МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет компьютерных наук

Кафедра цифровых технологий

Вероятностный подход к обработке последовательных данных

ВКР Бакалаврская работа

02.03.01 Математика и компьютерные науки

Распределенные системы и искусственный интеллект

Допущено к защите в ГЭК __.__.22__

Зав. кафедрой _____ *С. Д. Кургалин, д. ф.-м. н., профессор*
подпись

Обучающийся _____ *А. М. Гузенко, 4 курс, д/о*
подпись

Руководитель _____ *А. Ф. Клинских, д. ф.-м. н, профессор*
подпись *расшифровка подписи, ученая степень, звание, должность*

Воронеж 2022

Содержание

Введение

Применение анализа временных рядов и методов машинного обучения играют огромную роль в жизни человека и делает большой шаг в технологическом прогрессе. Машинное обучение применяется во многих сферах, от сетевых магазинов до промышленных производств. Где-то они носят рекомендательный характер для человека, а в каких-то уже могут заменить человека полностью. В пример можно привести систему автоматического торможения от Volvo, учитывая дистанцию и скорость, при помощи передней камеры, в режиме реального времени рассчитывая тормозной путь, система может принять решение об экстренном торможении, с реакцией, которая не сравнится с человеческой.

Анализ временных рядов и машинное обучение получили свое быстрое развитие в основном по двум причинам: увеличение количества информации для анализа и развитие вычислительных мощностей, но всему есть предел. Если для информации одна из основных проблем это правильный сбор и последующее интерпретирование информации, то для машинного обучения актуален вопрос эффективности алгоритмов. Для решения данной проблемы используются как другие вычислительные модели, например, квантовые вычисления, так и поиск более эффективных алгоритмов для обучения модели. Один из таких алгоритмов является вероятностный подход, или же Байесовский подход, названный в честь Томаса Байеса, который разработал одну из важнейших теорем в теории вероятности, теорему Байеса о зависимых вероятностях.

В последнее время вероятностных подход нашел отклик и стремительно развивается, созданы несколько библиотек для создания моделей, анализа данных. Решаются проблемы эффективности и точности моделей, созданием улучшенных методов для работы с вероятностными распределениями.

Главные задачи в данной работе:

1. Взять временной ряд, в котором требуется решить задачу классификации по одному или нескольким переменным.

2. Реализовать две модели для решения задач классификации.

3. Сравнить обычный подход в решении задачи классификации и вероятностный подход в эффективности на разных выборках.

4. Сравнить скорости обучения моделей на равном объеме данных.

5. Сделать выводы об областях применения данного подхода, его преимуществах и недостатках.

Литературный обзор

В первую очередь стоит отметить, что рассматриваемая область быстроразвивающаяся, и, соответственно, источники для исследования в данной области должны быть как можно более актуальными, но не стоит забывать о качестве этих источников, так как некоторые поспешные выводы могут быть ошибочны, но данный тезис в большей мере касается научных статей, результатов недавно проведенных исследований. Основополагающим для данной работы я выбрал три литературных источника, остальные же либо в какой-то мере дополняют их, либо более углубленно раскрывают некоторые темы.

Для анализа временных рядов за основу была взята книга «Практический анализ временных рядов: Прогнозирование со статистикой и машинное обучение» Эйлин Нильсен 2019 года. Эйлин кандидат прикладной физики Колумбийского университета, занималась исследованиями влияния развития технологий на право и медицину. В данный момент она занимается разработкой нейронной сети для прогнозирования в сфере финансов. В этой книге предлагается практическое знакомство с временными рядами, их анализом, обработкой и последующим применением.

Для Байесовского подхода за основу были взяты книги «Байесовские модели. Байесовская статистика на языке Python» Аллен Б. Дауни 2013 года и «Байесовский анализ на Python: введение в статистическое моделирование и вероятностное программирование с использованием PyMC3 и ArviZ» Освальдо Мартин 2018 года.

Аллен Б. Дауни профессор компьютерных наук в инженерном колледже штата Массачусетс, выпустил множество книг по обработке данных и программированию в целом. Книга по Байесовской статистике является руководством к применению языка программирования Python для данной сферы, без углубления в математические основы.

Освальдо Мартин ученый-исследователь агентства National Scientific and Technical Research Council (CONICET) в Аргентине, работающий в области структурной биоинформатики, является одним из основных разработчиков библиотек PyMC3 и ArviZ, преподаватель курса по Байесовскому анализу данных. Книга представляет собой введение в Байесовский анализ, излагается методический подход моделирования вероятностных моделей, применение теоремы Байеса для вывода логических следствий из используемых моделей и данных.

