**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики и информатики**

**Кафедра математического моделирования и анализа данных**

**Курсовая работа**

по теме

«Разработка байесовских авторегрессионных моделей для прогнозирования стационарных временных рядов»

Студента 3 курса 7 группы

Карповича Артёма

**Преподаватель**

Лобач В.И.

Минск, 2023

Оглавление

[Введение 3](#_Toc153282927)

[ГЛАВА 1. Применение машинного обучения для прогнозирования 5](#_Toc153282928)

[ГЛАВА 2. Статистические методы 9](#_Toc153282929)

[Глава 2.1. Метод Байеса 9](#_Toc153282930)

[Глава 2.2. Дискриминантный анализ 12](#_Toc153282931)

[Глава 2.3. Логистическая регрессия 14](#_Toc153282932)

[Приложение 1 16](#_Toc153282933)

[СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ 18](#_Toc153282934)

# 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: детерменированные 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* – наука об извлечении закономерностей из

ограниченного количества данных.

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 *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 :

(6.1)

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

, (). (6.2)

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*:

(6.3)

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, который минимизирует функционалы (6.5) или (6.6), 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 (6.5) – (6.6)) 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

## **Глава 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*:

*,* (6.7)

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 (6.8) 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 (6.9) 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.

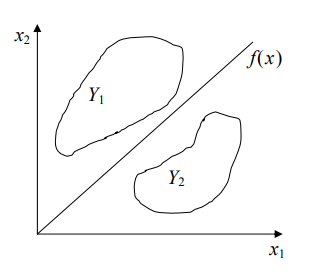
Пример использования приведен в Приложении 1.

## **Глава 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. К какому классу относится этот объект?

*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. 6.1). Необходимо разделить эти классы некоторой функцией *f*(*x*), которая называется *каноничекой дискриминантной функцией.* На рисунке 6.1 показана линейная функция (линейный дискриминант Фишера):



**Рис. 6.1.** 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

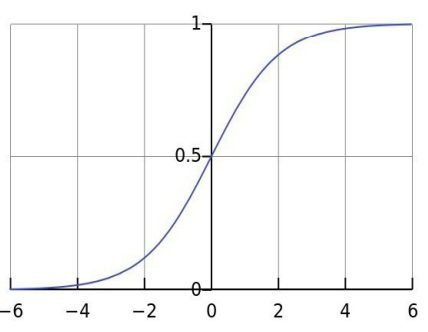
то объект относится к классу 1 (исправен), в противном случае – к классу 2 (неисправен).

## **Глава 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 6.2, значения функции лежат в диапазоне от нуля до 1.



**Рис. 6.2.** 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 (6.12): при объект относится к одному классу, в противном случае – к противоположному.

# 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. Статистические методы прогнозирования; Ю. Е. Кувайскова, В. Н. Клячкин. — Минск: УГТУ, 2019.
2. Эконометрический анализ временных рядов; Клайв У. Дж. Грэйнджер.
3. Временной ряд — Википедия: [Электронный ресурс]. URL: https://ru.wikipedia.org/ (Дата обращения: 26.11.2023).
4. Градиентный спуск: всё, что нужно знать: [Электронный ресурс]. URL: https://neurohive.io/ (Дата обращения: 10.12.2023).
5. Теорема Байеса [3Blue1Brown] – YouTube: [Электронный ресурс]. URL: https://www.youtube.com/ (Дата обращения: 02.12.2023)