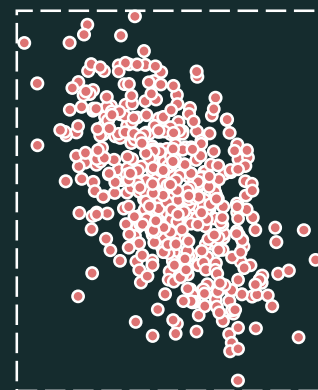
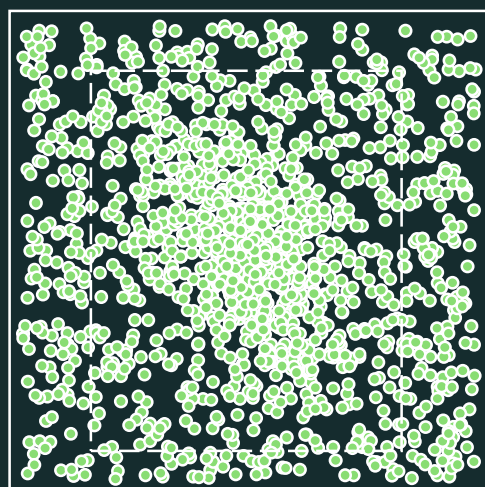


Метод унарной классификации

Исследование проводится при поддержке Академии криптографии в рамках НИР Доверие, Доверие 2



Интерактивный
визуализатор



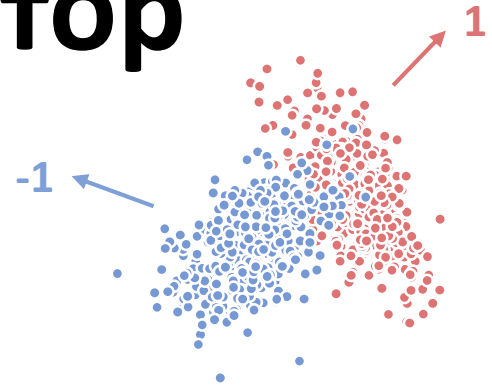
Перминов А. И.
ИСП РАН им. В. П.
Иванникова



Коваленко А. П.
Академия
криптографии

Бинарный байесовский классификатор

- $D = (X, Y)$ – случайный вектор с распределением P ;
- $X \subset [0, 1]^d, Y \in \{\pm 1\}$;
- $s: [0, 1]^d \rightarrow \{\pm 1\}$ – классификатор;
- S – носитель распределения вектора X .



Общая задача классификации:

$$P(Y \neq s(X)) \rightarrow \min_s$$

Байесовский классификатор $s^*(x)$ строится на основе **оптимальной дискриминантной функции $g(x)$** задачи классификации:

$$s^*(x) = \begin{cases} 1, & \text{if } g(x) > 0 \text{ и } x \in S \\ \text{любое из } \pm 1, & \text{if } g(x) = 0 \text{ или } x \notin S \\ -1, & \text{if } g(x) < 0 \text{ и } x \in S \end{cases}$$

