МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УО «БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра математических методов в экономике

Специальность «Оптимальное планирование и управление в экономике»

Допущена к защите

Заведующий кафедрой

д-р экон. наук, доц.

­­­­\_\_\_\_\_\_\_\_ Г.О. Читая

31.05.2020

**ДИПЛОМНАЯ РАБОТА**

на тему: **Разработка нейронных сетей и их использование при принятии решений о выдаче кредита (на примере ОАО «Белинвестбанк»)**

|  |  |  |
| --- | --- | --- |
| Студент  ФЦЭ, 4-й курс, ДКК-1 |  | Ф.А. Кобак |
|  |  |  |
| Руководитель  кнад. экон. наук,  профессор |  | Э.М. Аксень |
|  |  |  |
| Нормоконтролер |  | И.В. Денисейко |

МИНСК 2022

**РЕФЕРАТ**

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc99060264)

[1. Теория моделей искусственной нейронной сети при задаче классификации 5](#_Toc99060265)

[ЗАКЛЮЧЕНИЕ 6](#_Toc99060266)

# ВВЕДЕНИЕ

# Теория моделей искусственной нейронной сети в задаче классификации

# 1.2 Постановка задачи

Рассматриваемые в этой работе методы, на ряду с некоторыми другими, предназначены для решения задач классификации. Задачей классификации называется задача в которой требуется определить способ отнесения некоторых объектов к некоторым группам (классам).

Такие задачи разбиваются на два вида – те для которых известно число классов и те для которых число классов заранее неизвестно. В этой работе рассматриваются только задачи первого вида.

Особенностью описного типа задач является наличие предсказываемого фактора Y. То есть, для каждого объекта из изучаемой совокупности имеется признак который может принимать значения :

|  |  |
| --- | --- |
|  | (1.1) |

Предположим, имеются наблюдения за некоторым явлением или процессом которые можно представить подобно таблице 1.1.

**Таблица 1.1 – Исходная система данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| i | *Y* | *X*1 | *X*2 | … | *X*m |
| 1 | *y*1 | *x*11 | *x*12 | … | *x*1*m* |
| 2 | *y*2 | *x*21 | *x*22 | … | *x*2*m* |
| … | … | … | … | … | … |
| *n* | *y*n | *xn*1 | *xn2* | … | *xnm* |

Примечание – Источник: собственная разработка.

Где по строкам расположились наблюдения, а по столбцам некоторые переменные для этих наблюдений. На основе этих данных формируется правило, которое позволяет, получив произвольный набор значении переменных , наиболее точно, в некотором смысле, предсказать значения . В литературе это правило нередко обозначается так:

Это очень распространённая задача для прикладной статистики и машинного обучения, встречающаяся во многих сферах жизни: задача определения диагноза по симптомам и анализам; разбиение электронной почты на действительную и спам; задачи распознавания рукописного текста и множество других.

Итак, была получена задача, в которой имеется набор наблюдений с откликом *Y*, аналогичная задача решается классическим регрессионным анализом, разница лишь в типе отклика – для обычной регрессионной модели он численная переменная, в нашем же случае номинативная (категориальная).

Номинативной называют переменную, для которой не определены ни порядок, ни шкала[1 с.316]. Под отсудившем шкалы понимается, что мы не можем как-либо определить расстояние между двумя различными значениями переменной, например, если предсказывается факт возврата или невозврата кредитополучателем задолженности, нет никакой возможности показать на сколько возврат «выше» или «ниже» невозврата. Понятие порядка переменной будет более подробно раскрыто в третьем подразделе, когда речь пойдет о упорядоченной модели логистической регрессии.

Описанная разница в типе отклика и порождает непригодность использования классической регрессионной модели.

Рассмотрим задачу бинарной классификации – отклик может принимать лишь два значения (*K=2*), для простоты обозначим события (1.1) каким-либо числами, обычно используется:

После оценивания будет получено регрессионное уравнение следующего вида:

где – оценка коэффициента регрессии;

– оценка свободного члена;

– случайная ошибка.

Отдельного обсуждения достойна переменная , она представляет собой оценку вероятности того, что исследуемый объект принадлежит к классу советующему числу 1. Тут возникает первая проблема такого подхода – нет никаких оснований чтобы , что в корне неверно для понятия вероятности. Наглядная иллюстрация этой проблемы представлена на рисунке 1.1 слева – две выборки распределены вдоль некоторой переменной x, притом для тех для которых , значения *х* заметно выше, построив регрессионную прямую наглядно убеждаемся в том, что ряд наблюдений получают оценки вероятности за границами нуля и единицы.



**Рисунок 1.1 – Оценки вероятностей через линейную регрессионную модель**

Примечание ­­­– Источник: собственная разработка.

