**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики и информатики**

**Кафедра математического моделирования и анализа данных**

Курсовая работа

**Разработка байесовских авторегрессионных моделей для прогнозирования стационарных временных рядов**

Карповича Артёма Дмитриевича

студента 3 курса, специальность

«Прикладная математика»

Научный руководитель

Лобач Виктор Иванович

доцент кафедры ММАД

кандидат физ.-мат. наук

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра математического моделирования и анализа данных

# ЗАДАНИЕ НА КУРСОВОЙ ПРОЕКТ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | | Карпович Артём Дмитриевич | | | |
| 1. Тема | | Разработка байесовских авторегрессионных моделей для прогнозирования стационарных временных рядов | | | |
| 2. Срок представления курсового проекта к защите 15 декабря 2023 г. | | | | | |
| 3. Исходные данные для научного проектирования  3.1. Кувайскова, Ю. Е Статистические методы прогнозирования / Ю. Е. Кувайскова, В. Н. Клячкин/ – Минск: УГТУ, 2019. | | | | | |
| 3.2. Ивченко, Г. И. Математическая статистика: Учеб. Пособие для втузов / Ивченко Г. И., Медведев Ю. И./ – Москва: МВТУ им. Н. Э. Баумана, 1984. | | | | | |
| 3.3. Грэйнджер, Клайв У. Дж. Эконометрический анализ временных рядов / Клайв У. Дж. Грэйнджер. | | | | | |
| 3.4. Абрамова, А. Е. Метод Монте Карло по схеме марковской цепи для оценки вероятности редких событий в задачах биоинформатики / Абрамова А. Н./ – Санкт-Петербург: СПБГУ, 2017. | | | | | |
| 3.5. Четыркин, Е. М. Статистические методы прогнозирования / Четыркин Е. М. // издание «Статистика». – 1975. | | | | | |
| 3.6. Hamilton, J. D. Time Series Analysis / James D. Hamilton/ – Princeton, New Jersey: Princeton University, 1994. | | | | | |
|  | | | | | |
| 4. Содержание курсового проекта | | | | | |
| 4.1. | Подготовить обзор по теме «разработка байесовских авторегрессионных моделей для прогнозирования стационарных временных рядов». | | | | |
| 4.2. | Подготовить математическое описание байесовской модели машинного обучения | | | | |
| 4.3. | Привести пример использования теоремы Байеса | | | | |
| 4.4. | Подготовить отчет по курсовому проекту. | | | | |
| Руководитель курсового проекта | | |  | Лобач В.И. |
|  | | | *подпись, дата* | *фамилия, инициалы* |
| Задание принял к выполнению | | | Карпович А.Д. | |

*подпись, дата фамилия, инициалы*

Оглавление

[Введение 3](#_Toc153406632)

[ГЛАВА 1. Применение машинного обучения для прогнозирования 6](#_Toc153406633)

[ГЛАВА 2. Статистические методы 11](#_Toc153406634)

[2.1. Метод Байеса 11](#_Toc153406635)

[2.2. Дискриминантный анализ 13](#_Toc153406636)

[2.3. Логистическая регрессия 15](#_Toc153406637)

[ГЛАВА 3. Применение теоремы Байеса в машинном обучении. 18](#_Toc153406638)

[3.1 Метод MAP 18](#_Toc153406639)

[3.2. МСМС 19](#_Toc153406640)

[3.2.1.Cхема Метрополиса-Хастинга 20](#_Toc153406641)

[3.2.2.Схема Гиббса 21](#_Toc153406642)

[Приложение 1 24](#_Toc153406643)

[СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ 26](#_Toc153406644)

# Введение

В современном мире большое количество явлений можно представить в виде последовательности наблюдений, измеренных в разные моменты времени и упорядоченных в хронологическом порядке, такие последовательности называются временными рядами. Прогнозирование временных рядов играет важную роль во многих областях, включая финансы, экономику, климатологию, медицину и другие. Эффективный анализ временных рядов является ключевым инструментом для принятия обоснованных решений в различных сферах деятельности.

Анализ временных рядов включает в себя исследование статистических свойств ряда, выявление трендов, сезонности и других особенностей, а также построение моделей для прогнозирования будущих значений и выявления структуры временного ряда. Выявление структуры временного ряда необходимо для того, чтобы построить математическую модель того явления, которое является источником анализируемого временного ряда. Прогноз будущих значений временного ряда используется для эффективного принятия решений.

Временные ряды состоят из двух элементов:

* периода времени, за который или по состоянию на который приводятся числовые значения;
* числовых значений того или иного показателя, называемых уровнями ряда.

Классифицируются же временные ряды по следующим признакам:

* по форме представления уровней: ряды абсолютных показателей; относительных показателей; средних величин;
* по количеству показателей, для которых определяются уровни в каждый момент времени: одномерные и многомерные временные ряды;
* по характеру временного параметра: моментные и интервальные временные ряды;
* по расстоянию между датами и интервалами времени: равноотстоящие и неравноотстоящие (неполные);
* по наличию пропущенных значений: полные и неполные временные ряды;
* по принципу получения временного ряда: детерминированные и случайные;
* по наличию основной тенденции: стационарные и нестационарные;

На стационарных временных рядах остановимся поподробней. Стационарные временной ряд – это ряд, чьи статистические свойства не меняются со временем. Формально, временной ряд считается стационарным, если выполнены следующие свойства:

* стационарность по среднему: среднее значение ряда не зависит от времени и остается постоянным на протяжении всего ряда;
* стационарность по дисперсии: дисперсия ряда не зависит от времени и остается постоянной на протяжении всего ряда;
* стационарность по автоковариации: автоковариация между значениями ряда на разных временных отрезках зависит только от длины этих отрезках, но не от их положения во времени.