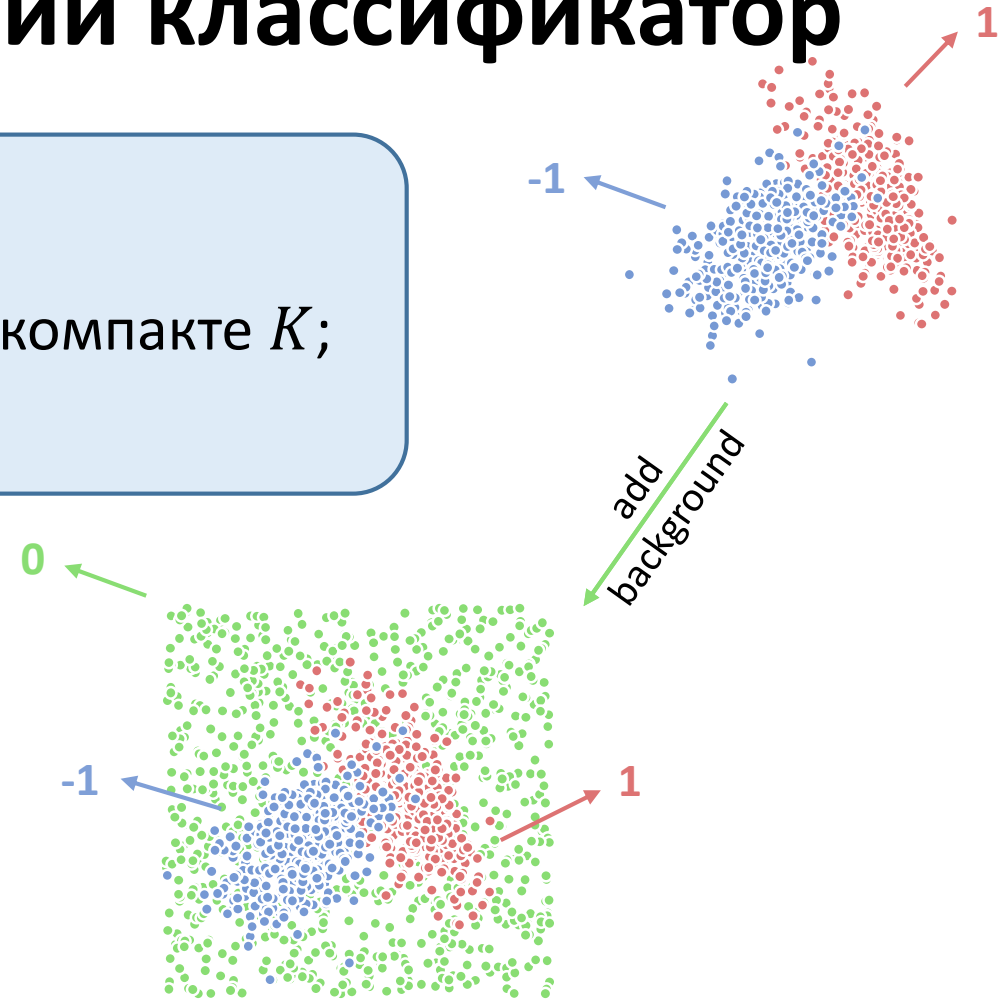
Модифицированный байесовский классификатор

Добавим t "фоновых" векторов $(\check{X}, \check{Y})$:

- $\check{X} \subset [0, 1]^d$ – равномерно распределённый шум на компакте K ;
- $\check{Y} \in \{0\}$ – метка "фонового" класса.

Смесь двух распределений:

- $(X, Y) \in [0, 1]^d \times \{\pm 1, 0\}$ – новый вектор;
- $P_\alpha = \alpha P + (1 - \alpha)\check{P}, \alpha \in (0, 1)$;
- $\check{P}$ – распределение фона.



Оптимальное решение $\check{S}^*(x)$ основанное на модифицированном наборе данных **не ухудшает** $g(x)$, но $\check{S}^*(x)$ **равна 0 вне носителя.**

Унарная классификация: смесь данных

- (X, Y) – случайная переменная;
- $X \subset [0, 1]^d$ – вектор с плотностью смеси $\alpha f(x) + (1 - \alpha)p(x)$;
- $Y = 0$ для фонового распределения и $Y = 1$ для целевого класса;
- $f(x)$ – равномерно непрерывная плотность целевого класса;
- $p(x)$ – плотность равномерного распределения на компакте K ;
- α – весовой коэффициент, $\alpha \in [0, 1)$;

Апостериорная вероятность (или регрессионная функция) Y при заданном X :

$$g(x) = P(Y = 1 | X = x) = E(Y | X = x) = \frac{\alpha f(x)}{\alpha f(x) + (1 - \alpha)p(x)}$$

Унарная классификация: решение задачи

- $g(x)$ известна (ого!):

x из распределения, если $g(x) > 0$.

- $g(x)$ неизвестна (обычно):

Строим аппроксимацию по имеющимся данным.

Унарная классификация: аппроксимация

- $c(x)$ – непрерывная функция, определенная на компакте K ;

Задача средне-квадратичной аппроксимации (минимизация идёт по всем $c(x)$):

$$c^*(x) = \operatorname{argmin} E(c(x) - Y)^2 \quad (1)$$

$$E(c(x) - Y)^2 = E(c(x) - g(x) + g(x) - Y)^2$$

$$E(c(x) - Y)^2 = E(c(x) - g(x))^2 + \underbrace{E(g(x) - Y)^2}_{\text{не зависит от } c(x)}$$

Задача **сводится** к аппроксимации функции регрессии:

