МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УО «БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра математических методов в экономике
Специальность «Экономическая кибернетика»

Допущена к защите
Заведующий кафедрой
д-р экон. наук, доц.
_____ Г.О. Читая
31.05.2022

ДИПЛОМНАЯ РАБОТА

на тему: **Разработка моделей 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
доктор экон. наук,
профессор _____ Э.М. Аксень

Нормоконтролер _____ И.В. Денисейко

МИНСК 2022

РЕФЕРАТ

Дипломная работа: 112 с., 16 табл., 33 рис., 7 лист., 22 ист., 11 прил.

КЛАССИФИКАЦИЯ, КРЕДИТНЫЙ СКОРИНГ, ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ, АЛГОРИМ ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБКИ, ВАЛИДАЦИЯ МОДЕЛИ, ОТБОР ПОКАЗАТЕЛЕЙ, AUC, KS, ТЕСТ КОЛМОГОРОВА-СМИРНОВА

Объект исследования – правило классификации кредитополучателей в ОАО «Белинвестбанк».

Предмет исследования – модели искусственных нейронных сетей в приложении к задаче классификации.

Цель работы: построение нелинейного классификатора для целей кредитного скоринга.

Методы исследования: компьютерный анализ данных, машинное обучение.

Исследования и разработки: при отборе показателей и преобразованиях данных для модели задействован особый метод, использующий ROC анализ, подготовлен ряд инструментов для автоматизации процесса подготовки данных.

Элементы научной новизны: новый метод оценки индивидуальной классифицирующей способности каждого показателя.

Область возможного практического применения: задачи требующие проведения классификации клиентов по собранным данным.

Технико-экономическая и социальная значимость: построение моделей, подобных рассмотренным в работе, позволит решить задачи в которых производительности классических методов классификации недостаточно.

Автор работы подтверждает, что приведенный в ней расчетно-аналитический материал правильно и объективно отражает состояние исследуемого процесса, а все заимствованные из литературных и других источников теоритические, методологические и методические положения и концепции сопровождаются ссылками на их авторов.

ABSTRACT

Term work: 112 p., 16 tables, 33 fig., 7 listings, 22 res., 11 supp.

CLASSIFICATION, CREDIT SCORING, ARTIFICIAL NEURAL NETWORKS, BACKWARD ERROR PROPAGATION ALGORITHM, MODEL VALIDATION, FEATURES SELECTION, AUC, KS, KOLMOGOROV-SMIRNOV TEST

The object of study – borrowers classification rule of JSC «Belinvestbank».

The subject of research – models of artificial neural networks attached to classification task.

Objective: fitting of nonlinear classifier for purposes of credit scoring.

Methods: computer data analysis, machine learning.

Research and development: when selecting features and transforming data for the model, a special method is used, the method uses ROC analysis; a number of tools have been prepared to automate the data preparation process.

Elements of scientific novelty: developed evaluation method of individual classifying ability of each feature.

The area of possible practical applications: tasks required classification of clients based on data.

Technical, economic and social significance: fitting models, closed to researched in work, gives an opportunity to solve problems which requires more performance than classic classification algorithms can give.

The author of the work confirms that the calculation and analytical material presented in it correctly and objectively reflects the state of the process under study, and all theoretical and methodological provisions and concepts borrowed from literary and other sources are accompanied by references to their authors.

.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Задача классификации – одна из центральный задач решаемых с помощью алгоритмов машинного обучения. Именно к задаче классификации сводиться целый ряд задач из самых разных прикладных областей: в медицине – определение заболевания по набору симптомов, в кибербезопастности – определение потенциально мошеннических сообщений, в компьютерном зрении – по набору пикселей отличать один объект от другого и, наконец, в прикладной экономической науке – классификация клиентов. В данной работе рассмотрена еще более узкая область – кредитный скоринг. Это направление которое ставит своей задачей формирование системы оценки клиентов относительно их способности выполнять обязательства перед банком.

В самом деле, в любой сфере деятельности при работе с клиентами было бы очень полезно знать наперед, как поведет себя тот или иной клиент. Свое решение предлагают статистические методы. Если обобщить и упростить, предлагается накопить большой объем данных описывающий каждый частный случай, а затем отследить как в каждом конкретном случае проявлялось исследуемое явление, а, при принятии решений о новых случаях, учитывать выводы полученные на тех данных. Применительно к кредитному скорингу это обычно реализовано так: при подаче заявки на получение кредита потенциальный кредитополучатель должен указать некоторые данные, по этим данный принимается решение о выдаче или удержании кредита, эти же данные учитываются для того, чтобы сделать новые выводы.