На рисунке 1.1 справа наглядно представлена еще одна проблема описанного подхода – к той же выборке, которая обсуждалась выше, был добавлен ряд наблюдений с уровнем исследуемого фактора соответствующим , с заметным смещением в большую сторону по фактору *х*. Регрессионная прямая в таком случае сместилась и заметно хуже предсказывает вероятности для старых наблюдений – вся группа советующая , получила оценки вероятностей того, что они принадлежат группе , в районе 0,4, что вообще говоря достаточно плохо для такого простого примера. Пример того как с аналогичной задачей справиться логистическая регрессия будет представлен в следующем подразделе.

Более того, такая модель плохо обобщается для случая, когда требуется решение для задачи не бинарной классификации. [2 c.145]

Наличие всех описанных проблем при использовании регрессионного анализа для моделирования процесса с номинативным откликом и вызвало потребность в разработке специальных методов классификации, одним из которых является логистическая регрессия, рассматриваемая в следующем подразделе.

# 1.2 Модель логистической регрессии – линейный классификатор

Отталкиваясь от того, что было сказано в прошлом разделе, проведем ряд рассуждений, которые приведут к модели логистической регрессии, для задачи бинарной классификации.

Обозначим вероятность того, что исследуемый признак примет значение как . Одной из ключевых проблем в данном случае является то что , в то время как отклик в модели регрессионного анализа принимает значение .

Введём понятие шанса события, шансом появления некоторого события называется отношение вероятности появления этого события к вероятности появления любого другого совместного события. В нашем случае справедливо:

Несложно показать, что . Пользуясь свойствами логарифма получим, что , а такая переменная уже хороший кандидат, для того чтобы быть описанной методом линейной регрессии. Таким образом идея логистической регрессии предлагает предсказывать не вероятность того что , а логарифм отношения шансов этого события. Логарифм отношения вероятностей в дальнейшем будет называть логит функцией.

Для краткости последующих записей обозначим:

Получим аналитическую запись рассматриваемой модели. Запишем теоретическую линейную модель предсказывающую логарифм отношения шансов:

Имея значение логарифма отношения шансов легко получить искомую вероятность. Сначала проведем потенцирование рассматриваемого выражения:

Используя свойства натурального логарифма, получим:

Решив это уравнение относительно , получим общую запись модели логистической регрессии:

|  |  |
| --- | --- |
|  | (1.2) |

Выражение (1.2) и есть модель логистической регрессии для случая бинарной классификации[3 c.32]. Функция, лежащая в основании этой модели, соответствует функции логистического распределения и имеет два ключевых полезных для рассматриваемой задачи свойства: во-первых, она принимает значения в диапазоне от нуля до единицы, во-вторых она принадлежит к классу сигмовидных функций, то есть это гладкая, возрастающая функция, имеющая на графике форму буквы «S». В литературе распространено название – сигма функция или сигмоида и в общем она записывается так:

Вернемся к примеру, из предыдущего раздела и посмотрим, как логистическая регрессия справиться с поставленной задачей. На рисунке 1.2 синими точками по-прежнему обозначены сгенерированные наблюдения, красной линией теперь обозначаются оценки вероятностей для различных значений предиктора *х*, полученные по модели подобной (1.2). Как видно обе обозначенные проблемы решены, предсказания лежат строго в пределах от нуля до единицы, и смещенная выборка почти не влияет на качество модели.



**Рисунок 1.2 – Оценки вероятностей через логистическую регрессионную модель**

Примечание ­­­– Источник: собственная разработка.

Рассмотрим более обобщённый вариант – когда предсказываемая переменная не бинарна, а имеет более двух уровней. Такую модель принято называть мультиномиальной логистической регрессией.

В данном случае предсказываемый фактор будет закодирован, так:

В данном случае нам понадобиться *K*-1 логит функции. Кроме того, надо выбрать базовый уровень выходной переменной, пусть, не нарушая общности, это будет *Y*=0. Для такого случая логиты можно записать так:

|  |  |
| --- | --- |
|  | (1.3) |

Притом, каждый из них будет описываться предикторами в соответствии со следующим правилом:

Чтобы получить в данном случае запись, подобную выражению (1.2) для биномиальной регрессии, потребуется поработать с системой (1.3). Для начала проведем потенцирование каждого уравнения системы:

|  |  |
| --- | --- |
|  | (1.4) |

Теперь из одного уравнения выразим , пусть, не нарушая общности, это будет первое уравнение:

Используя определение полной вероятности, получим:

|  |  |
| --- | --- |
|  | (1.5) |

Из оставшихся уравнений системы (1.3) получим:

|  |  |
| --- | --- |
| . | (1.6) |

Перенеся левую часть выражения (1.5) приведя и общему знаменателю, положив его неравным и вынеся за скобки, получим:

