Сравнение генеративных моделей с традиционными подходами в задаче предсказания цен на фондовом рынке

Казакова Анастасия

17 декабря 2024 г.

Аннотация

В последние годы в области финансовых технологий наблюдается быстрый рост интереса к применению генеративных моделей для решения задач предсказания цен на фондовом рынке. Традиционные подходы, такие как временные ряды - модели ARIMA и авторегрессионные модели - LSTM, зарекомендовали себя в предсказаниях временных рядов. Но эволюция моделей машинног обучения не стоит на месте - на смену традиционным методам приходят новые - генеративные подходы. Они хорошо показали себя в генерации изображений, и логичным будет протестировать их возможности на традиционных, но повседневных задачах. Генеративные модели способны выявлять скрытые паттерны в больших объемах данных (латентное простарнство) - возможно, они будут более перспективными для прогнозирования рыночных тенденций?

В данной работе проводится сравнение традиционных методов и генеративных моделей на примере предсказания цен акций, с целью выявления сильных и слабых сторон каждого подхода, а также оценки их точности, устойчивости и способности к адаптации в условиях меняющихся рыночных условий.

1 Лит обзор

В прошлых работах были исследованы традиционные методы LSTM и ARIMA[1], TS[2] . В посоледнее время есть попытки применять GAN[3] и VAE[4] к задачам данного типа.

2 Результаты работы моделей

Для задачи предсказания ценового ряда на 21-й день на основе данных за предыдущие 20 дней можно использовать различные методы машинного обучения. У вас есть ценовые данные, включающие 5 показателей: volume, open price, high price, low price, close price для каждого дня. Задача заключается в том, чтобы на основе этих данных предсказать упадёт или возрастёт цена закрытия на 21-й день.

В предобработке данных превращаем close price в diff close price или binary close price (1 если diff close price больше 0 и 0 в обратном случае).

2.1 LSTM

Модель LSTM построена для решения двух задач: регрессии (прогнозирование цены через 21 день) и бинарной классификации (определение роста или падения цены). Данные нормализуются, создается бинарная метка на основе изменения цен, и формируются последовательности для обучения. Модель имеет два выхода: для бинарной классификации (сигмоида) и для регрессии (линейная активация). После обучения модели оцениваются потери и точность предсказаний для обеих задач на тестовых данных. Экспериментально оказалось, что такой комбинированный лосс лучше, чем лосс по бинарной метрике / вещественной.

accuracy = 57

Такая точность достигается на неглубокой стандартной модели LSTM. Она является статичтически значимой : точность $baseline = 50$.

Значит, LSTM можно использовать для решения такой задачи. Логично предположить, что при увеличении глубины скор повысится.

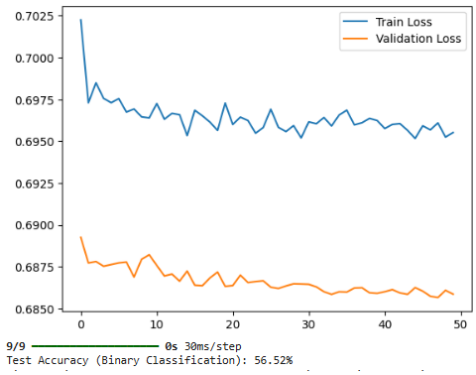

Рис. 1: lstm

2.2 ARIMA как *baseline* модель

ARIMA используется как *baseline* - простая модель, решающая задачу. ARIMA не может работать с многомерными временными рядами, поэтому экзогенные переменные в ней не учитывались. В этой задаче ARIMA используется для прогнозирования изменений цен ('$closediff$'). Данные разделены на обучающую и тестовую выборки. Модель с параметрами (5, 1, 0) обучается на истории и прогнозирует изменения для тестовых данных. Точность оценивается через процент правильных прогнозов роста/падения и метрики ошибок (MSE, MAE). accuracy = 51