Конечно, можно и без накопленной информации рассматривать каждую заявку в ручную и принимать решение о выдаче или удержании кредита. Но касательно кредитов физическим лицам и, даже, малому бизнесу, в крупный банк может поступать до ста тысяч заявок за год. Не возникает сомнений, что для обработки такого объема информации потребуются просто недопустимые людские затраты. Кроме того, такая ситуация ведет к накоплению большого объема информации, что является отличной почвой для реализации статистических методов. Этим объясняется актуальность выбранной темы.

Студент-дипломник проходил преддипломную практику в ОАО «Белинвестбанк». Там, как раз, проводили плановые обновления и валидации моделей кредитного скоринга клиентов физических лиц. Потому объектом исследования станет правило классификации кредитополучателей в ОАО «Белинвестбанк».

В процессе прохождения практики были предприняты попытки построения линейного классификатора идентификационной формы логистической регрессии. Качества построенного классификатора едва хватало для того, чтобы допустить модель к работе. Было решено усложнять форму модели для преодоления линейности и перейти к классу моделей искусственных нейронных сетей. Таким образом предметом исследования станут модели искусственных нейронные сети в приложении к задаче классификации.

От сюда сформируем задачи дипломной работы. В первой главе займемся теорией моделей искусственных нейронных сетей. Перейдем от самых простых идей к более сложным по ходу усложнения поступающих задач. Раскроем чем плоха привычная линейная регрессия и как ее проблемы решаются в модели логистической регрессии. Объясним, почему модель логистической регрессии является линейным классификатором несмотря на очевидную нелинейность идентификационной формы. Введем терминологию и, в развитии идей логистической регрессии, перейдем к искусственным нейронным сетям. Подробно рассмотрим принципы работы искусственных нейронных сетей и разберемся в особенностях построения данного вида моделей.

Во второй главе познакомился с имеющейся задачей и поработаем с массивом данных полученным для построения модели. Затронем вопросы преобразования данных, очистки от выбросов и невозможных значений, борьбу с пропущенными значениями. При построении моделей лучше избавляться от неинформативных данных путем понижения размерности. Это может заметно упростить процесс построения модели и снизить вычислительную нагрузку при обучении модели. При отборе наиболее информативных критериев был использован особый метод, потому одной из задач ставим его обоснование и понятное разъяснение.

Для проведения вычислений в работе использован язык программирования python3. Относительно простой синтаксис, развитое сообщество генерирующее бесплатные инструменты, особенно применительно к обработке данных и моделированию, и доступная документация делают его идеальным кандидатом для использования в подобных задачах. Весь код приведен в работе не будет. Безусловный плюс в пользу использования языков программирования для моделирования, по сравнению с инструментами не предполагающими программирования, состоит в возможности создания абстрактных инструментов которые могут быть множество раз повторно использованы. Такие инструменты были созданы и при реализации практической части этой работы мы приведем только их и результаты их выполнения.

Располагая теоритическими выводами первой главы и набором данных, подготовленным методами, описанными во второй главе, попробуем построить соответствующую модель. Результаты найдут отражение в третьей главе. Обязательно применим и опишем опыт представленный в специальной литературе, сконцентрируемся на технических моментах связанных с pytorch – одной из самых популярных библиотек для создания моделей искусственных нейронных сетей.

Уже упоминали ОАО «Белинвестбанк» (далее Банк), это организация благодаря которой и для которой были рассмотрены данные методы. Банк является одним из ведущих банков Республики Беларусь и специализируется на кредитовании конечного потребителя, потому для него такая разработка может быть особенно полезна.

1 Торические основы моделей нейронной сети

1.1 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 K 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 $i-$го объекта из изучаемой совокупности имеется признак y_i который может принимать значения $O_1, O_2, \dots, O_K$:

$$y_i = \begin{cases} O_1, \\ O_2, \\ \dots, \\ O_K. \end{cases} \tag{1.1}$$

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_{1m}
2	y_2	x_{21}	x_{22}	…	x_{2m}
…	…	…	…	…	…
n	y_n	x_{n1}	x_{n2}	…	x_{nm}

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 $\hat{X} = \{\hat{X}_1, \hat{X}_2, \dots, \hat{X}_m\}$, наиболее точно, в некотором смысле, предсказать значения $\hat{Y}$. В литературе это правило нередко обозначается так:

$$a: X \to Y.$$

Это очень распространенная 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:

$$Y = \begin{cases} 0; \\ 1. \end{cases}$$

После оценивания будет получено регрессионное уравнение следующего вида:

$$\hat{y}_i = a + \sum_{j=1}^{m} x_{ij} b_j + \varepsilon_i,$$

где b_i – оценка коэффициента регрессии;

a – оценка свободного члена;

ε_i – случайная ошибка.

