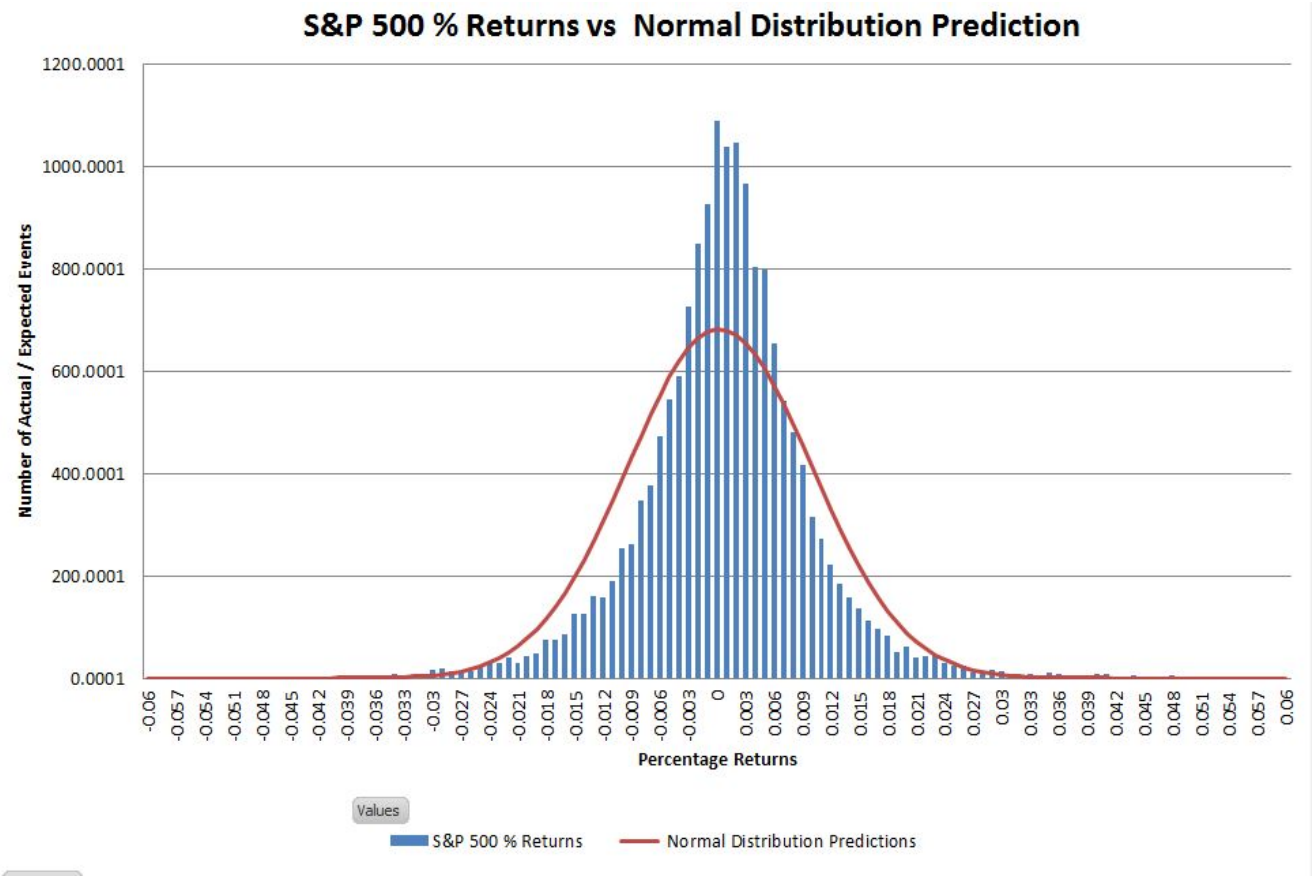
- 역사적 가격자료는 시계열, 봉 형태로 수집 (봉 : 시가, 종가, 고가, 저가, 거래량)
- 데이터는 2008~2015년, 브라질 증권거래소(우리나라 KRX) IBovespa index(우리나라 코스피)에 포함된 종목들
- 수익률은 log 변환한다. 시계열에서의 평균과 분산 정규화.