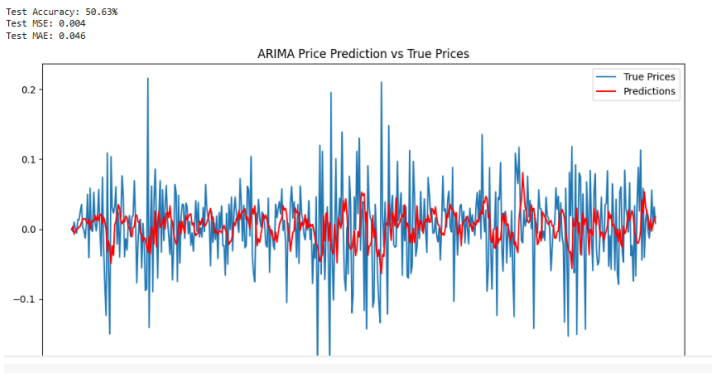

Рис. 2: arima

По рисунку видно, что модель пытается угадать зависимости (предсказать close diff), но не всегда попадает в знак, поэтому base line здесь не лучше, чем подбрасывание монетки : необходимо обучить модель с экзогенными переменными.

2.3 SARIMAX

В данной задаче модель SARIMAX используется для прогнозирования изменений цен ('$close_diff$') на основе исторических данных о ценах и экзогенных переменных (таких как 'open', 'high', 'low', 'volume'). Модель обучается на обучающей выборке, используя параметры ARIMA (p, d, q) для учета трендов и сезонности. Затем модель прогнозирует изменения для тестовой выборки, и точность её работы оценивается через процент правильных прогнозов (оценка роста или падения) и метрики ошибок (MSE).

accuracy = 60

2.4 VAE

В данном коде реализована модель вариационного автокодировщика (VAE) для предсказания направления изменения цен на основе временных рядов. Модель состоит из энкодера и декодера: энкодер преобразует входные данные в скрытые представления (latent variables), а декодер

Рис. 3: SARIMAX

восстанавливает исходные данные. В процессе обучения оптимизируется функция потерь, которая включает среднеквадратичную ошибку (MSE) для реконструкции и дивергенцию Кульбака-Лейблера (KL) для регуляризации латентного пространства. Модель обучается на обучающей выборке, и после завершения тренировки на тестовых данных предсказывается направление изменения цен с использованием изменения значения 'close'. Точность модели оценивается с помощью метрики accuracy.

accuracy = 71

Рис. 4: vae

2.5 GAN

GAN для этой задачи обучить не поулчилось : дискриминатор и генератор сложно настроить так, чтолбы обучалась модель - accuracy был низким. Дейсвительно, работ, в которых изучают применение gan к временным рядам немного - неспроста.

2.6 Выводы

Для быстрой ненастраиваемой генерации легче всего пользоваться SARIMA, вводя новую колонку – diff column

Если область деятельности узкая, то есть резон поискать на google scholar аналогичныее работы (например, медицинская тематика и VAE)

Для более детальной настройки можн воспользоваться LSTM / подобрать гиперпараметры SARIMAX

GAN для этой задачи обучить не поулчилось : дискриминатор и генератор сложно настроить так, чтолбы обучалась модель - accuracy был низким. Дейсвительно, работ, в которых изучают применение gan к временным рядам немного - неспроста.

2.7 links

[1] S. Siami-Namini and A. S. Namin, "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM," pp. 1–19, 2018

[2] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10614 LNCS, pp. 729–730, 2017.

[3] K. Zhang, G. Zhong, J. Dong, S. Wang, and Y. Wang, "Stock Market Prediction Based on Generative Adversarial Network," Procedia Computer Science, vol. 147, pp. 400–406, 2019.

[4] GP-VAE: Deep Probabilistic Time Series Imputation Vincent Fortuin†1, Dmitry Baranchuk†2, Gunnar Rätsch1, and Stephan Mandt3