**SK네트웍스 Family AI 과정 12기  
 데이터 전처리 인공지능 학습 결과서**

|  |  |
| --- | --- |
| **산출물 단계** | 데이터 전처리 |
| **평가 산출물** | 인공지능 학습 결과서 |
| **제출 일자** |  |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** | 윤 권 |

## 1. 모델 비교 및 선정 이유

### 1.1 비교 대상 모델

|  |  |  |
| --- | --- | --- |
| **모델명** | **종류** | **선정 이유 요약** |
| LSTM | 시계열 순환 신경망 | 과거 주가 흐름을 반영해 미래 가격이나 상태 예측에 유리함 |
| DQN | 강화학습 모델 | 수익률 기반 보상 구조를 통해 매매 판단을 최적화 가능함 |
| LSTM + DQN | 시계열 + 강화학습 복합 | LSTM을 통해 상태 벡터 생성 → DQN이 행동 결정 (정책 네트워크 분리 구조) |

#### 실험 모델 수: 총 3종

### 1.2 최종 선정 모델: LSTM 단일 모델

##### 시계열 기반의 가격 예측 모델로서, 단순 구조에도 불구하고 안정적인 학습 성능과 예측 정확도를 보였으며, 강화학습 기반 모델(DQN, LSTM + DQN)은 reward 구조에 민감하고 불안정한 학습 경향을 보여 최종적으로 LSTM 단일 모델을 선정함.

## 2. 모델 구조 및 아키텍처

### 2.1 모델 아키텍처 도식

##### LSTM 기반 시계열 예측 모델은 과거 주가 데이터를 입력으로 받아,미래의 가격 변화율 혹은 상태를 예측합니다.

##### 전체 구조는 다음과 같이 구성됩니다.

##### 입력층 (State Vector: 5~7일 시계열 데이터)

##### → LSTM 계층 (1~2층)

##### → Dropout (과적합 방지)

##### → Dense (Fully Connected) 출력층

##### → 예측값 (수익률 또는 가격)

### 2.2 구성 요소 설명

|  |  |  |
| --- | --- | --- |
| **계층명** | **역할** | **구성 요소/설명** |
| Input Layer | 과거 시계열 상태 벡터 입력 | OHLCV 및 파생 변수 포함 (윈도우 5~7일) |
| LSTM Layer | 시계열 정보 학습 및 요약 | 64~128 유닛, hidden state 추출 |
| Dropout | 과적합 방지 | Dropout 비율: 0.2 |
| Dense Layer | 출력층 | Fully Connected Layer → 수익률 또는 가격 예측 |
| Output Layer | 예측 결과 생성 | 회귀(Linear) 출력 |

#### LSTM은 시계열 패턴을 학습하여 다음 시점의 가격 움직임을 예측하며, 이 예측을 바탕으로 에이전트가 매매 결정을 할 수 있도록 합니다.

#### Dropout 과 같은 규제 기법은 학습 안정성 향상에 도움을 주며, Dense 출력층은 회귀 문제에 적합한 형태로 구성됩니다.

## 3. 학습 설정 및 하이퍼파라미터

### 

|  |  |
| --- | --- |
| 항목 | 값 |
| 학습 데이터 수 | 약 36,000건 (에피소드 단위 상태 벡터 기준) |
| 검증 데이터 수 | 약 9,000건 |
| 에폭(Epoch) 수 | 50 |
| 배치 크기 (Batch Size) | 64 |
| 학습률 (Learning Rate) | 0.001 |
| 옵티마이저 | Adam |
| 손실 함수 | MSELoss (회귀 예측) |
| 조기 종료 기준 | 검증 손실(validation loss) 5회 연속 증가 시 종료 |

#### 손실 함수는 수익률 또는 가격 예측이 목적이므로 회귀 기반 손실 함수(MSE) 사용

#### 검증 손실 기준으로 조기 종료를 설정해 과적합을 방지함

#### 배치 크기와 학습률은 실험적 최적화 과정을 거쳐 설정됨

#### 학습 데이터는 슬라이딩 윈도우로 구성된 시계열 상태(State) 벡터 기준으로 분할된 건수

## 4. 학습 결과 및 성능 평가

### 4.1 학습 결과 요약

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델** | **MSE(평균제곱오차)** | **MAE(평균절대오차)** | **RMSE(평균제곱근오차)** | **R^2 (결정계수)** | **평균 예측 수익률 오차 (%)** |
| LSTM | 0.0157 | 0.092 | 0.125 | 0.21 | ±4.38% |