Стационарность является важным свойством при анализе и моделировании, так как она позволяет применять статистические методы и модели, которые предполагают постоянство статистических свойств данных. В стационарном ряде проще выявить закономерности, построить модели и делать прогнозы на основе его статистических свойств, чем в нестационарном ряде, где эти свойства могут меняться со временем.

Основными моделями для работы со стационарными рядами можно упомянуть следующие виды: одномерные, линейные и гомоскедстичные (то есть модели, характеризующиеся постоянством дисперсий остатков). Основной же «рабочей» моделей была авторегрессионная интегрированная модель скользящей средней (ARMA(p, q)), в которой текущее значение исследуемой переменной объясняется ее *p* лаговым значениями и имеет компонент в форме скользящей средней порядка *q*; в виде формул все это описывается следующим образом:

Главным строительным блоком этих моделей является ряд белого шума с нулевой средней, так что ковариация () = 0, *s ≠ t.* Используя терминологию анализа временных рядов, можно сказать, что такой анализ включает три шага:

1. Идентификация, то есть подбор пригодной модели, что означает отбор «подходящих» *p* и *q;*
2. Получение оценки, которая заключается в оценивании других параметров модели, то есть значений *a* и *b*, а также дисперсии (причем предполагается, что она представляет собой константу);
3. Определения качества построенной модели.

Несмотря на совершенствование моделей и методов, эти три шага остаются важными компонентами эмпирического анализа. Однако теперь гораздо больший интерес проявляется к многомерным нелинейным гетероскедастичным моделям, пригодным для работы с нестационарными рядами, не имеющими в то же время «взрывного» характера.

# ГЛАВА 1. Применение машинного обучения для прогнозирования

*Машинное обучение* – это компьютерные методы выявления закономерностей в данных большого объема (Big Data).

Одной из задач, которую можно решить методами машинного обучения является задача *бинарной классификации*. Например, рассматривается технический объект, о котором известно, что при одном определенном наборе параметров функционирования объект был исправен, при другом – неисправен. Известны новые параметры функционирования. Спрогнозировать, будет ли объект исправен? Это и есть *задача бинарной классификации*, ориентированная на предсказание состояния объекта – прогнозирование.

Другой класс задач прогнозирования, который решается с помощью методов машинного обучения – *задачи регрессии*. Пусть, например, мы планируем открыть новый магазин. Имеющийся набор данных – это информация о характеристиках магазинов нужного профиля, работающих в рассматриваемом районе (место расположения, площадь, товарооборот и т.п.). Известна и прибыль магазинов за прошедший год. По этим данным можно построить модель для оценки предполагаемой прибыли в зависимости от характеристик магазина. Новый магазин можно открыть в нескольких местах с заданными характеристиками. Какое из них обеспечит максимальную прибыль?

Рассмотренные примеры относятся к классу задач машинного обучения с учителем, или по прецедентам: прецеденты – это известные данные о функционировании технических объектов, или о работе магазинов и т. п. В качестве исходных данных используется множество параметров - признаков объекта *Х*:

или ,

где – результат *i*-го наблюдения по *j*-му параметру-признаку; *i* = 1, …, l, j = *1, …, p*, *l* – количество строк – число наблюдений: сколько объектов исследовано, *p* – количество столбцов, или число признаков, например, параметров функционирования объекта; – вектор значений *j*-го параметра-признака (значения признаков могут быть как количественными, так и бинарными, номинальными, порядковыми), и вектор-столбец ответов *Y*, состоящий для задач бинарной классификации из *1* (для тех опытов, в которых объект исправен) и *0* при неисправном объекте.

В задачах регрессии, в частности, в задаче об открытии нового магазина, вектор-столбец ответов *Y* состоит из значений прибыли каждого из работающих магазинов.

Каждой строке множества *Х* соответствует определенное значение вектора *Y*. Совокупность пар () образует выборку

*исходных данных – прецедентов*.

Задача состоит в построении функции , которая предскажет ответ *Y* для любого заданного *Х*. Таким образом, функция *а* (ее называют еще и алгоритмом или моделью алгоритма) должна не только приближать искомый результат на выборке исходных данных, но и «работать» на всей генеральной совокупности, из которой получено множество Х. При числовых значениях признаков часто используют линейные модели с вектором параметров :

При этом в задачах бинарной классификации обычно вместо нуля и единицы используют множество ответов *Y* = {–1; +1}. В этом случае модель алгоритма примет вид

Параметры подбираются по исходным данным. Процесс подбора оптимальных параметров называется *обучением алгоритма*. Найденные параметры должны обеспечить оптимальное значение функционала качества.

Вводится *функция потерь L(a, X)*, характеризующая ошибку алгоритма *а* на объекте *X*. При нулевом значении этой функции ответ считается корректным. В задачах бинарной классификации в качестве функции потерь можно использовать *индикатор ошибки*:

Функционал качества алгоритма *а* на выборке объема *l* (среднее

значение потерь) называют эмпирическим риском:

В задаче бинарной классификации минимизируется функционал ошибок:

В задаче регрессии минимизируется среднеквадратичная ошибка:

Алгоритм *a*, минимизирующий функционалы (1.5) или (1.6), может не обеспечивать хорошее приближение на произвольной выборке из генеральной совокупности. Ситуация, когда качество работы алгоритма на новых объектах значительно хуже, чем на исходной выборке, свидетельствует о переобучении: алгоритм слишком хорошо подогнан под обучающую выборку и не способен к обобщению на другие выборки. Таким образом, построенный алгоритм не сможет предсказывать состояние исследуемого объекта при новых параметрах функционирования.