Дополнительно использовались следующие источники.

Для понимания математических процессов, то есть теории вероятности и математической статистики, идеально подойдут работы двух советских математиков, а именно «Курс теории вероятностей» Бориса Владимировича Гнеденко и «Теория вероятностей и математическая статистика» Владимира Ефимовича Гмурмана. Данные две книги дают систематические знания по теории вероятностей и математической статистике, предоставляют разбор реальных примеров и задач для самостоятельного решения.

Для понимания общего устройства и работы нейросетей подойдут некоторые разделы следующих книг «Глубокое обучение с точки зрения практика» Джош Паттерсон, «Pattern Recognition and Machine Learning» Christopher Michael Bishop, «Building Machine Learning and Deep Learning Models on Google Cloud Platform» Ekaba Bisong, «Глубокое обучение» Гудфеллоу Ян, «Прикладное глубокое обучение: подход к пониманию глубоких нейронных сетей на основе метода кейсов» Микелуччи Умберто.

Для лучшего математического понимания Байесовской статистики была использована книга «Bayesian Data Analysis» Andrew Gelman. Так же, данный автор описал в научном журнале алгоритм, который мы будем рассматривать в дальнейшем.

Для рассмотрения нового алгоритма Automatic Differentiation Variational Inference взята статья Alp Kucukelbir из научного журнал

Глава 1. Последовательные данные

1. 1 Понятие последовательных данных

Последовательные данные (или временные ряды) – это серия точек данных, проиндексированных во временном или ином порядке. Чаще всего временной ряд представляет собой множество, полученное в последовательных равностоящих точках времени.

Данные временных рядов и их анализ приобретают все большее значение из-за получения таких данных посредством, например, интернет вещей, цифровизации здравоохранения, развития умных городов и так далее. По мере того, как непрерывный мониторинг и сбор данных становятся все более распространёнными, потребность в анализе временных рядов с использованием как статистических методов, так и методов машинного обучения, повысится. Анализ временных рядов часто сводится к вопросу о причинно-следственной связи, как прошлое повлияло на будущее.

Практическое применение анализа временных рядов и машинного обучения еще в 1980-ых годах включал широкий спектр сценариев:

- Специалисты по компьютерной безопасности задействовали их для выявлений аномалий в качестве метода идентификации против взломов и вмешательств.
- Динамическое изменение масштаба времени, один из доминирующих методов измерения сходства между временными рядами. С увеличением вычислительной мощности стало возможно достаточно быстро вычислять «расстояние» между, например, аудиозаписями.
- Изобретены рекурсивные нейронные сети, которые показали свою эффективность для извлечения шаблонов из поврежденных данных.

Для анализа временных рядов используются различные графические отображения данных, такие как графики, сводные статистики, гистограммы, диаграммы рассеяния. Рассмотрим некоторые из них.

Во временных рядах, как и в данных, можно произвести операцию группировки, то есть выделить отдельные группы значений. Например, в перекрестных данных операции группировки позволяют определить средние значения для значений возраста, пола или места проживания респондентов. В анализе временных рядов так же востребованы групповые операции, позволяющие рассчитывать статистические показатели наборов, например, средимесячные или недельные медианы. Данные одного временного ряда становятся более понятными при разделении на несколько параллельных временных рядов.

Возьмем сведения о ежедневных ценах закрытия четырех основных европейских фондовых индексов, указанных с 1991 по 1998 годах, и отобразим их на графике.