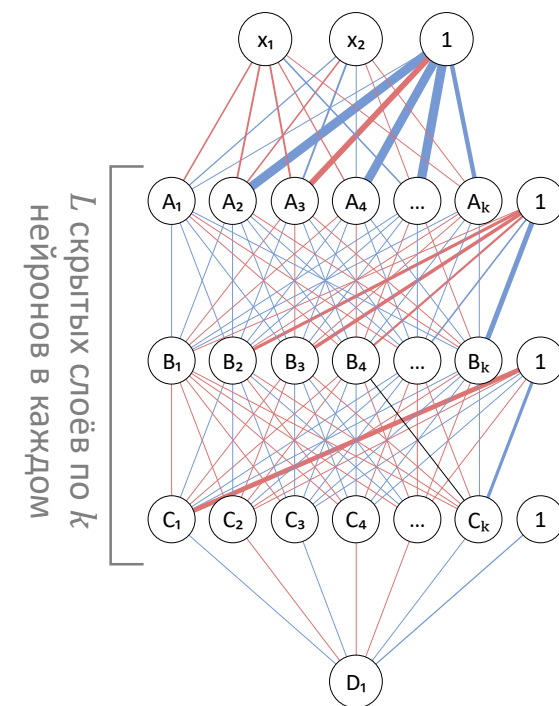
$$c^*(x) = \operatorname{argmin} E(c(x) - g(x))^2 \quad (2)$$

Унарная классификация: персептрон

- $c(x)$ – многослойный персептрон с L скрытыми слоями по k нейронов в каждом, одним нейроном в выходном слое и кусочно-линейной функцией активации (Abs , $ReLU$, ...).

Универсальная теорема аппроксимации: для любого $\epsilon > 0$ существуют такие k и L , что для любого $x \in K$:

$$\sup |c(x) - g(x)| < \epsilon$$



Таким образом, ϵ -приближенное решение **(1)** теоретически существует.

Унарная классификация: персептрон

- $\{X_i, 1\}_{i=1}^n$ – размеченный набор из n наблюдений с целевой плотностью $f(x)$;
- $\{X_j, 0\}_{j=n+1}^{n+m}$ – m добавленных **фоновых** наблюдений, $m = n \cdot \frac{1-\alpha}{\alpha}$;
- $C(L, k)$ – класс **MLP** с активацией **Abs** и заданными L, k ;

Оптимизационная задача (минимизация идёт по всем $c_n(x) \in C(L, k)$):

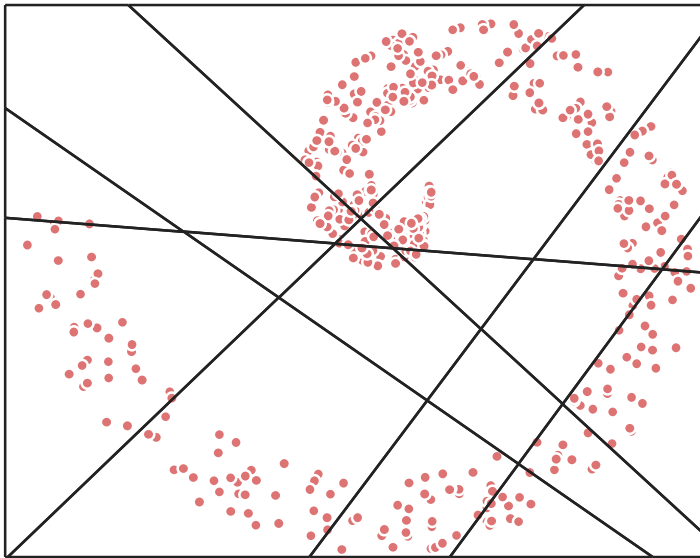
$$\sum_{i=1}^{n+m} (c_n(X_i) - Y_i)^2 \rightarrow \min \quad (3)$$

Пусть $c_n^*(x)$ – решение (3), называемое **нейросетевой функцией регрессии**.

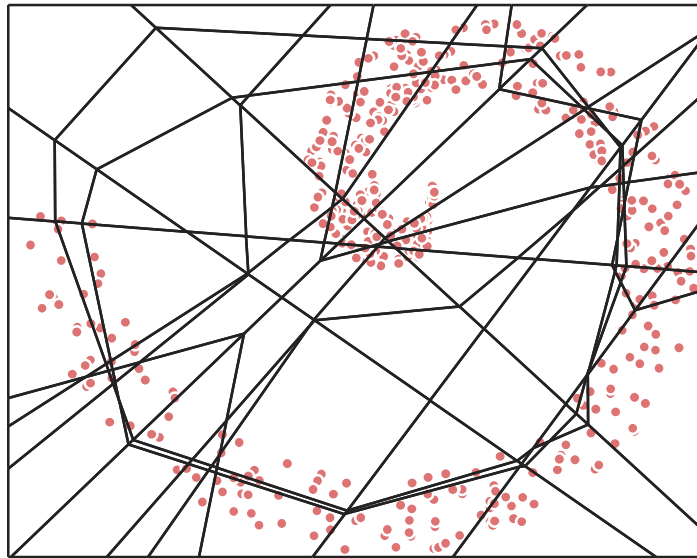
Унарная классификация: адаптивная гистограмма