$log(p_i) - log(p_{i-1})$



1980~2016년 코스피 수익률 자료 사진확대

미국 s&p500 수익률분포(1950~2017)



- 로그수익률을 사용하는 이유 (출처 : 박상우, 주식시장을 이긴 전략들 p.56)

> 수익률의 분포를 분석시 단순수익률이 아닌 로그수익률을 사용한다.
>
> 가격변동을 단순수익률로 계산하면 왜곡현상이 발생.
>
> (ex) 가상의 주식 주가가 10,000원에서 20,000원으로 상승하였다가 다시 하락하여 10,000원으로 돌아왔다.
>
> > 단순수익률계산 : 100%수익과 -50%손실 --> 누적수익률 50% 수익
> > 로그수익률계산 : 69.31%의 수익과 -69.31%의 손실로 계산 --> 누적수익률 0%

- 로그수익률 분포 확인 코드 sk 하이닉스 1월~7월

https://gist.github.com/taesiri89/8a90fcf2922c1ac12be76fcbc995aaf3

- 가격 시계열의 random variation과 노이즈를 줄이기 위해서 exponential smoothing 진행 : 지수적으로 이동 평균에 가중치 부여.

$$z_i = \lambda \bar{x} + (1 - \lambda) z_{i-1}$$

- TA-Lib library (금융시장 분석 라이브러리) : 기술적지표(주식관련 특성을 나타내는 feature) set 생성. 각 기간마다 175개 value. (ex: 선물가격, 예상거래량, 움직임 강도, 그래프 패턴)
- y : binary value, 1 : 다음 time step에 상승 , 0 : 상승X
- i : 현재상태, j : 다음상태 --> j = i + timestep (timestep : 15분)
- y의 값이 1이 나오면 --> "buy"

$$y = 1$$
$$if$$
$$close_j > close_i$$

$$y = 0$$
$$if$$
$$close_j \leq close_i$$

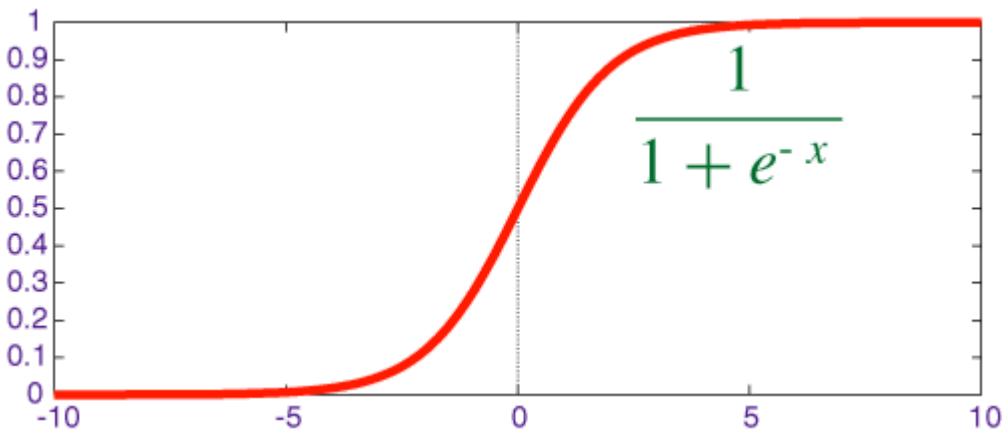
- 인공신경망은 X로부터의 k개의 인스턴스를 input으로 활용한다.
  $(X_{i-k}, \ldots, X_{i-2}, X_{i-1}, X_i)$
- X: 기술지표와 가격데이터로 구성, X로 y_j를 예측

5. 모델평가

- 각 거래일의 마지막에 인공신경망 새롭게 갱신 : 새로운 트레이닝, 벨리데이션 셋으로 가중치 재구성
- 학습 : 과거 10개월 부터 현재 거래일 앞까지의 데이터 사용
- 벨리데이션 : 저번주 자료 사용
- 내일 가격 변화 예측 : 가장 최근의 모델 사용
- 구글 tensorflow 사용
- 가격데이터와 기술적 지표를 사용하는 LSTM 인풋레이어 사용
- 활성함수는 sigmoid