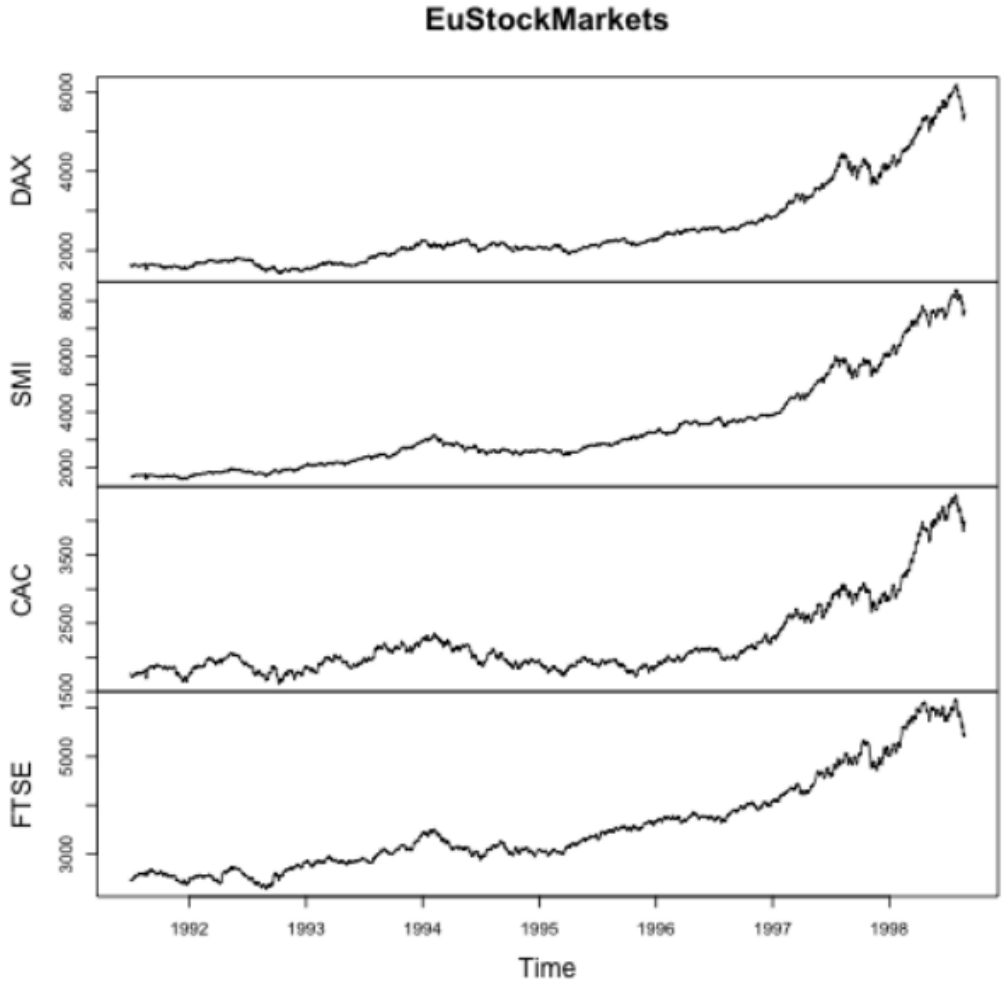

Рисунок 1 - График временного ряда

При помощи гистограммы мы можем оценить частоту появления данных, или, что более важно, проанализировать изменение значения во времени.

