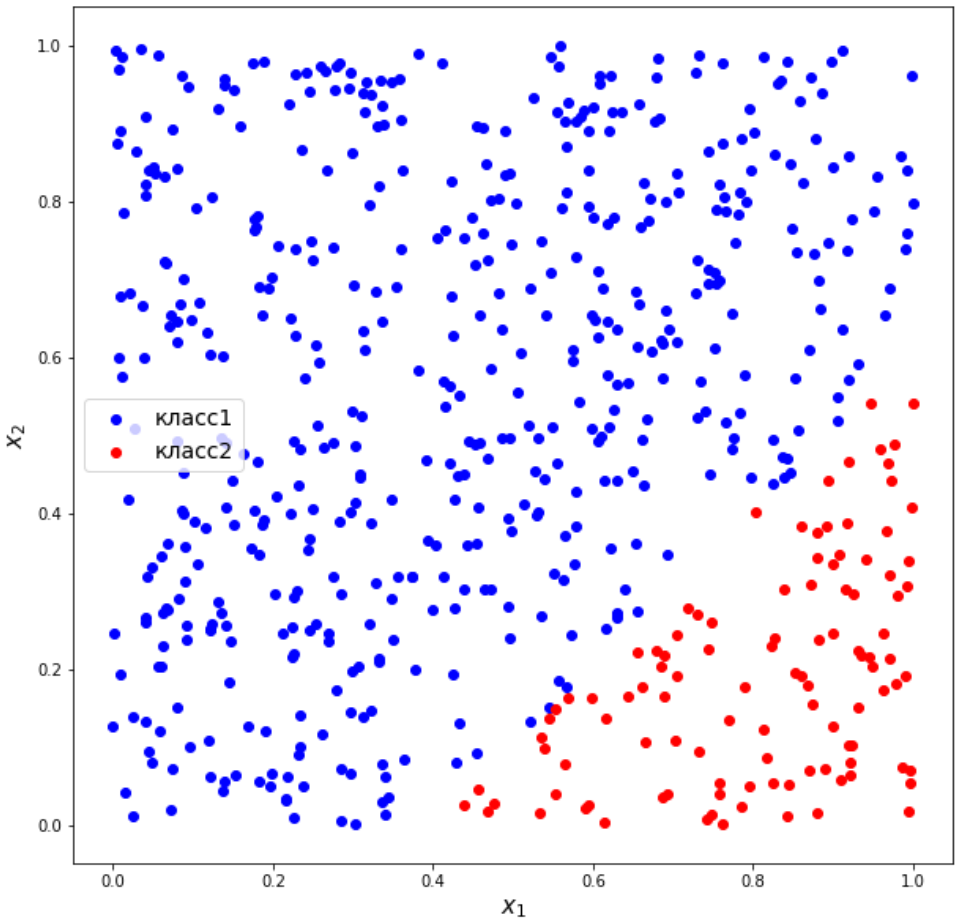

Рисунок 1.3 – Диаграмма рассеяния двумерных классифицированных данных
Примечание – Источник: собственная разработка.

Из рисунка понятно, что можно подобрать такую высоту p' 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 p':

$$p' = \frac{1}{1 + \exp(-\alpha - \sum_{j=1}^{m} x_j \beta_j)}. \quad (1.10)$$

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

$$\alpha + \sum_{j=1}^{m} x_j \beta_j = -\ln\left(\frac{1 - p'}{p'}\right).$$

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.

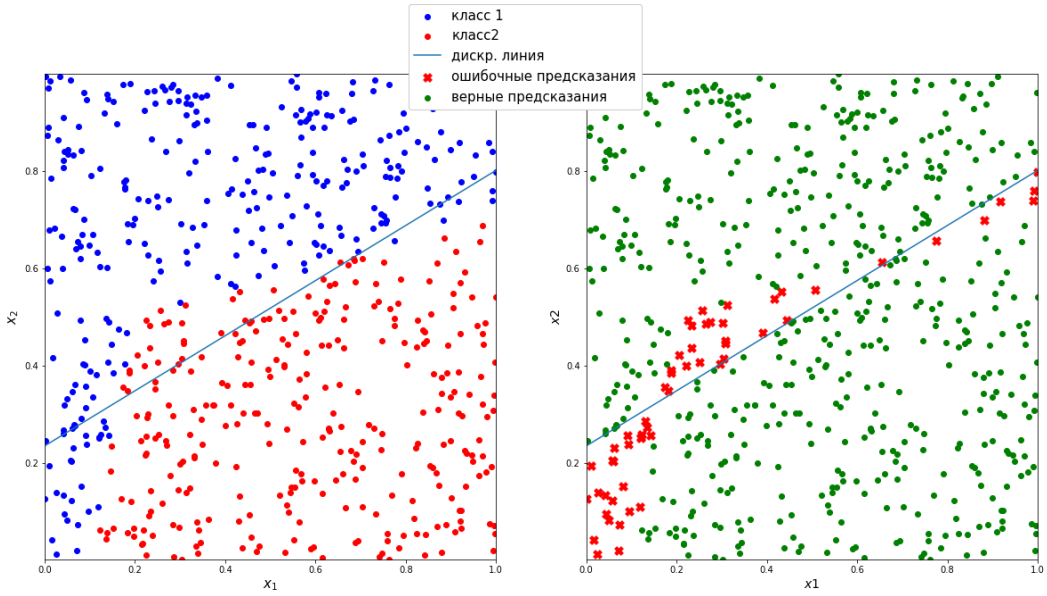

Рисунок 1.4 – Диаграмма рассеяния двумерных классифицированных данных при нелинейном принципе классификации

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 p' по принципу наибольшей разности между распределениями классов по p (этот принцип будет раскрыт в последующих главах) 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.