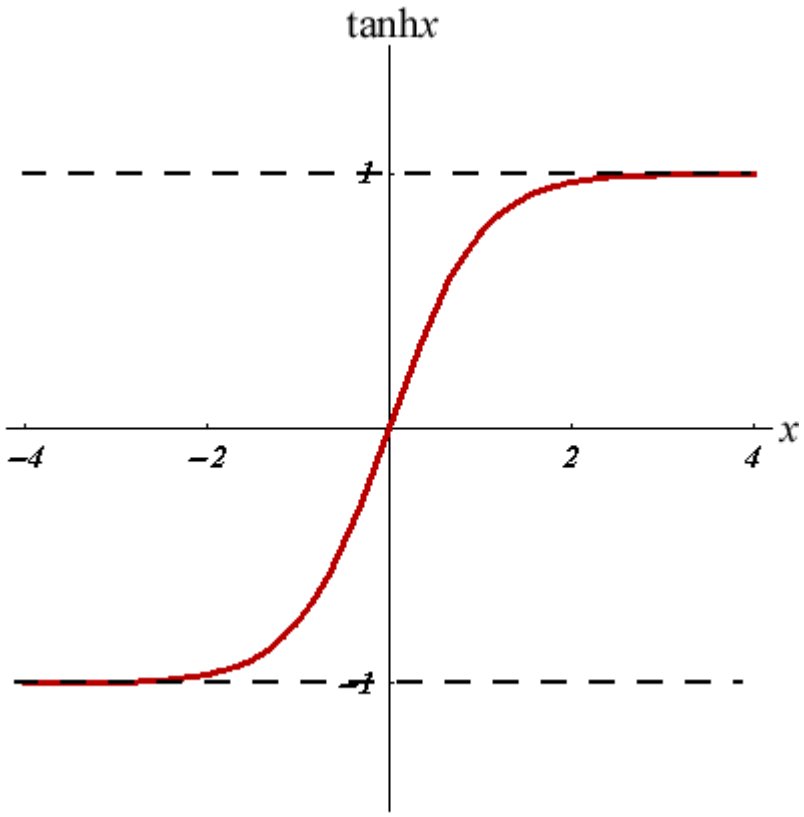
$$S(t) = \frac{1}{1+e^{-t}}$$

sigmoid 함수



- input layer는 180개 features (TA-Lib로 생성한 175개 기술적지표 포함 + 시가, 종가, 최소값, 최대값, 거래량)
- output은 $tanh$함수

tanh함수



- 2014년 브라질 거래소 상장 주식 종목 BOVA11, BBDC4, CIEL3, ITUB4 and PETR4 을 대상으로 실험
- 평가지표 : accuracy, precision, recall, F-score(P와R의 조화평균)

Predictive Model: Evaluation

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

predictive result / classification		actual result / classification: yes	actual result / classification: no
	yes	tp (true positive)	fp (false positive) ← Type 1 error
	no	fn (false negative)	tn (true negative)

$$\text{Precision} = \frac{tp}{tp+fp} \qquad \text{Recall} = \frac{tp}{tp+fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \qquad \text{True Negative Rate} = \frac{tn}{tn+fp}$$

투자전략

> y가 1로 예상되었다면 현재 i상태에서 "buy" position 그리고 j상태에서 종료 (매수, long)

실험결과

실험데이터 특성

종목	시가	종가	차이	%times go up
BOVA11	53.54	48.54	-5.00	0.471
BBDC4	35.68	30.79	-4.89	0.435
CIEL3	32.34	33.98	1.64	0.453
ITUB4	29.65	30.73	1.08	0.442
PETR4	18.39	10.03	-8.36	0.479

실험데이터 performance

종목	accuracy	precision	recall	F1
BOVA11	0.546	0.560	0.350	0.431
BBDC4	0.559	0.553	0.129	0.209
CIEL3	0.545	0.475	0.134	0.209
ITUB4	0.530	0.476	0.137	0.213
PETR4	0.533	0.563	0.231	0.327