Для оценки качества модели с точки зрения возможности прогнозирования исходную обучающую выборку из l опытов разбивают на два непересекающихся подмножества: собственно обучающую выборку объема (с помощью которой и решается задача обучения (1.5) – (1.6)) и контрольную (или тестовую) объема , не используемую для обучения.

При использовании кросс-валидации выборка разбивается на *N*

частей (на практике обычно принимают *N = 5* или *N = 10*). (*N – 1*) часть используется для обучения, одна – для контроля. Последовательно перебираются все варианты, например, при разбиении на пять частей вначале в качестве обучающей выборки используются части 1 – 4, а часть 5 – тестирования.

На следующем шаге в качестве обучающей используются части 2 – 5, а часть 1 – для тестирования, и т.д. Для каждого разбиения решается задача обучения по выборке и вычисляется функция ошибок *Q(a, X)* на контрольной выборке . Среднее значение этой функции по всем вариантам разбиения и характеризует обобщающую способность алгоритма.

Для построения качественных моделей необходим предварительный анализ исходных данных: проверка значимости признаков, исключение выбросов, иногда необходима нормировка(стандартизация). При большом числе признаков целесообразно сократить размерность, используя, например, метод главных компонент.

Таким образом, основные этапы машинного обучения таковы:

1. Постановка задачи.

2. Выделение признаков (параметров функционирования), оказывающих влияние на состояние объекта.

3. Формирование выборки исходных данных и способа ее разбиения на обучающую и контрольную части.

4. Выбор функционала качества.

5. Предварительная обработка данных – отбор значимых признаков, обработка выбросов и пропусков, проведение стандартизации.

6. Построение модели по обучающей выборке.

7. Оценка качества модели по тестовой выборке.

8. Использование модели для прогнозирования состояния объекта по известным параметрам функционирования.

# ГЛАВА 2. Статистические методы

## **2.1. Метод Байеса**

Этот метод базируется на известной из теории вероятностей *формуле Байеса*. Предположим, что событие *А* может произойти лишь при появлении одного из несовместных событий . Вероятности этих событий *Р*(*Нk*) известны и в сумме равны единице: , известны также условные вероятности *Р*(*А*/*Нk*) события *А*.

Тогда вероятность события *А* определяется по *формуле полной вероятности*:

События *Нk* можно рассматривать как гипотезы, тогда *Р*(*Нk*) называют *априорными,* или доопытными, вероятностями гипотез.

Пусть событие *А* произошло, тогда априорные вероятности могут измениться. *Апостериорные,* или послеопытные, вероятности *Р*(*Нk/А*) вычисляются по *формуле Байеса*:

Эта формула, в частности, используется и в задачах технической диагностики, когда априорные вероятности гипотез о причинах появления неисправности переоцениваются после поступления дополнительной информации.

Предположим, что состояние технического объекта характеризуется совокупностью параметров (показателей функционирования) , образующих p-мерный вектор . Чаще всего параметры имеют непрерывные распределения. Иногда их удобно представить в виде комплекса дискретных параметров Например, непрерывный признак «температура» можно представить в виде трехразрядного дискретного признака (пониженная, нормальная, повышенная). При этом состоянии объекта известны ().

Обобщенная формула Байеса при обследовании объекта по комплексу признаков в соответствии с (2.2) примет вид

где предварительная (априорная) вероятность : если у предварительно обследованных *N* объектов диагноз оказался у объектов, то ;

вероятность появления комплекса признаков K у объекта с состоянием (при решении практических задач даже при наличии значимых корреляционных связей между признаками часто принимается условие их независимости); если среди объектов с диагнозом признак проявился у , то соответствующая вероятность .

Для удобства расчетов по формуле (2.3) на основе статистических данных строится диагностическая матрица, каждая строка которой содержит вероятности возможных разрядов признаков ( количество разрядов дискретного признака ) при различных диагнозах . Формирование диагностической матрицы из априорных вероятностей диагнозов, по существу, представляет процесс обучения в этом методе.

При использовании метода Байеса объект с комплексом признаков K относится к диагнозу с наибольшей апостериорной вероятностью, если

количество возможных состояний объекта (диагнозов).

На практике обычно вводится пороговое значение для вероятности диагноза:

где заранее выбранный уровень распознавания.

Например, принимают тогда при решение о диагнозе не принимается: требуется дополнительная информация.

Недостатки метода Байеса связаны с большим объемом предварительной информации и значительными погрешностями при распознавании редких диагнозов. Однако при достаточно большом объеме статистических данных этот метод считается одним из самых надежных

В методах машинного обучения метод Байеса называют наивным байесовским классификатором. «Наивность» состоит в том, что предположение о статистической независимости признаков на практике выполняется достаточно редко. В тех же случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален.

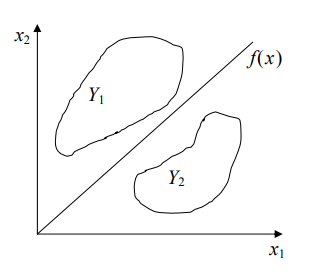
Пример использования приведен в Приложении 1.

## **2.2. Дискриминантный анализ**

Пусть, как и ранее, состояние технического объекта характеризуется совокупностью параметров (показателей функционирования) *xj*, образующих *р*-мерный вектор . При этом состояния объекта известны Заданы показатели функционирования нового объекта.

