1 Сигнатурные методы

1.1 Слайд 1

Добрый день, уважаемые слушатели! Меня зовут Мащенко Кирилл, я являюсь студентом 609 актуарно-финансовой группы. Сегодня я хочу Вам рассказать про свою дипломную работу по теме Симулирование случайных процессов с использованием сигнатурных методов, выполненную под научным руководством Житлухина Михаила Валентиновича.

Настоящая дипломная работа посвящена исследованию и обобщению одного генеративного метода для случайных процессов, основанного на теории сигнатур. Пусть имеется результат наблюдения за одной или несколькими траекториями некоторого случайного процесса, который описывает изменение какой-либо величины - тогда возникает вопрос, как численно произвести достаточно большое число новых траекторий, которые были бы в некотором смысле “похожи” на наблюдаемый процесс?

Центральную роль в предложенном методе играет преобразование траектории случайного процесса, состоящее в вычислении его сигнатуры.

В последнее десятилетие сигнатурные методы находят применение в машинном обучении, и в сочетании с другими моделями выигрывали первые места в различных соревнованиях, например, по распознаванию изолированных китайских символов, а также по вычислительной технике в кардиологии. Помимо этого, в последние годы сигнатурные методы нашли применение в задачах финансовой математики, связанных с хеджированием производных инструментов.

1.2 Слайд 2

Сформулируем некоторые определения.

Путём в $\mathbb{R}^d$ будем называть непрерывное отображение X из некоторого интервала $[a, b]$ в $\mathbb{R}^d$, зависящее от времени.

Индуктивно введем формальное определение повторного интеграла по компонентам пути, что будет являться одним элементом сигнатуры.

1.3 Слайд 3

Сигнатурой пути является бесконечное семейство таких повторных интегралов от пути по всевозможным индексам, однако на практике, поскольку повторные интегралы являются определенными, это является набором вещественных чисел, хорошо обобщающих и полноценно описывающих путь, так как сигнатуры содержат в себе полную информацию о его аналитических и геометрических свойствах.

1.4 Слайд 4

Рассмотрим несколько теорем о сигнатурах.

Важное свойство сигнатуры состоит в том, что произведение двух её элементов может быть всегда представлено в виде суммы других элементов, которая зависит только от их мульти-индексов, а именно от их шафл-произведения, что является множеством перестановок совместных индексов с определенными условиями.

Это свойство показывает, что члены сигнатуры не являются алгебраически независимыми, а также дает возможность работать с линейными объектами вместо произведений.

Еще одной важной теоремой о сигнатурах является тождество Чена, позволяющее вычислять сигнатуры для данных, которые являются результатом наблюдения за некоторой величиной, и чей путь является кусочно-линейной функцией и обладает неудобной для вычисления параметризацией.

Это тождество позволяет нам, зная значения сигнатур на двух отрезках, легко вычислить сигнатуру на отрезке, являющимся их объединением.

Таким образом, можно легко вычислить все сигнатуры на каждом линейном отрезке кусочно-линейной функции, которые обладают очень удобной параметризацией, а потом вычислить сигнатуру всего пути, по очереди присоединяя каждый следующий отрезок с помощью этого тождества.

1.5 Слайд 5

Сформулируем еще несколько свойств. Сигнатуры не зависят от сдвига или начальной точки, а также от репараметризации времени.

Эти свойства показывают, что сигнатура зависит от геометрических свойств пути. Можно также заметить, что первый уровень сигнатуры по определению является приращением пути по каждому аргументу, а второй уровень сигнатуры связан с таким понятием, как площадь Леви.

В общем случае путь не определяется однозначно его сигнатурой. Например, мы смогли увидеть, что по сигнатуре нельзя восстановить точную скорость, с которой этот путь проходится (из-за инвариантности сигнатуры при репараметризации времени). Однако для непересекающихся путей по сигнатуре можно полностью определить все точки, через которые пройдёт путь, и порядок их обхода.

1.6 Слайд 6

Из теоремы о шафл-произведении мы смогли увидеть, что члены сигнатуры не являются алгебраически независимыми. Этой проблемы можно избежать, если вместо сигнатур рассматривать логарифмические сигнатуры, которые представляют из себя независимый набор признаков.

Для определения понятия логарифмической сигнатуры необходимо использовать алгебру формальных степенных рядов.

Сигнатура может быть “закодирована” как элемент этого пространства, то есть элементы сигнатуры рассматриваются как соответствующие коэффициенты ряда, что позволяет “перейти” в алгебру рядов. В ней можно выполнять все операции с рядами, а коэффициенты полученного ряда интерпретировать как элементы полученной сигнатуры.

Используя это, введем понятие логарифической сигнатуры пути следующим образом, где данный знак обозначает операцию произведения рядов.

2 Генеративные модели

2.1 Слайд 7

Перейдем непосредственно к генеративным моделям.