비교분석

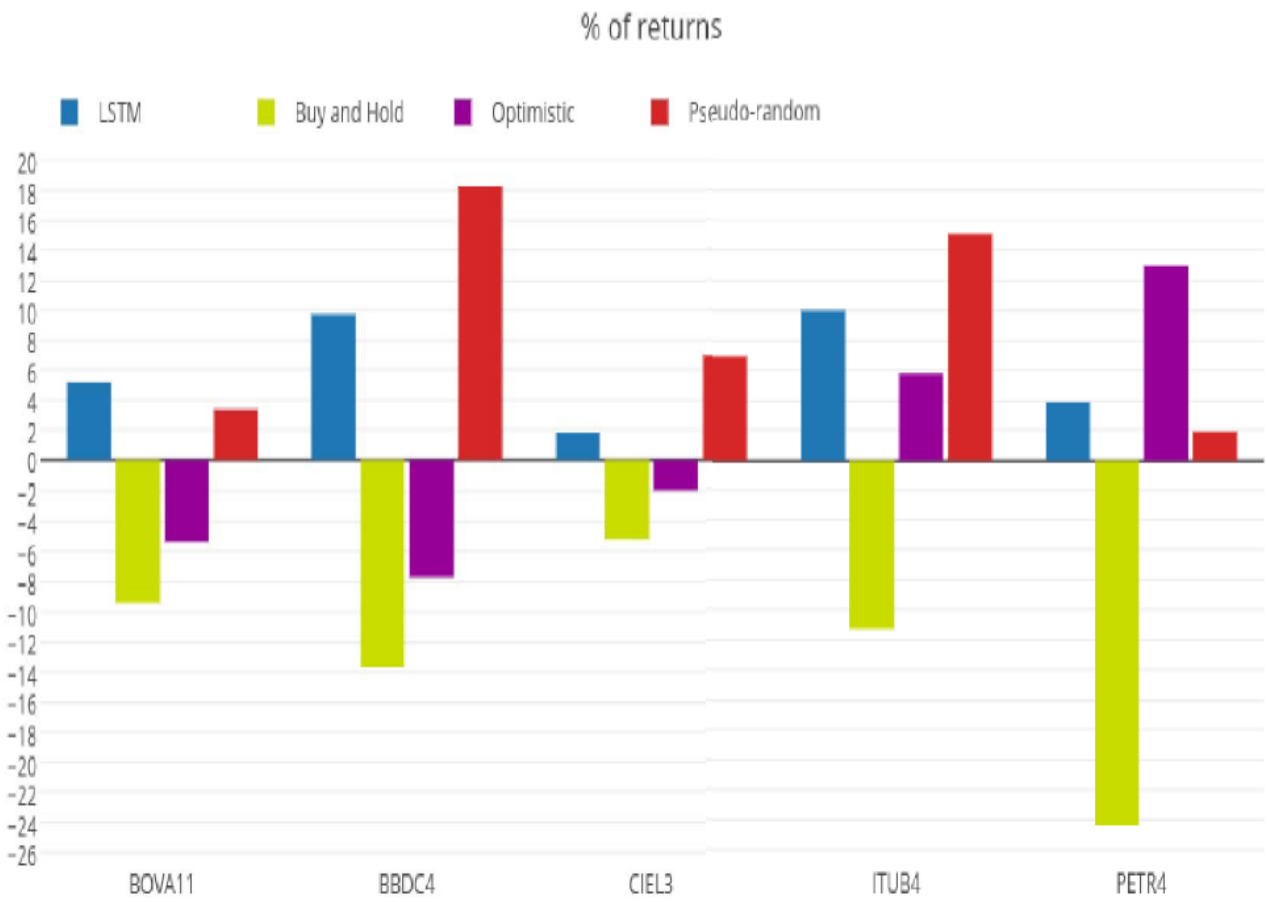


Fig. 6. Returns

- LSTM
- Buy and hold : 첫 기에 buy ... 마지막 기에 sell (중간에 거래X)
- Optimistic : 현재 시점 기준으로 어제 상승했으면 buy, 다음 시점에 sell
- Pseudo-random : class distribution (y의 분포??) 를 따른 확률에 기반해서 거래할지말지 결정

my opinion ...

참고자료

- LSTM : http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- LSTM : http://byungjin-study.blogspot.kr/2016/08/recurrent-neural-network.html
- LSTM : http://blog.naver.com/anthouse28/221026536458

참고할자료

- 머신러닝 주가방향 예측 : http://blog.naver.com/PostView.nhn?blogId=silvury&logNo=220721981693