*Дискриминантный анализ* – это метод классификации с обучением, при этом в качестве обучающих выборок, как правило, используются выборки объектов из нормально распределенных генеральных совокупностей – классов. Решение принимается, исходя из минимума средних потерь от ошибочной классификации, если они могут быть оценены, или из условия минимума вероятности ложной классификации. Таким образом, в частности, технические объекты могут быть разделены на исправные и неисправные.

Рассмотрим множество объектов, которые характеризуются двумя переменными *х*1 и *х*2, эти объекты разделены на два класса (состояния) *Y*1 и *Y*2 (рис. 2.1). Необходимо разделить эти классы некоторой функцией *f*(*x*), которая называется *каноничекой дискриминантной функцией.* На рисунке 2.1 показана линейная функция (линейный дискриминант Фишера):



**Рис. 2.1.** Линейный дискриминант

В общем случае функция может быть и нелинейной. Задача сводится к оценке коэффициентов дискриминантной функции *b*1 и *b*2.

В общем случае параметрического линейного дискриминантного анализа предполагается, что имеется *q* классов объекта .

Каждый класс имеет *р*-мерное нормальное распределение с математическим ожиданием и ковариационной матрицей . Задана обучающая выборка элементов каждого класс где количество объектов внутри *m*-го класса, или объем *m*-й выборки), *j=1,2,…,p* (количество переменных, характеризующий каждый объект).

Рассмотрим случай, когда ковариационные матрицы для различных классов одинаковы. В этом случае алгоритм дискриминантного анализа при количестве классов *q* = 2 (*m* = 1 – объект исправен, *m* = 2 – объект неисправен) следующий.

1. Вычисляем оценки математических ожиданий для каждого класса:

и формируем векторы:

1. Вычисляем () и 0,5().
2. Находим оценки элементов ковариационной матрицы :
3. Находим обратную ковариационную матрицу .
4. Определим вектор коэффициентов линейной дискриминантной функции

(предполагается, что новое наблюдение относится к классу ):

1. Вычисляем величину .
2. Принимаем решение: если выполняется условие

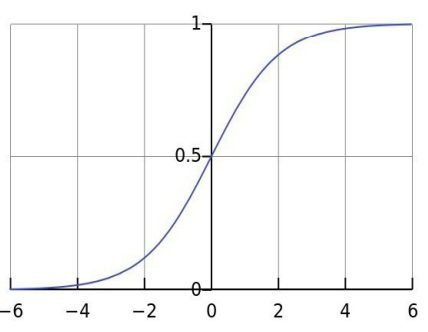
то объект относится к классу 1 (исправен), в противном случае – к классу 2 (неисправен).

## **2.3. Логистическая регрессия**

Пусть *f*(*z*) – это вероятность некоторого события. Тогда значения *f*(*z*) должны лежать в диапазоне от 0 до 1. Один из широко распространенных вариантов в этой ситуации – использование *логистической регрессии.*

Логистическая (сигмоидная) функция имеет вид

Ее график показан на рисунке 2.2, значения функции лежат в диапазоне от нуля до 1



**Рис. 2.2.** Логистическая функция

Соответствующая модель множественной регрессии

Ее можно линеаризовать, используя преобразование

тогда получим стандартную линейную модель множественной регрессии.

Однако в процессе преобразования нарушаются основные предпосылки метода наименьших квадратов, поэтому на практике такой подход не используется. Для оценки параметров логистической регрессии обычно применяют метод максимального правдоподобия. Логарифм функции правдоподобия максимизируется, например, методом градиентного спуска.

Для использования метода градиентного спуска при оценке параметров логистической регрессии с помощью метода максимального правдоподобия, следуйте следующим шагам:

1. Определите функцию правдоподобия (likelihood function) для логистической регрессии. Функция правдоподобия показывает вероятность получения наблюдаемых данных при заданных параметрах модели. Для логистической регрессии, логарифм функции правдоподобия записывается как:

где - наблюдаемое значение целевой переменной, – вероятность события при заданных параметрах модели и входных данных.

1. Вычислите градиент логарифма функции правдоподобия по параметрам модели. Градиент показывает направление наибольшего возрастания функции правдоподобия и используется для обновления параметров модели. Градиент для логистической регрессии выглядит следующим образом:

где – наблюдаемое значение целевой переменной, – вероятность события при текущих параметрах модели, – входные данные (признаки) для -го наблюдения.

1. Инициализация начальных значений параметров модели.
2. Повторение следующих шагов до выполнения условий сходимости:
   1. Вычислите градиент логарифма функции правдоподобия по текущим параметрам модели.
   2. Обновите параметры модели, используя градиентный спуск:

где – новые значения параметров модели, – текущие значения параметров модели, – скорость обучения (шаг градиентного спуска), – вычисленный градиент.

* 1. Повторите шаги 4.1 и 4.2 до достижения сходимости или выполнения другого условия остановки.

Метод градиентного спуска позволяет обновлять параметры модели в направлении, противоположном градиенту, чтобы максимизировать логарифм функции правдоподобия. Повторяя этот процесс и обновляя параметры, вы можете найти оптимальные значения параметров модели, которые наилучшим образом соответствуют данным.

Логистическая регрессия может использоваться как один из методов бинарной классификации: при заданной обучающей выборке оцениваются параметры этой регрессии, характеристики исследуемого объекта подставляются в уравнение модели (2.5): при объект относится к одному классу, в противном случае – к противоположному.

