Прогнозирование цен на акции с использованием нейронных сетей и анализа новостей

Можаев Роман

Постановка задачи

Задача заключается в создании системы, которая предсказывает цены на акции на основе исторических данных и анализа новостей. Это необходимо для помощи инвесторам и трейдерам в принятии решений, а также для автоматизации торговых стратегий. Система будет использовать нейронные сети для анализа временных рядов и обработки текстовых данных из новостей, чтобы учитывать влияние событий на рынок.

Формат входных и выходных данных

Входные данные:

1. Исторические данные по ценам акций (например, открытие, закрытие, максимум, минимум, объем торгов) в формате временных рядов.

   Размерность: (N, T, F), где N — количество акций, T — временные шаги, F — количество признаков (например, 5 для OHLCV).

2. Текстовые данные новостей, связанных с компаниями или рынком.

   Размерность: (N, T, S), где S — длина текста (после предобработки).

Выходные данные:

1. Прогнозируемая цена акции на следующий временной шаг (например, цена закрытия).

   Размерность: (N, 1).

Метрики

1. **Mean Absolute Error (MAE)** — измеряет среднюю абсолютную ошибку в предсказаниях цен в долларах. Целевое значение: MAE < 1-2% от средней цены актива. Это выбрано, так как MAE интерпретируема в реальных единицах и важна для финансовых приложений.

2. **Root Mean Squared Error (RMSE)** — штрафует большие отклонения сильнее, что полезно для оценки стабильности модели. Целевое значение: RMSE < 2-3% от средней цены.
3. **Accuracy (точность направления)** — процент случаев, когда модель правильно предсказала направление изменения цены. Ожидаемое значение: 60-70%.

Эти метрики выбраны, так как они позволяют оценить как точность прогноза, так и его полезность для принятия решений. Они понятны и полезны для оценки экономического эффекта, а значения выбраны как достижимые для качественных моделей на реальных данных.

Валидация

Способ разделения: Используется временное разделение (time-based split), так как данные — временной ряд. Например, 80% данных для обучения (train), 10% для валидации (val), 10% для тестирования (test), с сохранением хронологического порядка.

Воспроизводимость: Фиксация random seed для всех операций (например, `np.random.seed(42)`), использование конкретных дат для разделения (например, данные до 2022-01-01 — train, 2022-01-01 до 2023-01-01 — val, после 2023-01-01 — test). Это гарантирует одинаковые выборки при повторных запусках.

Данные

1. **Исторические данные по акциям:**
   - Источник: Yahoo Finance или Alpha Vantage.
   - Особенности: Данные могут содержать пропуски, шумы или аномалии.
2. **Новостные данные:**
   - Источник: NewsAPI, Reuters, или Google News RSS.
   - Особенности: Тексты могут быть неструктурированными, содержать шум (например, рекламу).

Проблемы:

- Недостаток данных для некоторых акций.
- Зависимость качества прогноза от доступности новостей.

Моделирование

Бейзлайн

Простейшее решение — использование линейной регрессии для предсказания цены на основе исторических данных. Модель обучается на признаках, таких как скользящие средние и разницы цен.

Основная модель

1. **Архитектура:**
    - Для временных рядов: LSTM.
    - Для текстовых данных: BERT для извлечения эмбеддингов новостей.
    - Объединение: Concatenate эмбеддингов новостей с выходом LSTM, затем полносвязные слои для предсказания.
2. **Обучение:**
    - Используется Adam optimizer с learning rate 1e-4.
    - Функция потерь: Mean Squared Error (MSE).
    - Регуляризация: Dropout и L2-регуляризация.

Внедрение

1. **Формат:**
    - Модель будет упакована в Docker-контейнер и развернута как REST API сервис.
2. **Дополнительные компоненты:**
    - База данных для хранения исторических данных и новостей.
    - Планировщик задач (например, Celery) для регулярного обновления данных и переобучения модели.
3. **Интерфейс:**
    - Веб-интерфейс для визуализации прогнозов.
    - Возможность интеграции с торговыми платформами через API.