호준

3. 코드분석

결과

로지스틱 회귀분석 결과

한계 및 보완점

-

-

* 이중분류가 아닌 다중분류로 가게되면 전혀 학습하지 못 함.

로지스틱 회귀 모델의 또다른 단점은 가중치 해석이 덧셈식이나 곱셈식이 아니기 때문에 해석이 더 어렵다는 것입니다.

로지스틱 회귀는 **완전한 분리**가 생길 수 있습니다. 만약 하나의 특성이 두 개의 클래스로 완전히 분리한다면 로지스틱 회귀 모델은 더이상 학습되지 않습니다. 이유는 최적의 가중치가 무한해서 특성에 대한 가중치가 수렴하지 않기 때문입니다. 간단한 규칙으로 두 클래스를 나눌 수 있다면 머신러닝이 필요하지 않습니다. 완전한 분리의 문제는 가중치에 대한 패널티를 주거나 가중치에 대한 사전 확률 분포를 줌으로써 해결할 수 있습니다.

로지스틱 회귀는 많은 사람들에 의해 널리 사용되었지만 제한적인 표현력(상호작용이 수동으로 추가되어야 함)과 분투하며, 다른 모델들은 더 나은 예측 성능을 나타낼 수도 있습니다. 로지스틱 회귀 모델의 또 다른 단점은 가중치를 더하는 것이 아니라 곱하는 식기 때문에 가중치의 해석이 매우 어렵습니다. 로지스틱 회귀는 **완전히 분리**될 수 있습니다. 두 가지의 클래스를 완벽하게 분리할 수 있는 특성값이 있다면, 로지스틱 회귀 모델을 더이상 훈련시킬 수 없습니다. 최적의 가중치값이 무한이 될 수 있기 때문에 그 특성값에 대한 가중치가 수렴되지 않기 때문입니다. 사실 특성값은 실제로 유용하기 때문에 이러한 경우는 운이 매우 좋지 않을 경우에 해당됩니다. 하지만 두 가지의 클래스를 구분하는 간단한 규칙이 있다면 기계 학습이 필요하지 않습니다. 완전한 분리의 문제는 가중치에 패널티를 도입하거나 가중치의 사전 확률 분포를 정의함으로써 해결할 수 있습니다.

https://eair.tistory.com/12

https://dodonam.tistory.com/251

0. (완)모듈 사용 이유

- sklearn : 머신러닝을 구현하기 위해 필요한 알고리즘 호출 라이브러리

- tensorflow : 데이터 플로우 그래프를 구현하기 위한 라이브러리

- financedatareader : 한국과 미국의 주식 관련 데이터 수집 라이브러리

- matplotlib : 파이썬 프로그래밍 언어 및 수학적 확장, numpy 라이브러리 활용한 플로팅 라이브러리

1. (완) 모델 선정 이유

- LSTM(Long Short-Term Memory) : RNN의 한 종류로 장, 단기 기억을 가능하게 설계한 신경망의 구조를 뜻한다. 주로 시계열 처리나, 자연어 처리에 사용

- 로지스틱회귀(logistic regression) : 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법, 범주형 출력변수를 설명 또는 예측하기 위한 모형이 필요한 경우에 사용

**어느 한 집단에 속하는 새로운 관찰치에 대해서 예측변수를 이용하여 특정 집단으로 분류하기위해 사용함. 이는 주가의 상승, 하락에 대한 예상을 할 수 있는 모델이다.**

2. (완) 종목 선정 이유

- 에스엠 (041510) : 현재의 산업군 중에서 가장 주목받고, 전세계적으로 인정받는 분야인 엔터테인먼트를 선정. 그 중에서 가장 대표적이고 오래된 상장기간(2000.4.27)을 가진 에스엠 선정

4. (완)보조지표 사용이유

(주식 매매전략에 따라 아래의 보조지표를 추가하였음.)

- SMA (단순이동평균, Simple Moving Average) : 주가의 방향성을 제시하는 추세지표

- 볼린저밴드(Bollinger Bands) : 주가의 변동에 따라 상, 하 밴드의 폭이 같이 움직이게하여 주가의 움직임을 밴드 내에서 판단하고자 고안된 지표

- RSI(상대강도지수, Relative Strength Index) : 가격의 상승압력과 하락압력 간의 상대적인 강도를 표시

- MACD(Moving Average Convergence & Divergence) : 주가의 단기 이동평균선과 장기 이동평균선의 수렴과 확산을 나타내는 지표  
- 스토캐스틱(Stochastic) : 현 주가 수준이 일정 기간 동안 변동했던 범위 내에서 상대적으로 어느 수준에 위치했는가를 판단하는 지표

5. 평가지표

- 오차행렬(confusion matrix, 혼동행렬) : 학습된 분류 모델이 예측을 수행하면서 얼마나 헷갈리고 있는지를 함께 보여주며 이진 분류에서의 성능지표

- F1(Precision, Recall) : 분류 모델에서 사용되는 머신러닝 평가지표(특 : precision과 recall이 모두 높으면 0.0 ~ 1.0 사이의 값을 가진다)