# ГЛАВА 3. Применение теоремы Байеса в машинном обучении.

## **3.1 Метод MAP**

Давайте обозначим наши данные — , а наши гипотезы — . Тогда нам надо найти вероятность гипотезы для наших данных , которая по теореме Байеса равна:

Нас интересует только соотношение вероятностей, поэтому мы можем выкинуть из этого выражения ( не зависит от ) и (будем считать, что все гипотезы равновероятны; строго говоря, это не всегда так, мы ещё вернёмся к этому вопросу, но во многих случаях это справедливо).

Получается, что нам надо найти такую гипотезу , для которой максимально . Говоря математическим языком, мы должны для каждой гипотезы вычислить её апостериорную вероятность и выбрать гипотезу, для которой эта вероятность максимальна. Этот подход называется maximum a posteriori probability (MAP). Если строго следовать ему, то вы получите лучшую гипотезу. И это математически доказано.

К сожалению, полное следование MAP возможно далеко не всегда. Например, данных слишком много. Или решение надо дать быстро. Поэтому, в реальной жизни используются различные модификации MAP.

Однако MAP обладает некоторыми недостатками:

1. Честное Байесовское обучение предполагает, что мы вычисляем вероятности **всех** возможных гипотез для **всех** обучающих данных. То есть, если у нас 1000 гипотез и 1000 обучающих паттернов, то нам придётся рассчитать 1000000 апостериорых вероятностей. В реальной жизни, обычно, числа больше на порядки. Выполнить полный рассчёт просто невозможно;
2. Вторая важная проблема возникает при создании классификаторов на основе Байесовского обучения. Мы хотим решить: мужчина перед нами, или женщина. У нас есть 4 гипотезы. Их вероятности распределяются так:

Если мы выберем MAP-гипотезу, то это будет h₄. И она скажет нам, что перед нами мужчина. Однако, сумма остальных гипотез — 0.6, что больше, чем 0.4. То есть, следовало бы ответить, что перед нами женщина. Проблема не в том, что Байесовское обучение плохо. Просто в данном случае мы должны рассматривать не отдельные гипотезы, а все возможные сочетание гипотез. Тогда мы получим идеальный Байесовский классификатор. Основная его проблема даже не в том, как его получить, а уже в том, как его использовать. Ведь даже имея идеальный классификатор, чтобы им воспользоваться нам надо вычислить некую идеальную комбинацию всех гипотез.

## **3.2. МСМС**

Рассмотрим вероятностное распределение . Методы Монте Карло (методы статистических испытаний) предполагают генерацию выборки из этого распределения:

Данная выборка может быть использована для оценки вероятностных интегралов вида

К интегралам такого вида сводятся многие шаги при осуществлении байесовского вывода в вероятностных моделях. Например, такими интегралами являются функция обоснованности ) и прогнозное распределение в модели RVM, а также функционал на М-шаге ЕМ-алгоритма Здесь стоит отметить, что плотность в подобных вероятностных интегралах часто известна с точностью до нормировочной константы

Выборка может быть также использована для оценки моды распределения

т.к. появление точек выборки наиболее вероятно в областях больших значений плотности.

Основной вопрос, раскрываемый в дальнейшем, состоит в том, как эффективно сгенерировать выборку из вероятностного распределения, заданного своей плотностью или своей ненормированной плотностью

Рассмотрим теперь вопрос генерации выборки из распределения в многомерном пространстве с помощью методов Монте Карло по схеме марковской цепи (MCMC). В этих методах вводится некоторая марковская цепь с априорным распределением и вероятностями перехода в момент времени а генерация выборки происходит следующим образом:

Заметим, что при таком подходе генерируемая выборка не является набором независимых случайных величин. Однако, она подходит для оценки вероятностных интегралов вида (3.1) или оценки моды распределения. В том случае, если необходимо получить набор независимых величин, достаточно проредить полученный набор , взяв каждый -ый отсчет, где достаточно велико.

В дальнейшем рассматривается вопрос о том, как выбрать вероятности перехода таким образом, чтобы выборка, генерируемая по схеме (3.2), была бы выборкой из интересующего нас распределения

### **Cхема Метрополиса-Хастинга**

Рассмотрим схему генерации нужной выборки *Метрополиса-Хастингса*

Пусть необходимо сгенерировать выборку из распределения p(T), известного с точностью до нормировочной константы:

Рассмотрим шаг генерации по схеме Метрополиса-Хастингса. Пусть на шаге сгенерирована конфигурация . Тогда на шаге сначала генерируется конфигурация из некоторого предложного распределения Затем вычисляется величина

и точка принимается в качестве следующей точки с вероятностью В противном случае, . Таким образом, мы ввели марковскую цепь с вероятностью перехода

Покажем, что распределение p(T) является инвариантным относительно введенной марковской цепочки. Если , то инвариантность сохраняется, т.к. значение не изменяется. Для случая проверим выполнимость уравнения детального баланса:

Для эргодичности введенной марковской цепи достаточно потребовать выполнение .

В том случае, если предложное распределение является симметричным, т.е. то схема Метрополиса-Хастингса переходит в классическую схему Метрополиса. Согласно этой схеме, если значение плотности в новой точке оказалось выше, чем значение плотности в предыдущей точке , то эта точка гарантированно принимается в качестве следующей точки выборки. Если плотность в новой точке оказалась меньше, то такая точка тоже может быть принята, но с вероятностью, пропорциональной величине уменьшения плотности.

### **Схема Гиббса**

Пусть необходимо сгенерировать выборку из многомерного распределения где Рассмотрим шаг генерации по схеме Гиббса. Пусть на шаге сгенерирована конфигурация Тогда генерация следующей точки выборки происходит следующим образом:

Здесь через обозначено маргинальное одномерное распределение значений -ой компоненты при условии всех остальных. Таким образом, согласно схеме Гиббса генерация выборки из многомерного распределения заменяется на итерационную генерацию точек из одномерных распределений. По аналогии с методами одномерной оптимизации генерация выборки из одномерного распределения является существенно более простой задачей, чем генерация выборки из многомерного распределения.