Выражая от сюда получим:

|  |  |  |
| --- | --- | --- |
|  |  | (1.7) |

Вообще говоря, выражение (1.6) будет справедливо и при *k*=1, потому подставляя туда (1.7), легко получим вероятности появления других уровней исследуемого признака:

|  |  |  |
| --- | --- | --- |
|  |  | (1.8) |

Выражения (1.7) и (1.8) – самая общая запись логистической регрессии. Для удобства последующих записей, предполагая что , запишем одной формулой:

|  |  |  |
| --- | --- | --- |
|  |  | (1.9) |

Сразу повторюсь, для заострения внимания – вероятность проявления каждого возможного уровня отклика имеет свою функцию от параметров и входных данных . Заметим, что оценки коэффициентов и , получают методом максимального правдоподобия.

Теперь поговорим о том, почему методы машинного обучения в задачах классификации не остановились на логистической регрессии. В литературе можно встретить утверждение: «модель логистической регрессии – линейный классификатор». На первый взгляд может показаться, что это утверждение неверно, ведь выражение (1.2) никак не назвать линейным входного набора данных и сигмоида на рисунке 1.2 в целом кривая. Дело тут кроется в принципе принятия решения о отнесении наблюдения к тому или иному классу. Данное явление лучше всего рассматривать на примере двумерной задачи.

На рисунке 1.3 представлена диаграмма рассеяния опять же сгенерированных случайных данных. Данные двумерные и разделяются на два класса.



**Рисунок 1.3 – Диаграмма рассеяния двумерных классифицированных данных**

Примечание ­­­– Источник: собственная разработка.

По прежнему, нет никакой сложности в том, чтобы провести вручную линию с полной точностью отделяющую один класс от другого (хотя, в отличии от одномерно примера, уже понадобиться знания аналитической геометрии чтобы записать классифицирующее правило). Однако, в целях изучения метода, взглянем на то как с этой задачей справиться логистическая регрессия. На рисунке А.1 двумерный аналог рисунка 1.2 – соответствующая рассматриваемому примеру двумерная сигмоида. На нее нанесены наблюдения с рисунка 1.3.

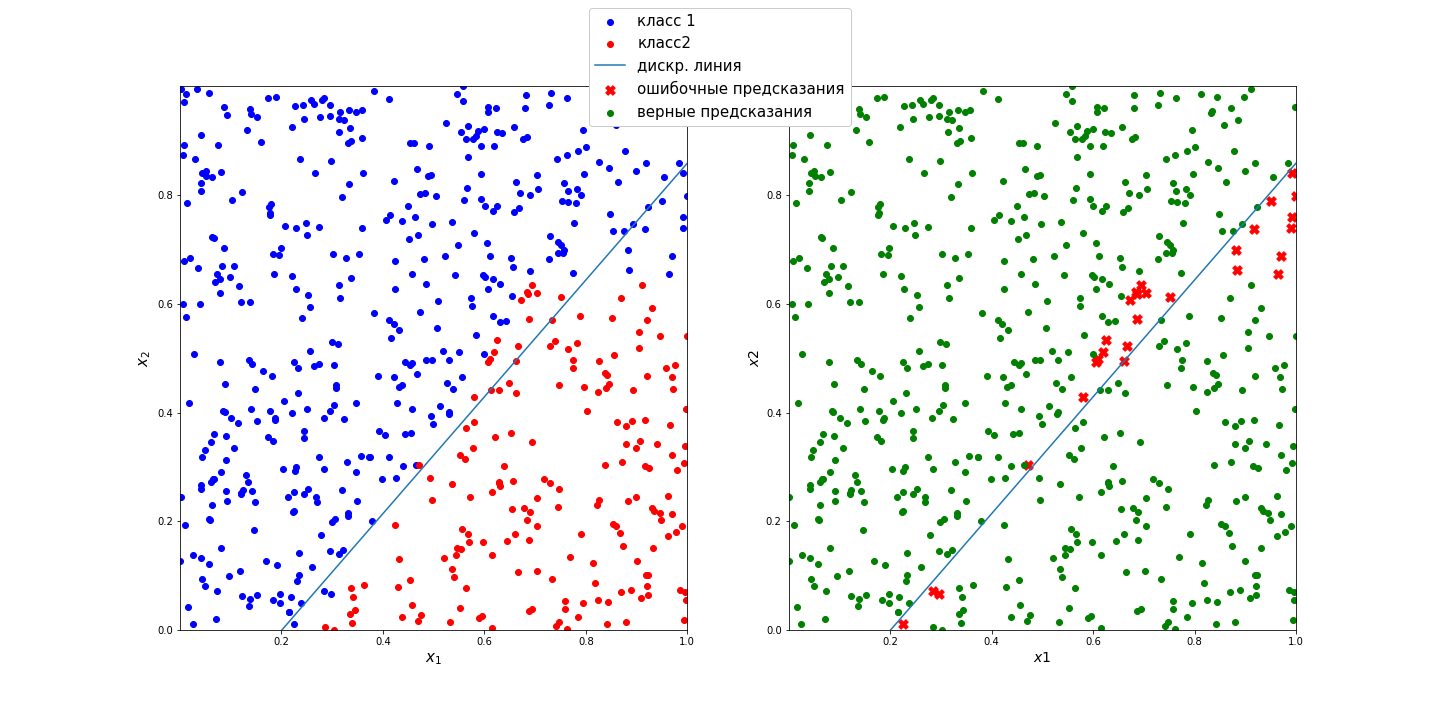
Из рисунка понятно, что можно подобрать такую высоту что бы точки принадлежащие отдельным классам были разделены, а соответствующая линия уровня «упадет» на рисунок 1.3 так что идеально разделит классы и в отрыве от высоты сигмоиды. Покажем, что такая линия уровня будет линейной – она возникает при сигмоиде равной :