- $c_n^*(x)$ строит иерархическое разбиение K на N непересекающихся ячеек: $K = \{K_1, K_2, \dots, K_N\}$

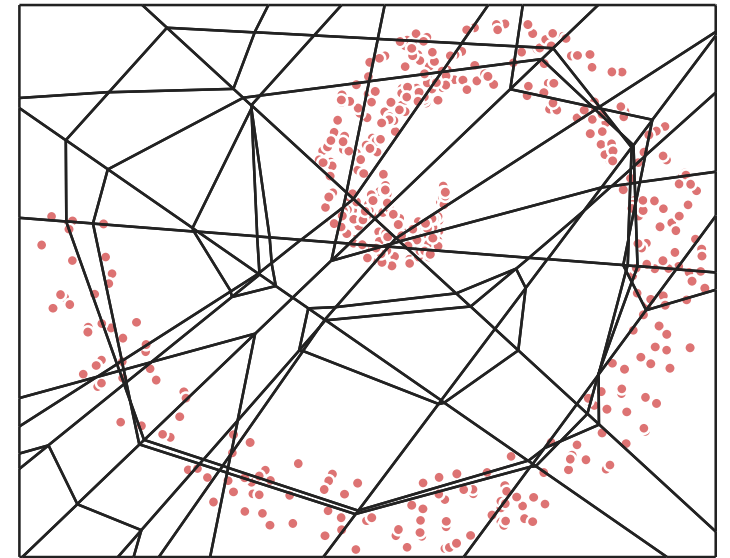
Пример разбиения некоторым MLP с $L = 2$, $k = 6$ и Abs активацией:



Ячейки первого слоя:
18 ячеек



Ячейки второго слоя:
121 ячейка



Ячейки выходного нейрона:
148 ячеек

Унарная классификация: адаптивная гистограмма

- $h_n(x)$ – **кусочно-постоянная** функция гистограммной регрессии, определённая путём минимизации:

$$\sum_{i=1}^{n+m} (h_n(X_i) - Y_i)^2 \rightarrow \min \quad (4)$$

Минимизация осуществляется по всем **кусочно-постоянным** функциям, определённым на ячейках K_r компакта K .

- Пусть $X \in K_r \rightarrow$ для каждой ячейки K_r задача (4) сводится к

$$n_1(X) \cdot (h_n(X) - 1)^2 + n_0(X) \cdot (h_n(X) - 0)^2 \rightarrow \min$$

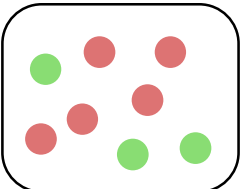
- $n_j(X) = \sum_{i=1}^{n+m} I_{\{X_i \in K_r, Y_i = j\}};$

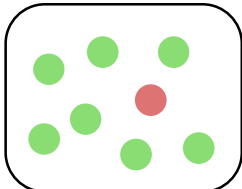
Унарная классификация: адаптивная гистограмма

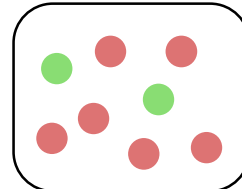
Дифференцируя по $h_n(x)$ получаем решение (4):

$$h_n^*(x) = \frac{n_1(X)}{n_1(X) + n_0(X)} = \frac{f_n(X)}{f_n(X) + \frac{1-\alpha}{\alpha} p_n(X)}$$

- $f_n(X) = \frac{n_1(X)}{n \cdot V(K_r)}$ – гистограммная оценка плотности $f(x)$ в ячейке K_r ;
- $p_n(X) = \frac{n_0(X)}{n \cdot V(K_r)}$ – гистограммная оценка равномерной плотности $p(x)$ в ячейке K_r ;
- $V(K_r)$ – мера ячейки K_r .