Докажем, что распределение является инвариантным относительно введенной марковской цепи. Рассмотрим один шаг генерации очередной компоненты . По предположению индукции . Тогда совместная конфигурация . Отсюда, совместное распределение является инвариантным относительно одного шага процесса генерации (3.3). Следовательно, оно является инвариантным и относительно всего процесса (3.3).

При реализации схемы Гиббса на практике часто допускается следующая ошибка:

вместо шага

делается шаг

т.е. в условие подставляются значения компонент только с предыдущей итерации. При таком подходе вероятность перехода в марковской цепи определяется как

Распределение не является инвариантным относительно данной марковской цепи. Эту ситуацию легко исправить, если взять схему Метрополиса-Хастингса, где в качестве предложного распределения фигурирует распределение (3.4). Заметим, что в отличие от схемы Гиббса, схема Метрополиса-Хастингса с предложным распределением (3.4) легко распараллеливается и на практике в некоторых ситуациях может работать быстрее, чем схема Гиббса.

# 

# ГЛАВА 4. Прогнозирование временного ряда на основе моделей авторегрессии

## **4.1. Прогнозирование по модели авторегресии**

Случайный процесс называется *стационарным* (в широком

смысле), если его математическое ожидание и дисперсия постоянны (значение *μ* определяет постоянную, вокруг которой лежат значения временного ряда, а *σ* характеризует ширину полосы, в пределах которой рассеяны эти значения), а автоковариационная функция зависит лишь от разности аргументов:

Стационарность процесса может быть оценена визуально по графику временного ряда или с помощью специальных тестов.

В предположении о нормальности распределения можно разбить ряд на две части, проверить гипотезы о равенстве дисперсий в каждой части по критерию Фишера, затем – о равенстве средних по критерию Стьюдента.

Иногда целесообразно представить значение ряда в момент *t* в

виде *авторегрессионной* зависимости: линейной функции от предыдущего наблюдения плюс случайный компонент.

*Модель авторегрессии* (AutoRegressive) *первого порядка* АR*(1)*

имеет вид

где – числовые коэффициенты – последовательность

величин, образующих *белый шум* (так называется последовательность

некоррелированных случайных величин с нулевым математическим

ожиданием и постоянной дисперсией).

Свободный член часто приравнивается нулю, то есть

рассматриваются центрированные процессы, средний уровень

которых равен нулю – .

Прогноз

и т.д.

Модель авторегрессии порядка *p* AR*(p)* определяется выражением:

Прогноз по модели авторегрессии порядка *p*:

и т. п.

Коэффициенты модели авторегрессии определяются методом наименьших квадратов. Значимость модели авторегрессии соответствующего порядка проверяется по критерию Стьюдента.

## **4.2. Прогнозирование по модели скользящего среднего**

Случайную составляющую временного ряда иногда целесообразно представить и в виде взвешенной суммы настоящего и прошлых значений белого шума:

Такой процесс называется *моделью скользящего среднего* (Moving Average) порядка *q* и обозначается *MA(q).*

Как и в случае авторегрессионных моделей, наибольшее распространение на практике имеют модели первого и второго порядка:

Часто полезно включить в модель и слагаемые, описывающие авторегрессию, и члены, соответствующие скользящим средним. *Модель авторегрессии – скользящего среднего ARMA(p,q)* имеет вид

В частности, *ARMA(1,1):*

Прогноз выполняется так:

где

и т.д.

Модель *ARMA*(2,2):

Прогноз

и т.д.

## **4.3. Модели авторегрессии – проинтегрированного скользящего среднего для нестационарных временных рядов**

Мы рассмотрели некоторые модели стационарного временного ряда. Реальные временные ряды обычно *нестационарны*. Однако на практике часто встречаются такие временные ряды, которые стационарны после вычитания из уровней ряда его неслучайной составляющей.

Назовем последовательной разностью первого порядка ряда новый ряд:

последовательной разностью второго порядка – разность последовательных разностей первого порядка:

*последовательной разностью k-го порядка* –

Предположим, что нестационарную составляющую можно представить в виде полинома степени *(k – 1).* Можно доказать, что переход к последовательным разностям *k*-го порядка исключает неслучайную составляющую. Будем предполагать, что оставшийся ряд, включающий только случайную составляющую, может быть представлен в виде модели *ARMA(p,q).*

Тогда получим модель

называемую моделью *авторегрессии – проинтегрированного скользящего среднего* (Auto Regressive Integrated Moving Average), или сокращенно, *ARIMA(p,k,q).*

Свободный член в модели (4.8) часто приравнивается нулю.

В общем случае сначала подбирают параметр *k* (обычно *k* ≤2), затем после перехода к стационарной модели проводится идентификация модели *ARMA(p,q)*, при этом, как правило, и *p* ≤ 2, и *q* ≤ 2.

Прогноз, например, для модели *ARIMA*(1,1,1), можно найти так:

тогда

и т.д.

Для модели *ARIMA*(2, 2, 1) по аналогии:

тогда

и т.п.

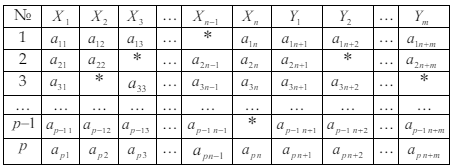
# ГЛАВА 5. Практическое задание

## **5.1. Математическая постановка задачи восстановления пропусков в таблице данных**

В общей постановке задача восстановления пропусков в таблицах данных является такой: Пусть – вектор входных факторов, – вектор результирующих характеристик, – количество экспериментов или периодов рестроспективы, матрица исходных данных. Среди ее элементов есть пропуски.