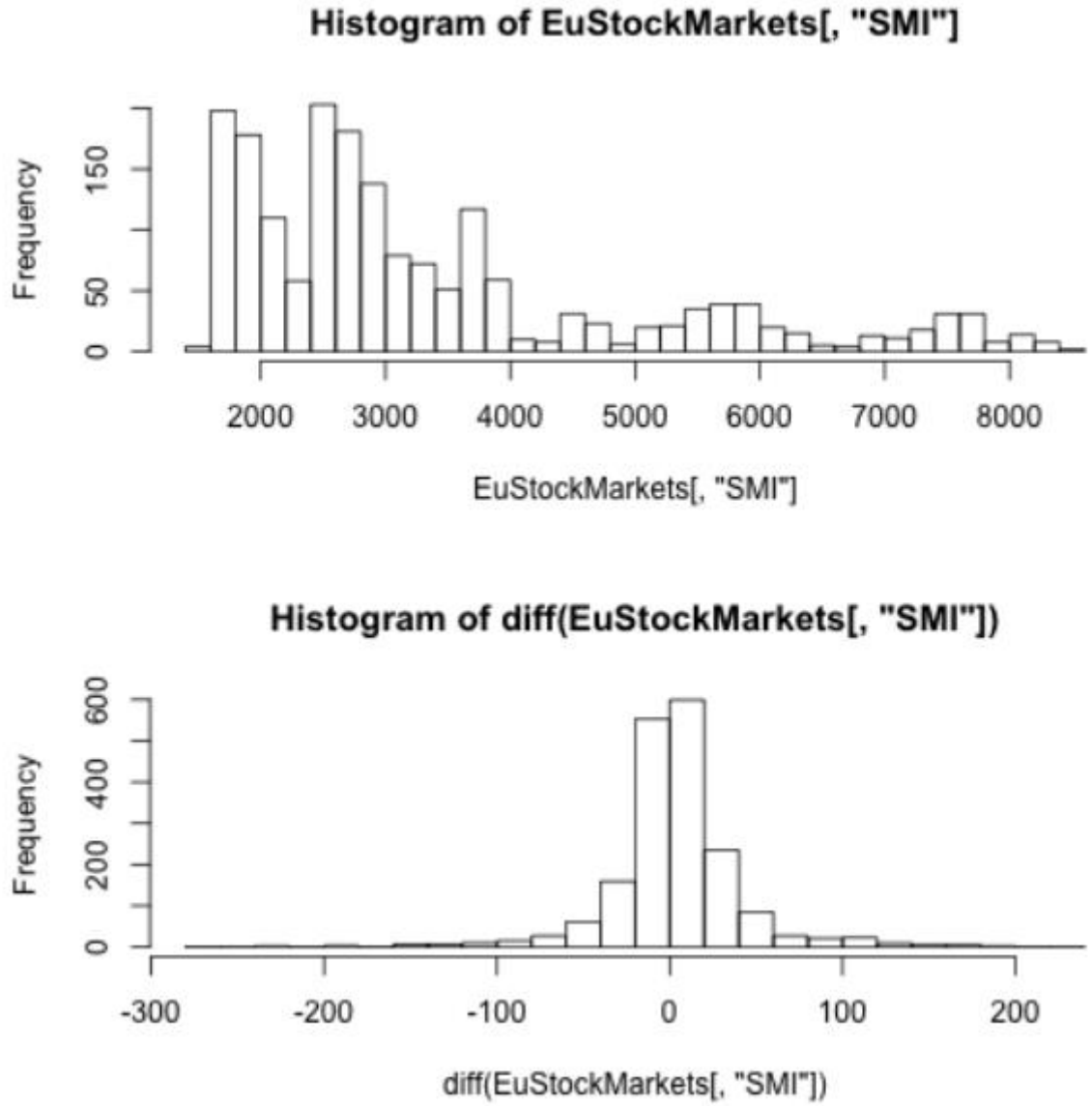

Рисунок 2 - Гистограмма временного ряда. Первая гистограмма – частота значений, вторая гистограмма – частота разностей

Диаграммы рассеяния мы можем использовать для определения взаимосвязи между ценами двух акций в отдельные моменты времени, а также отслеживания их временных изменений. На следующем рисунке присутствуют оба варианта.