#### MSE/MAE: 회귀 기반 오차 기준. 수익률 예측 정확도가 낮음

#### R^2: 0.21로, 대부분의 변동성을 설명하지 못함

#### 평균 수익률 오차: ±4.38% → 실거래 기준으로는 과도한 오차 범위

### 4.2 해석 및 분석

#### 전반적인 성능 저조: LSTM은 시계열 정보를 일부 학습하였으나, 가격 급등락이나 외부 이벤트 반영이 어려움

#### R^2=0.21은 입력 피처들이 수익률 변동의 대부분을 설명하지 못함을 시사

#### 수익률 예측값 분포는 실제보다 과도하게 보수적 → 시장 급변 시 예측 실패 빈도가 높음

#### 모델의 한계: 단순 LSTM 구조는 고차원 시장 정보를 충분히 반영하지 못하며, 학습 대상인 주식 가격의 비선형성과 고변동성을 처리하는 데 한계가 있음

## 5. 과적합/과소적합 대응

### 5.1 적용 기법 요약

|  |  |  |  |
| --- | --- | --- | --- |
| **기법** | **설명** | **적용 여부** | **효과 평가** |
| Dropout | 과적합 방지를 위한 무작위 뉴런 제거 | O (0.2 적용) | 과적합 억제에 기여했으나 성능 자체는 개선되지 않음 |
| 초기 종료 (Early Stopping) | 검증 성능 약화 시 학습 중단 | O | 과적합은 방지했으나 수렴 속도는 불안정함 |
| 학습률 감소(ReduceLROnPlateau) | Loss 감소 정체 시 학습률 자동 감소 | O | 일부 개선 효과 있었지만 한계적 |
| 교차 검증 | 데이터 분포의 안정적 평가 | X | 시계열 특성상 사용하지 않음 (time-series split 필요) |

### 5.2 과적합 여부 판단

##### 학습 손실과 검증 손실 차이는 크지 않았음

##### 그러나 Epoch 10~30 구간에서는 loss 진동(불안정한 수렴)이 관측됨

##### 이는 과적합이 아닌 과소적합 또는 데이터 표현력 부족의 가능성을 시사

### 5.3 분석 요약

#### 구조적으로 간단한 LSTM 단일 모델은 일정 수준 이상으로 성능을 끌어올리기 어려움

#### Dropout, 조기 종료, 학습률 스케줄링 등 일반적인 정규화 기법은 적용되었으나, 예측 정확도 향상에는 기여하지 못함

#### 원인은 피처 부족, 외생 변수 미포함, 데이터 전처리 과정의 부실함 또는 학습 인자 설정의 미숙함으로 분석됨

## 6. 결론 및 향후 계획

### 6.1 최종 선정 모델: LSTM 단일 모델

#### 단일 LSTM 구조는 시계열 데이터를 다루는 데 있어 구조적 단순성과 직관성을 갖추고 있으나, 실제 주가 데이터의 복잡성과 비선형성에 대한 대응력은 부족한 것을 나타났다.

### 6.2 활용 방안

#### 주가 수익률 또는 변동성 예측을 통해 강화학습 에이전트의 보상 기반 정보로 제공

#### 향후 투자 전략 구성 시 보조 지표 또는 필터링 조건으로 활용 가능

#### 가격 예측 자체보다는 변화 방향성 예측 (상승/하락) 분류로 전환하는 것도 검토

### 6.3 향후 계획 (모델 성능 개선)

|  |  |
| --- | --- |
| **개선 전략** | **설명** |
| 입력 변수 확장 | 기술적 지표(RSI, MACD 등), 거래량 기반 파생 변수, 종목별 변동성 등 추가로 표현력 강화 |
| 다중 시계열 윈도우 실험 | 다양한 길이의 슬라이딩 윈도우(3일, 7일, 14일 등) 비교 실험 진행 |
| 학습 데이터 전처리 고도화 | 이상치 탐지 기준 정교화, 비정상 구간 클리핑, 수익률 중심 정규화 방식 개선 |
| 출력 구조 변경 | 회귀 → 방향성 분류(binary classification) 전환 시 정확도 개선 가능성 탐색 |
| 정규화 방식 실험 | MinMax, Z-score 외에 log return, robust scaler 등 대안 정규화 방식 비교 |
| 성능 평가지표 다변화 | RMSE 외에 directional accuracy, profit factor 등 투자 관점 성능지표 추가 |

### 6.4 중장기 계획

#### LSTM 단일 모델을 기반으로 입력/출력 구조를 최적화하며 반복 실험 진행

#### 이후 성능이 임계치 이상으로 개선되지 않을 경우, LSTM + DQN 구조로 확장하는 방안 고려

#### 최종적으로 실거래 시나리오 테스트 및 리스크 기반 백테스트를 통한 전략 검증 수행 예정