Предположим, что между входными факторами и результирующими характеристиками существуют зависимости

Рассмотрим структуру исходной информации (Таблица 4.1).



**Таблица 5.1.** Структура исходной информации

Тогда задача восстановления пропусков в данных состоит в поиске

где и – векторы значений,

которые получены по идентифицированным зависимостям и

приведены в табл. 7.1, соответственно. Задачу минимизации детализируем и перепишем в виде

или

Если предположить, что зависимости (4.1) линейные, то есть

тогда задача восстановления пропусков состоит в поиске

где .

Решение задач (4.2)–(4.5) имеет первый этап, который, в

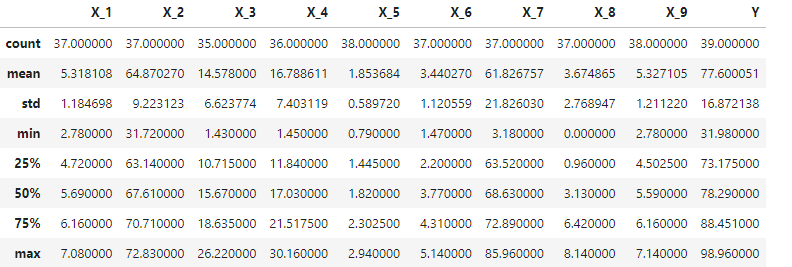
общем случае, состоит в идентификации зависимостей Заметим, что в задаче восстановления пропусков в таблицах данных процедуры идентификации и оптимизации итеративно повторяются.

Рассмотрим набор данных для нашей задачи

****

**Таблица 5.2.** Набор данных

Рассмотрим дескриптивные статистики для этого набора



**Таблица 5.3.** Дескриптивные статистики

Как можем заметить, количество значений в каждом столбце отличается, ориентиром возьмем столбец *“Y”,* в котором нет пропусков.

Посчитаем количество пропусков в каждом столбце.



**Таблица 5.4.** Количество пропусков

## **5.2. Эвристические методы обработки некомплектных данных**

Выполним анализ методов восстановления утраченных данных, применяемых в большинстве случаев. Матрица, строка или столбец, имеющие пропуски данных, называются некомплектными.

*1. Метод заполнения средним значением.*

Согласно этому методу, отсутствующее значение на пересечении *i*-ой строки и *j*-го столбца рассчитывается по формуле:

где *q-*количество заполненных элементов в *j*-м столбце.

Преимуществом метода является простота.

Недостатки: в одном столбце может быть большое количество пропусков и все они будут заполнении одинаковыми значениями; не учитывается связь некомплектной строки с другими строками, что приводит к смещению и недостоверной оценке неизвестного значения. Этот вывод справедлив и для случая осуществления перерасчета с учетом каждого заполненного пропуска.

Воспользуемся этим методом для нашей задачи, для его применения рассчитаем средние значения каждого столбца



**Таблица 5.5.** Средние значения каждого столбца

Заполним этими значениями пропуски в соответствующих столбцах





**Таблица 5.6.** Пропуски заполнены средними значениями

*2. Метод исключения некомплектных строк.*

Применяется в случае незначительного количества пропусков. Метод является простым, но изъятие данных увеличивает энтропию прогнозных значений и ведет к смещенности параметров модели.

*3. Метод множественной линейной регрессии*

Применяется в предположении, что зависимость (4.1) является линейной. Для ее идентификации используют только комплектные данные и с использованием МНК получают (4.5). Очевидно, что в дальнейшем зависимости (4.5) используются для восстановления пропусков, но адекватно это можно делать только в случае одного пропуска среди значений строки .

Если таких пропусков два и больше, то задача решается при дополнительных предположениях и ограничениях. Метод требует выполнения ряда предусловий и проверок входных факторов на мультиколлинеарность, гетероскедастичность, автокорреляцию и применения модифицированных версий МНК.

## **5.3. Восстановление пропусков значений зависимой переменной**

Рассмотрим случай, когда проводится активный эксперимент и значения факторов заданы исследователем, а *Y* – зависимая от этих факторов переменная. Очевидно, что тогда пропусков среди значений входных факторов намного меньше, чем среди значений результирующей характеристики.

1. *Resampling-метод*

Предположим, что пропущенные значения есть только среди значений

результирующей характеристики Y и строки, которые им отвечают, находятся вверху таблицы исходных данных. Применение метода *resampling* может быть выполнено двумя способами:

*Resampling-1*

*Шаг 1*. Формируем матрицу полных наблюдений , количество строк в которой равно .

*Шаг 2*. Случайным образом выбираем *j*-ю строку из матрицы *H* и заменяем *i*-ю строку начальной матрицы, . Эта строка может выбираться случайно или по порядку, начиная с -й.

*Шаг 3*. Если все пропуски заполнены, то по МНК находим коэффициенты регрессионного уравнения в противном случае выполняем переход на шаг 2.

*Шаг 4*. Если получено *r* векторов то находим средние значения коэффициентов регрессионной модели:

.

*Resampling-2*

*Шаг 1.* По матрице *H* строим регрессионную модель и находим оценки коэффициентов

*Шаг 2*. Рассчитываем оценку по регрессионной модели для

*Шаг 3.* Находим ошибки

*Шаг 4.* Для каждого пропуска, подставляя значения сопутствующих переменных в полученное регрессионное уравнение, находим оценку

*Шаг 5*. Значения, которыми заменяют пропуска, получаем по формуле:

где выбирают случайным образом из результатов шага 3.