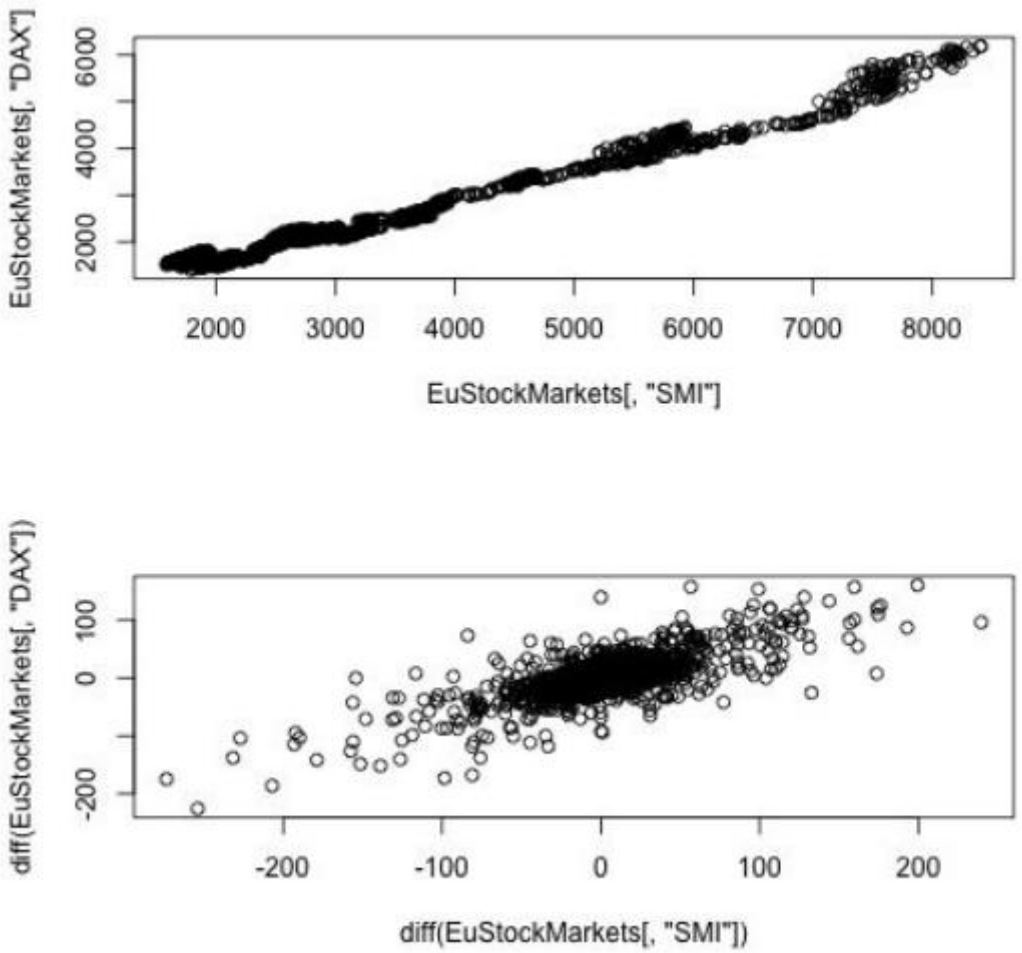

Рисунок 3 – Диаграмма рассеяния. На первой диаграмме – цена двух акций с течением времени, на второй – разница цен двух акций с течением времени

Помимо графического анализа стоит обращать внимание на некоторые характеристики временных рядов, такие как стационарность, корреляция, ложная корреляция.

Один из важных вопросов при анализе временного ряда – это какую систему он описывает, стабильную или изменчивую? Уровень стабильности, или стационарности, важен для оценки того, как долгосрочное поведение в прошлом отражает ее поведение в будущем. После оценки уровня стабильности мы пытаемся определить, присуще ли ему динамическое поведение. Простое определение стационарности процесса заключается в следующем: процесс считается стационарным, если для всех возможных смещений k распределение $y_t, y_{t+1}, \dots, y_{t+k}$ не зависит от t.

Самокорреляция – значение временного ряда в отдельные моменты времени могут коррелировать с его значениями в другие моменты времени.

Автокорреляция – общий случай самокорреляции, лишенному привязки к конкретному моменту времени. Автокорреляция сводится к решению задачи поиска взаимосвязи между любыми двумя точками общего временного ряда, расположенными на строго заданном расстоянии друг от друга, то есть автокорреляция дает представление о линейной взаимосвязи точек данных, полученных в разные моменты времени, как о функции разницы времени их получения.

Ложные корреляции – математическая зависимость, в которой переменные или события связаны, но в следствии совпадения. Ложные корреляции остаются важной проблемой, требующей самого пристального изучения и опровержения.

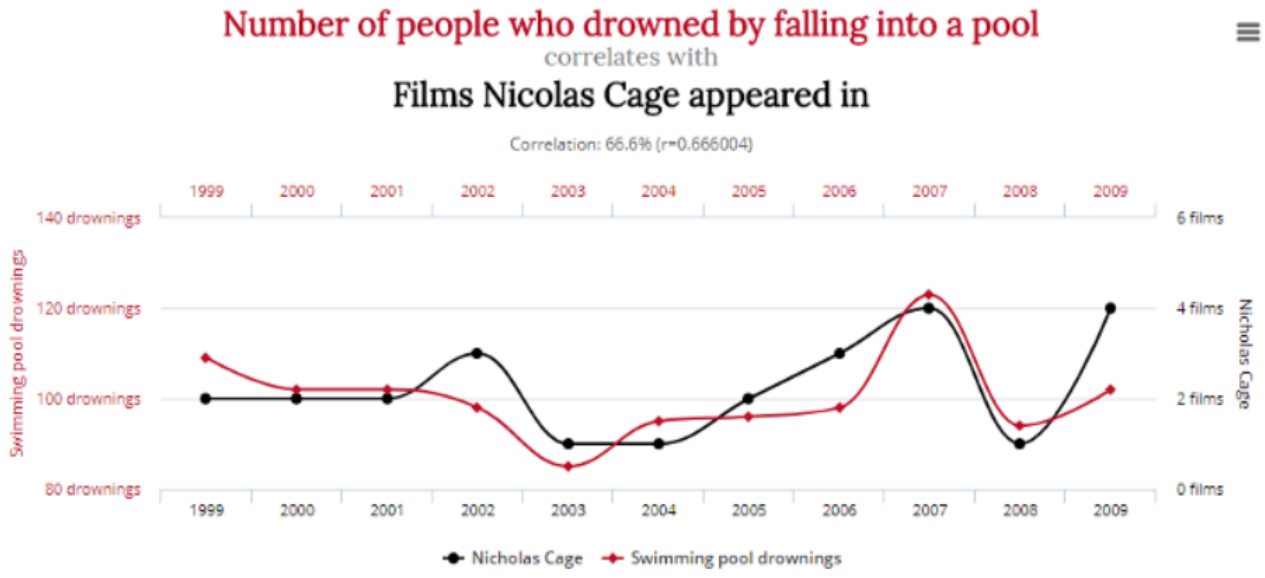

Рисунок 4 – График выхода фильмов с Николасом Кейджем и количество утонувших людей в бассейне