Отдельного обсуждения достойна переменная $\hat{y}_i$, 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 $\hat{y}_i \in [0; 1]$, 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 $y = 1$, значения x заметно выше, построив регрессионную прямую наглядно убеждаемся в том, что ряд наблюдений получают оценки вероятности за границами нуля и единицы.

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 $y = 1$, с заметным смещением в большую сторону по фактору x. Регрессионная прямая в таком случае сместилась и заметно хуже предсказывает вероятности

для старых наблюдений – вся группа советующая $y = 0$, получила оценки вероятностей того, что они принадлежат группе $y = 1$, 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.

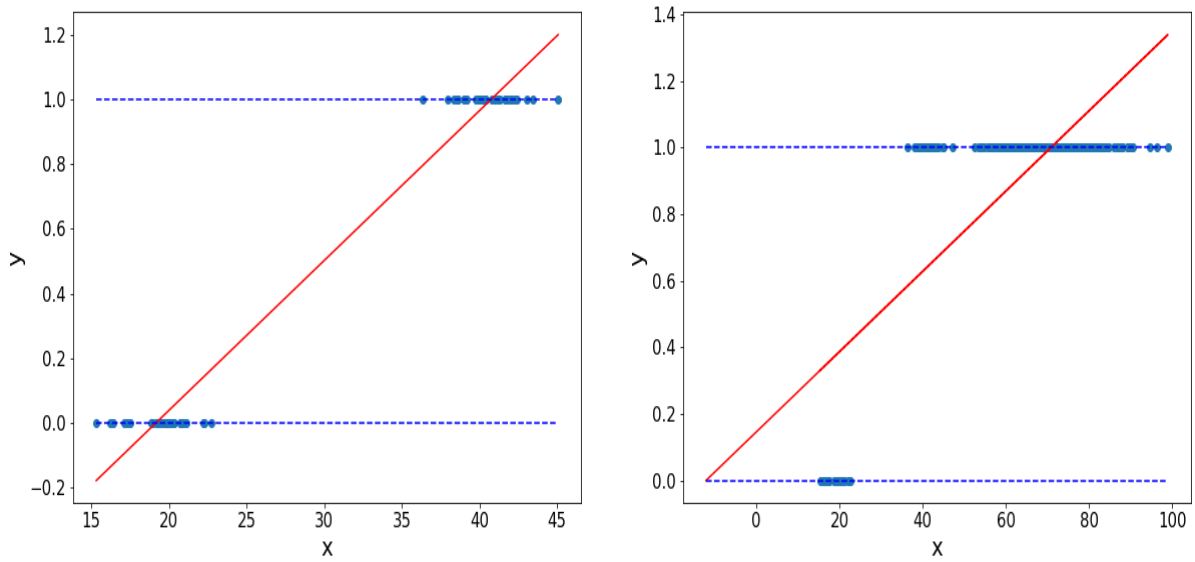

Рисунок 1.1 – Оценки вероятностей через линейную регрессионную модель
Примечание – Источник: собственная разработка.

Более того, такая модель плохо обобщается для случая, когда требуется решение для задачи не бинарной классификации. [2 с. 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 $y = 1$ как $P(y = 1)$. Одной из ключевых проблем в данном случае является то что $P(y = 1) \in [0,1]$, в то время как отклик в модели регрессионного анализа принимает значение $\hat{y}_i \in (-\infty, +\infty)$.

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

$$odds(y = 1) = \frac{P(y = 1)}{1 - P(y = 1)}.$$

Несложно показать, что $odds(y = 1) \in [0, +\infty)$. Пользуясь свойствами логарифма получим, что $\ln[odds(y = 1)] \in (-\infty, +\infty)$, 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 $y = 1$, а логарифм отношения шансов этого события. Логарифм отношения вероятностей в дальнейшем будет называть логит функцией.

Для краткости последующих записей обозначим:

$$\ln[odds(y = 1)] = y_{odds}.$$

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

$$y_{odds} = \alpha + \sum_{j=1}^{m} x_j \beta_j.$$

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

$$e^{y_{odds}} = e^{\alpha + \sum_{j=1}^{m} x_j \beta_j}.$$

Используя свойства натурального логарифма, получим:

$$\frac{P(y = 1)}{1 - P(y = 1)} = e^{\alpha + \sum_{j=1}^{m} x_j \beta_j}.$$

Решив это уравнение относительно $P(y = 1)$, получим общую запись модели логистической регрессии:

$$P(y = 1) = \frac{e^{\alpha + \sum_{j=1}^{m} x_j \beta_j}}{1 + e^{\alpha + \sum_{j=1}^{m} x_j \beta_j}}. \quad (1.2)$$

Выражение (1.2) и есть модель логистической регрессии для случая бинарной классификации[3 с. 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:

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

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 x, 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.