$$\begin{aligned} n_1 &= 5, n_0 = 3 \\ h_n^*(X) &= \frac{5}{5+3} = 0.625 \end{aligned}$$


$$\begin{aligned} n_1 &= 1, n_0 = 7 \\ h_n^*(X) &= \frac{1}{1+7} = 0.125 \end{aligned}$$

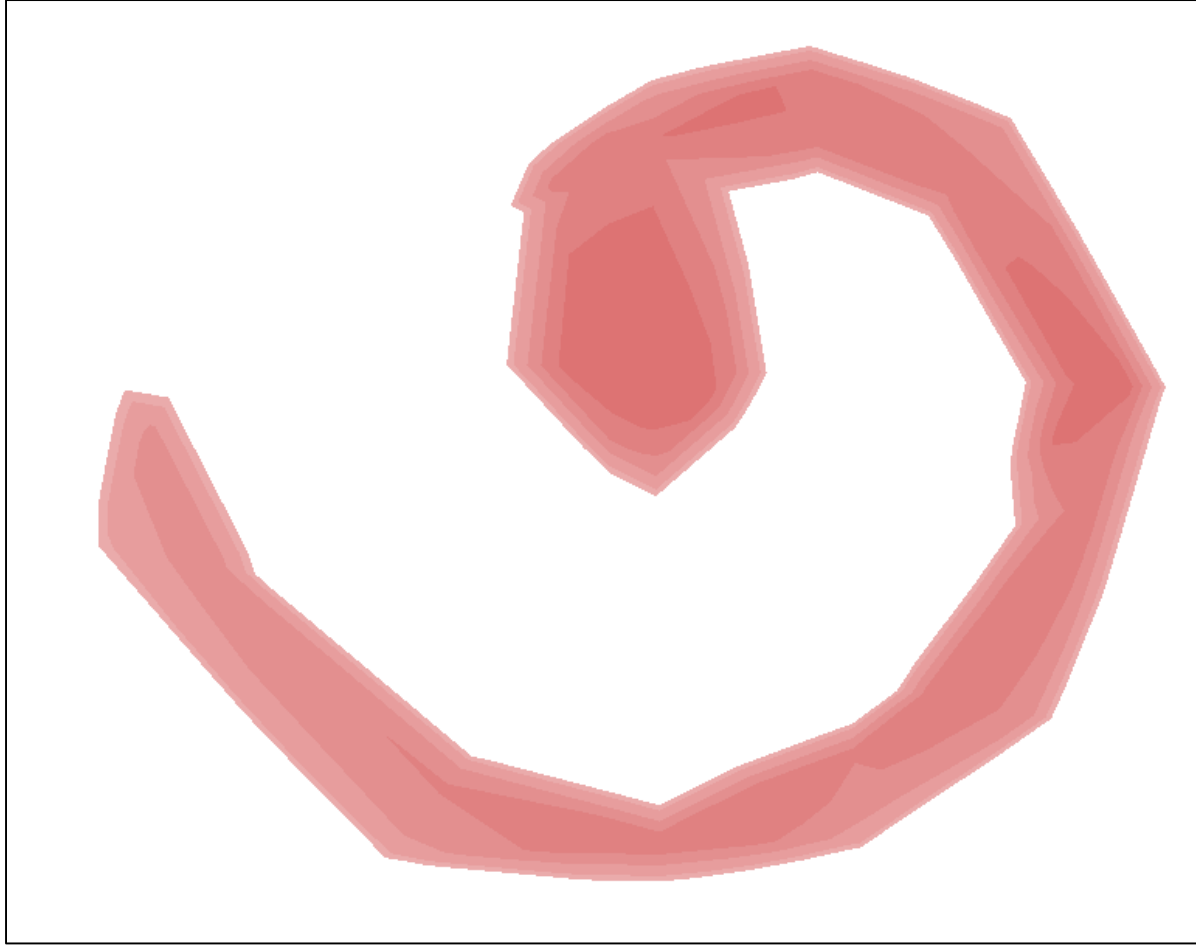

$$\begin{aligned} n_1 &= 6, n_0 = 2 \\ h_n^*(X) &= \frac{6}{6+2} = 0.75 \end{aligned}$$