Моделирование данных выступает разновидностью анализа данных, который находит широкое применение при работе с временными рядами. Это следует из одного недостатка временных рядов: никакие две точки данных в одном и том же временном ряду не могут быть точно сопоставимы, поскольку они относятся к разному времени. Как только мы задумаемся о

том, что могло бы произойти в данный момент времени, мы придём к моделированию.

Моделирование временных рядов методами глубокого обучения – это новая, но весьма многообещающая дисциплина науки о данных. С помощью технологии глубокого обучения во временных рядах можно обнаружить сложные динамические процессы, зачастую скрытые для понимания даже опытных специалистов по анализу данных.

Моделирование и прогнозирование – схожие задачи. В обоих случаях сначала нужно сформулировать гипотезу о параметрах и поведении базовой системы, а затем экстраполировать имеющиеся данные для получения новых точек.

Но нужно четко понимать различия между моделированием и прогнозированием:

- Иногда качественные наблюдения проще обрабатывать методами моделирования, а не прогнозирования.
- Моделирование выполняется в определённом масштабе, что позволяет увидеть множество альтернативных сценариев, в то время как прогнозы составляются предельно точно.

1. 2 Вероятностный подход к анализу последовательных данных.

Для начала стоит напомнить, что такое вероятность в общем понимании. Вероятность – это число между 0 и 1, которые представляют собой уровень уверенности, что некоторый факт справедлив. Числом 1 представляется абсолютная уверенность, что некоторый факт справедлив. Числом 0 – абсолютная уверенность, что этот факт не справедлив. Промежуточные числа в этом интервале определяют степень уверенности. Число 0,5 означает, что предсказанное событие в одинаковой степень может как осуществиться, так и не осуществиться.

Распределение вероятностей – это математический объект, который описывает, насколько возможными являются различные события. Эти

события ограничены каким-либо образом, то есть представляют собой набор возможных событий, например набор возможных чисел $\{1, 2, 3, 4, 5, 6\}$ для игральных костей (исключая неопределенные случаи). Общепринятым и полезным представлением концепции в статистике является интерпретация данных как генерируемых из некоторого истинного распределения вероятностей с неизвестными параметрами. То есть нам нужно найти значения этих параметров с использованием только частичной выборки из истинного распределения вероятностей. В обобщенном случае у нас нет доступа к такому истинному распределению вероятностей, поэтому необходимо создать модель, чтобы попытаться аппроксимировать это распределение. Вероятностные модели создаются при помощи правильного объединения распределений вероятностей.

