А. В. Сочивко

КРОСС-ВАЛИДАЦИЯ ФИНАНСОВЫХ МОДЕЛЕЙ

Описаны основные задачи кросс-валидации, проблемы её применения для задач финансов и адаптированные методы, решающие эти проблемы

Введение

У подавляющего большинства практикующих аналитиков рынков и авторов торговых стратегий, основным показателем эффективности торговых моделей является демонстрация её результатов на исторической симуляции. Однако, высокая эффективность на тестовой исторической выборке легко достигается за счёт тестирования альтернативных конфигураций торговой стратегии. Переобученная модель разочарует при использовании, ведь доходы окажутся меньше ожидаемых, сравнивая с тестовой выборкой. По этой причине, переобучение можно считать одной из основных причин провала отдельных торговых стратегий и целых фондов.[1] Чтобы избежать переобучения, специалисты по машинному обучению применяют кросс-валидацию. Простейшая кросс-валидация - это дробление наблюдений на два подмножества: тренировочное и тестовое. Каждое наблюдение в полной совокупности данных принадлежит одному и только одному подмножеству[2]. Это сделано для того, чтобы избегать утечки из одного подмножества в другое. Существует ряд альтернативных проверок, из которых одной из самой популярных является k-блочная кросс-валидация[3]. На Рисунке 1 изображён принцип 4-блочной кросс-валидации.

Рис. 1 – 4-блочная кросс-валидация

На Рисунке 2 изображён пример как выглядит 4-блочная кросс-валидация для ценового ряда.

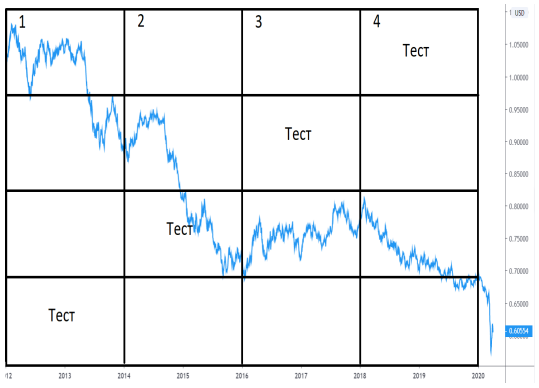

Рис. 2 – 4-блочная кросс-валидация на примере ценового ряда

I. Проблемы кросс-валидации в финансах

Применение k-блочной перекрёстной проверки оказывается безуспешной в финансах из-за двух особенностей.

Первая проблема: наблюдения не берутся из взаимно независимого случайного процесса, т.е. в один момент времени может быть открыто несколько торговых позиций, а т.к. решения по этим торговым позициям зависят от одних и тех же участков ценового ряда, эти наблюдения взаимозависимы. Таким образом возникает систематическая утечка данных. Проблема утечки или заглядывания в будущее тоже является распространённой и очень опасной проблемой при тестировании торговых стратегий.

Вторая проблема: постоянное использование одних и тех же тестовых подмножеств в процессе работы приводит к систематическому переобучению. Поэтому возникли специфичные для финансовой сферы методы кросс-валидации.

II. Прочищенная кросс-валидация

Чтобы справиться с проблемой взаимно зависимых наблюдений, можно применить метод прочищенной k-блочной перекрёстной проверки.[4] Метод заключается в том, чтобы удалить из тренировочного подмножества те наблюдения, временные метки которых накладываются по времени на метки наблюдений из тестового подмножества. Удалённые области помечены на Рисунке 3 жёлтым цветом. Также исключаем из тренировочного подмножества наблюдения, которые хронологически следуют сразу за наблюдениями тестового подмножества, чтобы уменьшить влияние внутрирядовой корреляции. Об-

ласть с высокой внутрирядовой корреляцией показана на Рисунке 3 красным цветом.

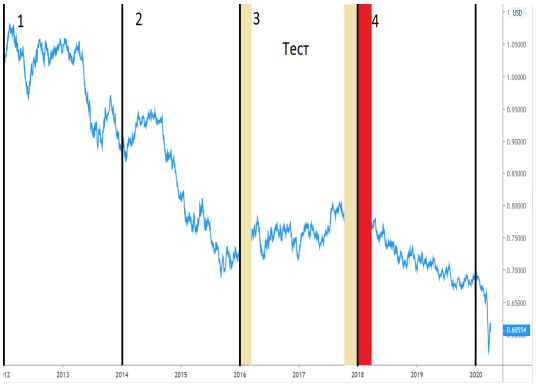

Рис. 3 – Пример прочищенной кросс-валидация

III. Комбинаторная кросс-валидация

Уменьшить влияние второй проблемы можно, если увеличить количество траекторий, по которым можно проводить кросс-валидацию. Метод комбинаторной кросс-валидации позволяет получить больше траекторий из тех же данных.[5] Разделим наблюдение на N групп, указав, что k из них будут тестовыми группами. Тогда количество возможных сочетаний подмножеств равно.

$$\binom{N}{N-k} = \frac{\prod_{i=0}^{k-1}(N-i)}{k!} \quad (1)$$

Тогда количество уникальных траекторий на этих сочетаниях равно.

$$\phi[N,k] = \frac{k}{N}\binom{N}{N-k} = \frac{\prod_{i=1}^{k-1}(N-i)}{(k-1)!} \quad (2)$$

Таким образом, поделив данные на 10 групп, 2 из которых будут тестовыми, получим $\frac{2}{10}\binom{10}{8} = 9$ траекторий. Ещё одной полезной особенностью этого решения является возможность подсчитать вероятность, что проверяемая модель переобучена.

IV. Выводы

Используя решения проблем кросс-валидации в финансах, которые предложены в описанных методах, можно адаптировать k-блочную и менее популярные методы кросс-валидации под финансовые задачи. Методы гибкие: их можно применять как по отдельности, так и в связке. Эти решение отлично дополняются тестированием на синтетических, сгенерированных данных, которые соответсвуют разным потенциальным сценариям движения цен. Используя комбинаторную кросс-валидацию появляется возможность подсчитывать вероятность переобучения, а значит можно отбирать модели с низким значением этой вероятности.

1. The 10 Reasons Most Machine Learning Funds Fail / Marcos Lopez de Prado// Journalof Portfolio Management, Forthcoming. – 2018.
2. Python Machine Learning / Sebastian Raschka // Packt Publishing(Birmingham-Mumbai). – 2015,– С. 173-176.
3. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition) / Trevor Hastie, Robert Tibshirani, Jerome Friedman // Springer Series in Statistics. – 2008, – С. 247.
4. The Probability of Backtest Overfitting / David H. Bailey, Jonathan Borwein, Marcos Lopez de Prado, Qiji Jim Zhu // Journal of Computational Finance (Risk Journals). – 2015, Forthcoming – С. 1-34.
5. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance / David H. Bailey, Jonathan Borwein, Marcos Lopez de Prado, Qiji Jim Zhu // Notices of the American Mathematical Society, 61(5), 2014, С. 458-471.

Сочивко Андрей Викторович, магистрант кафедры информационных технологий автоматизированных систем БГУИР, andreysoch@gmail.com.

Научный руководитель: Шилин Леонид Юрьевич, декан факультета информационных технологий и управления БГУИР, доктор технических наук, профессор, dekfitu@bsuir.by.