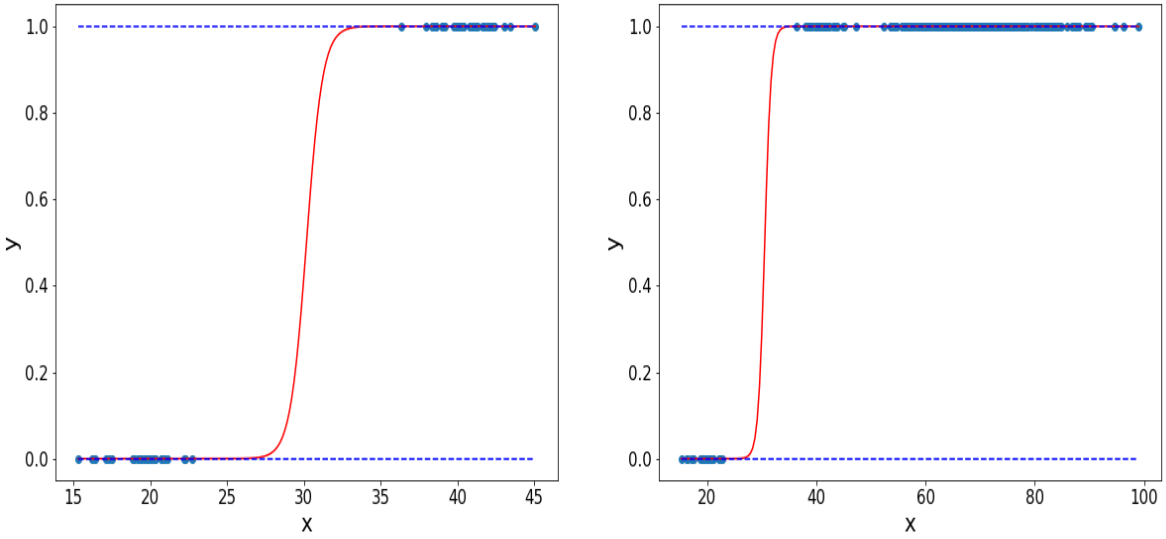

Рисунок 1.2 – Оценки вероятностей через логистическую регрессионную модель

Примечание – Источник: собственная разработка.

Рассмотрим более обобщенный 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:

$$Y = \begin{cases} 0; \\ 1; \\ \dots \\ K - 1. \end{cases}$$

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:

$$g_k(x) = \ln\left(\frac{P(y=k)}{P(y=0)}\right), k = \overline{1, K-1}. \quad (1.3)$$

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

$$g_k(x) = \alpha_k + \sum_{j=1}^{m} x_j \beta_{kj}, i = \overline{0, K-1}.$$

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:

$$e^{g_k(x)} = \frac{P(y=k)}{P(y=0)}, k = \overline{1, K-1}. \quad (1.4)$$

Теперь из одного уравнения выразим $P(y=0)$, пусть, не нарушая общности, это будет первое уравнение:

$$P(y=0) = \frac{P(y=1)}{e^{g_1(x)}}.$$

Используя определение полной вероятности, получим:

$$P(y=0) = \frac{1 - P(y=0) - \sum_{k=2}^{K-1} P(y=k)}{e^{g_1(x)}}. \quad (1.5)$$

Из оставшихся уравнений системы (1.3) получим:

$$P(y=k) = P(y=0)e^{g_k(x)}, k = \overline{2, K-1}. \quad (1.6)$$

Перенеся левую часть выражения (1.5) приведя и общему знаменателю, положив его неравным и вынеся $-P(y=0)$ за скобки, получим:

$$0 = 1 - P(y=0)\left(1 + \sum_{k=1}^{K-1} e^{g_k(x)}\right).$$

Выражая от сюда $P(y=0)$ получим:

$$P(y=0)=\frac{1}{1+\sum_{k=1}^{K-1}e^{g_k(x)}}. \quad (1.7)$$

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

$$P(y=k)=\frac{e^{g_k(x)}}{1+\sum_{j=1}^{K-1}e^{g_j(x)}}, k=\overline{1,K-1}. \quad (1.8)$$

Выражения (1.7) и (1.8) – самая общая запись логистической регрессии. Для удобства последующих записей, предполагая что $g_0(x)=0$, запишем одной формулой:

$$P(y=k)=\frac{e^{g_k(x)}}{1+\sum_{j=1}^{K-1}e^{g_j(x)}}, k=\overline{0,K-1}. \quad (1.9)$$

Сразу повторюсь, для заострения внимания – вероятность проявления каждого возможного уровня отклика имеет свою функцию от параметров и входных данных X. Заметим, что оценки коэффициентов α_k и $\beta_{kj}, k=\overline{0,K-1}, j=\overline{1,m}$, 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.

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.