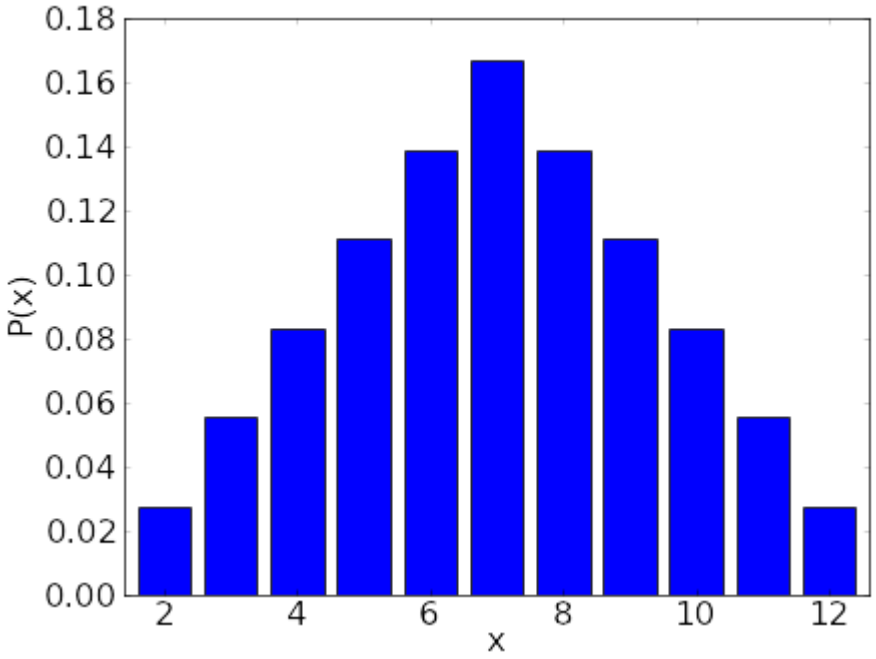

Рисунок 5 – Вероятностное распределение броска двух игральных костей

Для построения сложных моделей активно используется теорема Байеса

$$p(\theta, y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.1}$$

Если заменить элемент θ на «предположение» (гипотезу), а элемент y на «данные», то теорема Байеса показывает, как вычислить вероятность предположения θ при наличии данных y. Для того, чтобы превратить предположение в некоторый объект, нужно использовать вероятностные распределения. В вероятностном подходе для анализа данных мы часто будем использовать теорему Байеса. Ниже приведем названия для элементов данной теоремы:

- $p(\theta)$ – априорная вероятность
- $p(y|\theta)$ – правдоподобие
- $p(\theta|y)$ – апостериорная вероятность
- $p(y)$ – предельное правдоподобие

Априорная вероятность должна соответствовать тому, что нам известно о значении параметра θ перед рассмотрением данных y. Если нам ничего не известно, то можно использовать постоянные фиксированные априорные вероятности, которые не содержат какого-либо значимого объема информации.

Правдоподобие определяет, как будут представлены данные в дальнейшем анализе. Это выражение правдоподобности данных с учетом принятых параметров.

Апостериорная вероятность – это результат байесовского анализа, которые отображает все, что известно о задаче (проблеме) с учетом имеющихся данных и используемой модели. Апостериорная вероятность – это распределение вероятностей для параметра θ в используемой модели. Такое распределение это баланс между априорной вероятностью и правдоподобием. С теоретической концептуальной точки зрения апостериорную вероятность можно воспринимать как обновленную вероятность в свете новых данных. Апостериорная вероятность, полученная в результате одного процесса анализа, может использоваться как априорная вероятность для нового процесса анализа. Это свойство делает байесовский

анализ особенно подходящим для анализа данных, которые становятся доступными в определенном последовательном порядке. Примерами могут служить системы раннего оповещения о природных катастрофах, которые обрабатывают в режиме онлайн данные, поступающие с метеорологических станций и спутников.

Правдоподобие – это вероятность исследуемых данных, усредненная по всем возможным значениям, которые могут принимать параметры. В любом случае мы не уделяем особого внимания предельному правдоподобию и будем считать его простым фактором нормализации. Такой подход принят потому, что при анализе распределения апостериорной вероятности нас будут интересовать только относительные, а не абсолютные значения параметров.