Генерация новых данных, похожих по распределению на исходные, может быть

необходима во многих сферах. Например, для анонимизации данных, когда данные конфиденциальны и нужно сгенерировать другие примеры того же распределения. Примером этого могут являются финансовые и медицинские данные. Помимо этого, бывают ситуации, когда количество данных заведомо мало из-за ограничения на число экспериментов или доступ к данным. Также генеративные модели могут использоваться для тестирования стратегий, которое нельзя проводить на исторических данных во избежание переобучения.

2.2 Слайд 8

Рассмотрим работу 5 авторов, на основе которой проводилось исследование в рамках моей дипломной работы.

Целью работы является изучение методов моделирования финансовых временных рядов без предположений о лежащей в их основе стохастической динамики.

Авторами была разработана генеративная модель вариационного автокодировщика для финансовых временных рядов, основанная на сигнатурах, которая работает на маленьком количестве данных.

Помимо этого, в статье показывается различие между классическими подходами и подходом машинного обучения. В классических моделях известна стохастическая основа и математические свойства модели, но на практике в общем случае распределение финансовых данных неизвестно. В таких случаях может помочь подход машинного обучения, в случае которого мы генерируем новые данные, похожие по распределению на исходные. Тогда не нужно знать распределение исходных данных, мы генерируем новые и сравниваем их по распределению с исходными, что предлагается также делать с помощью сигнатурных методов.

Поскольку данная модель в качестве входных данных использует логарифмические сигнатуры - то и генерирует она тоже их. Полученные данные можно передать другой модели в качестве входных данных, оставив их в таком виде. Однако, если необходимо получить конкретные пути, в статье предлагается применить эволюционный алгоритм восстановления пути по его логарифмической сигнатуре.

Отметим, что область применимости данного алгоритма ограничивается одномерными путями, что делает невозможным его применение, например, для изучения

статистической зависимости между зависимыми временными рядами. Помимо этого, алгоритм обладал рядом ограничений в настройке различных параметров при работе всех моделей под другие случаи использования, кроме предполагаемого авторами.

Целью моей работы было убрать все эти ограничения, и в качестве результата приводится готовый программный код для генерации многомерных путей, протестированный на реальных ценах активов.

3 Основные результаты

3.1 Слайд 9

Для обращения логарифмической сигнатуры предлагается использовать эволюционный алгоритм подбора нужного пути, чтобы его логарифмическая сигнатура была наиболее похожа на изначальную.

Пусть мы ищем путь X, соответствующий известной нам логарифмической сигнатуре. В общем случае путь не является кусочно-линейным и может принимать любые вещественные значения.

Требуется найти кусочно-линейный путь $\widehat{X}$, который наилучшим образом приближает путь X. Изломы у этого пути могут быть только с фиксированным шагом Δt, а значения приращений каждой из координат пути $\widehat{X_i}$ возможно искать только из такой ограниченной сетки с шагом h_i.

3.2 Слайд 10

При этом при ограничении n_i накладывается ограничение на допустимый модуль производных соответствующих координат пути X_i в зависимости от значения Δt: чем меньше Δt, тем пути с большей производной возможно приблизить при фиксированных h_i и n_i.

Формально это описано в данной теореме.

Если провести достаточное количество итераций генерирования приращений каждой координаты пути $\widehat{X}_i$ из соответствующего арифметического распределения, то с вероятностью, равной единице, в один момент мы получим оптимальный путь

$\widehat{X}$. Однако на практике количество итераций алгоритма ограничено и на каждом шаге выбираются лучшие пути, чьи логарифмические сигнатуры наименее отличаются от исходной, после чего они различными методами размножаются для следующей итерации.

3.3 Слайд 11

Для проверки работы модели обращения логарифмической сигнатуры возьмём любой путь, посчитаем его логарифмическую сигнатуру, передадим в модель обращения и сравним исходный путь и полученный.

На данном рисунке синим цветом изображена совместная траектория цены двух активов, где по оси x и y обозначены их цены, а по оси z - время.

Оранжевым цветом изображен путь, восстановленный по логарифмической сигнатуре синего пути.

3.4 Слайд 12

Запустив по полученным из генеративной модели логарифмическим сигнатурам модель их обращения, получаем следующий набор сгенерированных путей.

На данном рисунке синим цветом изображен изначальный набор совместных траекторий цен акций компании ПАО Газпром и отраслевого индфекса нефти и газа Московской биржи, разделенные по времени на 20-ти дневные промежутки, что соответствует рабочему месяцу.

Красным цветом нарисованы восстановленные алгоритмом пути по сгенерированным моделью логарифмическим сигнатурам.

После того, как мы получили сгенерированные пути, необходимо проверить, что они из того же распределения, что исходный набор. Для этого авторами предлагается использовать тест с метрикой максимального среднего расхождения, использующий сигнатуры путей. Применение данного тест показало, что исходные и полученные данные действительно из одного вероятностного распределения.