|  |  |  |
| --- | --- | --- |
|  |  | (1.10) |

Требуется разрешить уравнение относительно выражения под сигмоидой. Такую операцию мы уже продевали раньше, только наоборот, потому просто запишем результат:

Заметив, что правая часть выражения константа, скажем, что разделяющая линия уровня линейна.

Тут и рождается «почва» для дальнейшей эволюции методов классификации связанных с логит функцией. Как и ранее рассмотрим пример с которым данный метод справится плохо – рисунок 1.4.



**Рисунок 1.4 – Диаграмма рассеяния двумерных классифицированных данных при нелинейном принципе классификации**

Примечание ­­­– Источник: собственная разработка.

Слева все та-же диаграмма рассеяния, только несколько изменены принципы классификации. По по-прежнему можно вручную записать правило разделяющее классы, только в этот раз это будет не единственные уравнение прямой, но кусочно-линейная функция.

На этих данных была построена модель логистической регрессии. Выбрав по принципу наибольшей разности между распределениями классов по (этот принцип будет раскрыт в главе 3 на практике) и использовав (1.10) мы получили уравнение описывающее дискриминирующую прямую, так же нанесенную на график.

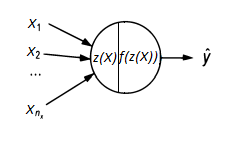
На рисунке справа более броско выделены ошибки модели. Как видно модель логистической регрессии, как и любая другая линейная модель с таким случаем справляется не идеально (хотя задача достаточно простая). Решение заключается в том, чтобы включить в модель некоторую нелинейность.

Этот подраздел посвящен логистической регрессии – модели которая не является целевой для данной работы, однако идеи в ней лежащие важны для полного раскрытия темы. Описаны плюсы по сравнению с линейной регрессией и выведены формулы позволяющие записать модель аналитически. Особо важной для дальнейшего повествования является логит функция (сигмоида, обратная функция логистического распределения). Показано на примере, почему логистическая регрессия является линейным классификатором и описаны причины дальнейшей эволюции методов классификации.

# 1.3 Модель искусственной нейронной сети в классификации

Как и любая другая математическая модель нейронная сеть может быть записана как система уравнений, но куда проще и понятнее рассматривать её как своеобразный граф по ребрам которого «текут» данные и преобразуются в узлах.

Базовой единицей выступает нейрон – узел графа. Раскроем его сруктуру. На рисунке 1.5 схематично представлен искусственный нейрон.



**Рисунок 1.5 – Схема искусственного нейрона**

Примечание ­­­– Источник: собственная разработка на основе [4 с.52].

# ЗАКЛЮЧЕНИЕ

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах; пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400с.
2. Введение в статистическое обучение с примерами на языке R / Г. Джеймс [и др.]; пер с англ. С.Э. Мастицкого. – М.: МДК Пресс, 2017. – 456 с.
3. Hosmer, D.W. Applied Logistic Regression / D.W. Hosmer, J.S. Lemeshow // University of Massachusetts. – 2nd ed. – 2000. – 376 p.
4. Микелуччи У. Прикладное глубокое обучение. Подход к пониманию глубоких нейронных сетей на основе метода кейсов/ У. Микелуччи; пер. с англ. – СПб: БХВ-Петербург, 2020. – 368 с.

# ПРИЛОЖЕНИЕ А

# Визуализация логистической регрессии



**Рисунок А.1 – Двумерная сигмоида для бинарного классификатора**

Примечание ­­­– Источник: собственная разработка.