*Шаг 6.* По данным, полученным после заполнения, строим регрессионную модель и находим оценки коэффициентов

*Шаг 7*. Аналогичен шагу 4 из *resampling-1*.

Преимуществом метода *resampling* является полное использование исходной информации, вместе с тем ее повторное использование уменьшает информативность данных.

## **5.4. EM-алгоритм**

*Алгоритм ЕМ* базируется на максимизации математического ожидания. В алгоритме реализована итерационная процедура вычисления

оценки максимального правдоподобия и состоит он из двух шагов:

*Шаг 1.* Множество данных неполной задачи и текущее

значение вектора параметров используются для получения

расширенного полного набора данных.

*Шаг 2.* Вычисляется новая оценка вектора параметров путем максимизации функции логарифмического правдоподобия полного множества данных.

Шаги 1 и 2 повторяются до полной сходимости.

Приведем математическую запись алгоритма ЕМ.

Пусть *r* – полный вектор данных, которым он должен был

бы быть, но не является; – вектор неполных данных, который фактически наблюдается. Таким образом, векторы *r* и *d* являются элементами соответствующих пространств: . Обозначим – функцию плотности условной вероятности *r* для данного вектора параметров . Отсюда следует, что функция плотности условной вероятности случайной переменной *d* для данного вектора определяется

следующим образом:

где – подпространство *R*, которое определяется равенством

Необходимо найти значение , которые максимизирует

функцию логарифмического правдоподобия на неполных

данных

Эта задача решается итеративным применением функций логарифмического правдоподобия на полных данных

которая является случайной переменной, поскольку отсутствующие данные являются неизвестными.

Если предположить, что значение вектора параметров на *n*-й итерации, то на первом ее шаге находим математическое ожидание

по . На втором шаге вычисляем максимум функции по в пространстве параметров таким образом, чтобы найти оценку вектора

Алгоритм начинается заданием некоторого начального значения

вектора параметров . Шаги алгоритма итеративно повторяются, пока различие между и не станет меньшим некоторого заведомо малого заданного значения. Тогда работа алгоритма завершается.

# ЗАКЛЮЧЕНИЕ

В ходе работы были получены следующие результаты

* Было рассмотрено понятие временного ряда
* Дан обзор методов машинного обучения
* Исследован байесовский подход в машинном обучении
* Приведен практический пример применения теоремы Байеса.

# Приложение 1

Пример использования формулы Байеса.

Людям дали следующее описание человека по имени, например, Иван: «*Иван – застенчивый, нелюдимый, стеснительный парень, редко контактирует с людьми.»* И попросили сказать, работает он фермером или библиотекарем.

Исследователи определили, что большинство опрошенных людей, получив тако описание, сказали, что он – библиотекарь, однако по статистике фермеров в США в двадцать раз больше, чем библиотекарей. Соответственно, выбор людей не был рационален, так как они не учли факт того, что гораздо больше шансов встретить фермера, чем библиотекаря.

Рассмотрим следующую таблицу

В полученной таблице имеем пять строк по двадцать одному элементу, где один библиотекарь и двадцать фермеров. Предположим, что 40% библиотекарей подходит под описание, то есть , где гипотеза о том, что выбранный человек – библиотекарь, гипотеза о том, что человек подходит под описание. А фермеров, подходящих под описание – 10%, то есть , где гипотеза о том, что выбранный человек – фермер.

Приняв это всё во внимание, мы получим, что вероятность того, что библиотекарь подойдет под описание равна:

то есть мы делим количество библиотекарей, подходящих под описание на общее число людей, подходящих под описание.

Вероятность того, что выбранный человек – библиотекарь

такая вероятность является априорной вероятностью.

Вероятность называется функцией правдоподобия.

Так же необходимо учитывать и противоположные исходы

Получим формулу

где количество библиотекарей, подходящих под описание; количество фермеров, подходящих под описание; общее число людей, составляющих выборку

Обозначим вероятность получения новых данных.

Вероятность называется апостериорной вероятностью (шанс того, что гипотеза верна с учетом новых данных).

# СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ

1. Кувайскова, Ю. Е Статистические методы прогнозирования / Ю. Е. Кувайскова, В. Н. Клячкин/ – Минск: УГТУ, 2019.
2. Ивченко, Г. И. Математическая статистика: Учеб. Пособие для втузов / Ивченко Г. И., Медведев Ю. И./ – Москва: МВТУ им. Н. Э. Баумана, 1984.
3. Грэйнджер, Клайв У. Дж. Эконометрический анализ временных рядов / Клайв У. Дж. Грэйнджер.
4. Абрамова, А. Е. Метод Монте Карло по схеме марковской цепи для оценки вероятности редких событий в задачах биоинформатики / Абрамова А. Н./ – Санкт-Петербург: СПБГУ, 2017.
5. Временной ряд – Википедия: [Электронный ресурс]. URL: https://ru.wikipedia.org/ (Дата обращения: 26.11.2023).
6. Градиентный спуск: всё, что нужно знать: [Электронный ресурс]. URL: https://neurohive.io/ (Дата обращения: 10.12.2023).
7. Теорема Байеса [3Blue1Brown] – YouTube: [Электронный ресурс]. URL: https://www.youtube.com/ (Дата обращения: 02.12.2023)
8. Hamilton, J. D. Time Series Analysis / James D. Hamilton/ – Princeton, New Jersey: Princeton University, 1994.