$$p(\theta, y) \propto p(y|\theta)p(\theta) \tag{1.2}$$

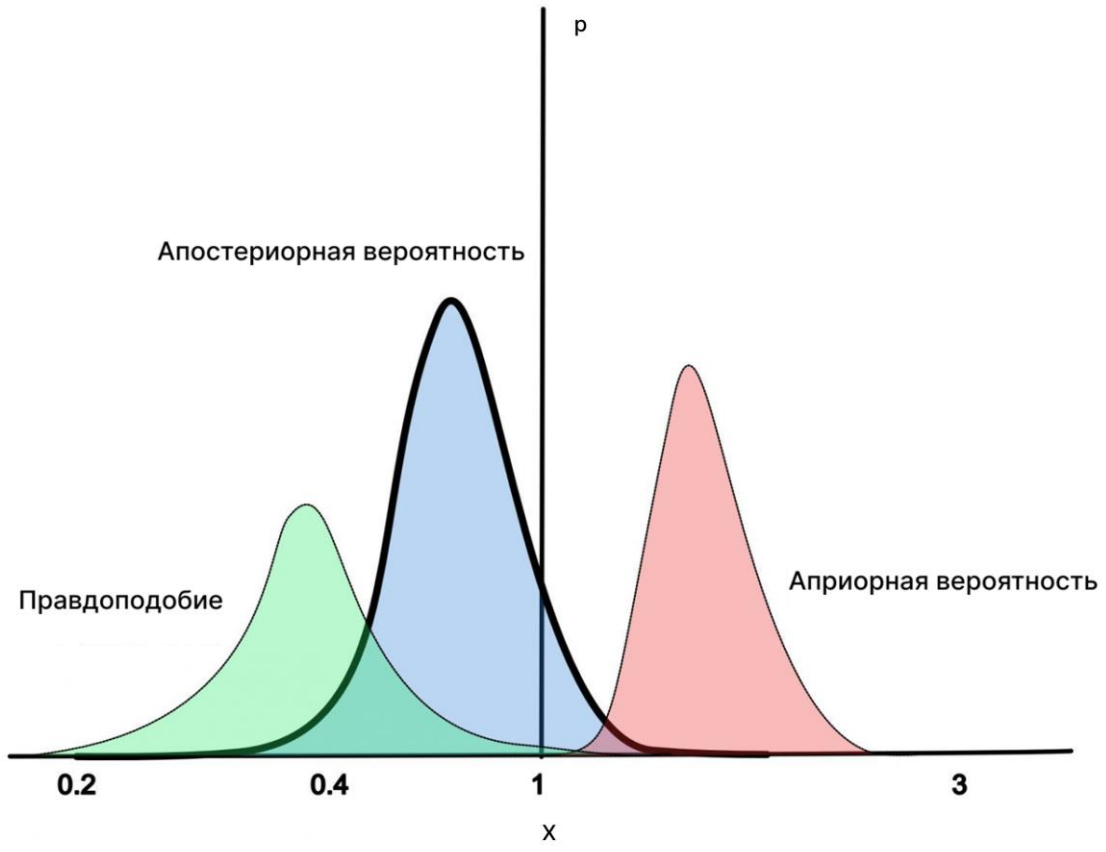

Рисунок 6 – Графический пример работы Байесовской теоремы

Список использованных источников

1. Мартин, Освальдо. Байесовский анализ на Python : введение в стати стическое моделирование и вероятностное программирование с использованием PyMC3 и ArviZ / Освальдо Мартин ; перевод с английского А. В. Снастина. – Москва : ДМК Пресс, 2020. - 339 с.

2. Дауни, Аллен Б. Байесовские модели : байесовская статистика на яз ыке программирования Python / Аллен Б. Дауни ; перевод с английского В. А. Яроцкого. - Москва : ДМК Пресс, 2018. - 181 с.

3. Нильсен, Эйлин. Практический анализ временных рядов : прогнози рование со статистикой и машинное обучение: перевод с английского / Нильсен, Эйлин. – Москва : Диалектика ; Санкт-Петербург Диалектика, 2021. - 538 с.

4. Гнеденко, Борис Владимирович. Курс теории вероятностей : [Учеб. Для мат. спец. ун-тов] / Б. В. Гнеденко. - 6-е изд., перераб. и доп. - М. : Наука, 1988. - 446 с.

5. Гмурман, Владимир Ефимович. Теория вероятностей и математическая статистика : учебное пособие для студентов вузов / В. Е. Гмурман. - 12-е изд. / перераб. - Москва : Юрайт : Высш. образование, 2009. - 478 с.

6. Паттерсон, Джош. Глубокое обучение с точки зрения практика / Джош Паттерсон, Адам Гибсон ; пер. с англ. А. А. Слинкина. - Москва : ДМК Пресс, 2018. - 415 с.

7. Christopher M. Bishop. Pattern Recognition and Machine Learning / Christopher Michael Bishop - New York : Springer New York, cop. 2006. - 738 p.

8. Ekaba Bisong. Building Machine Learning and Deep Learning Models on Google Cloud Platform / Ekaba Bisong - Berkeley, CA : Apress, cop. 2019. – 709 с.

9. Гудфеллоу Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; [пер. с англ. А. А. Слинкина]. - 2-е цв. изд., испр. - Москва : ДМК Пресс, 2018. - 651 с.

10. Микелуччи У. Прикладное глубокое обучение : подход к пониманию глубоких нсейронных сетей на основе метода кейсов / У. Микелуччи ; перевод с английского Андрея Логунова. - Санкт-Петербург : БХВ-Петербург, 2020. - 368 с.

11. Bayesian Data Analysis / Andrew Gelman, John B. Carlin, Hal S. Stern and etc.- third edition - Oxford, United States : CRC Press, cop. 2013. – 675 c.

12. Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo / Andrew Gelman // Journal of Machine Learning Research. – 2014. – №15. – С. 1593-1623.

13. Alp Kucukelbir. Automatic Differentiation Variational Inference / Alp Kucukelbir // Journal of Machine Learning Research. – 2017. - №18. – С. 1-45