Унарная классификация: состоятельность

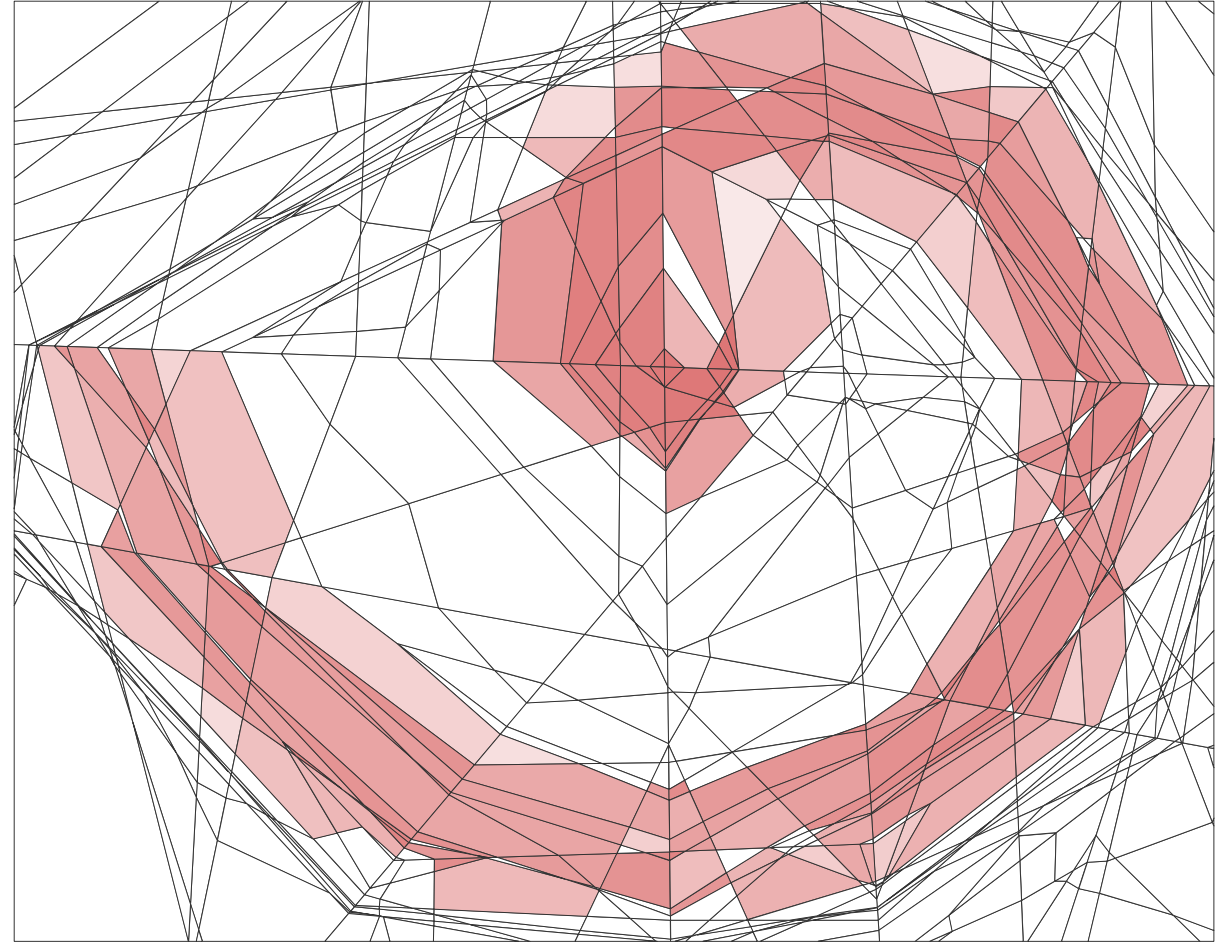
- **Уменьшение ячеек:** с уменьшением диаметра ячейки в ней должно сохраняться достаточное количество точек;
- **Практика:** для $d = 10, k = 10, L = 2, N > 10^4$ – нужны **миллионы** точек фона;
- **Заполненные ячейки** (есть целевые и фоновые точки): $h_n^*(X) \approx c_n^*(X)$;
- **Фоновые ячейки:** $h_n^*(X) = 0$, $c_n^*(X)$ интерполируется (непрерывно);
- **Области высокой плотности:** $c_n^*(X) \gg 0$;
- **Области с низкой плотностью:** $c_n^*(X) \rightarrow 0$;

При достаточно большом n и правильно выбранных k и L , нейросетевая функция регрессии $c_n^*(X)$ **может интерпретироваться** как состоятельная оценка $g(X)$.

Гистограммная и нейросетевая регрессия



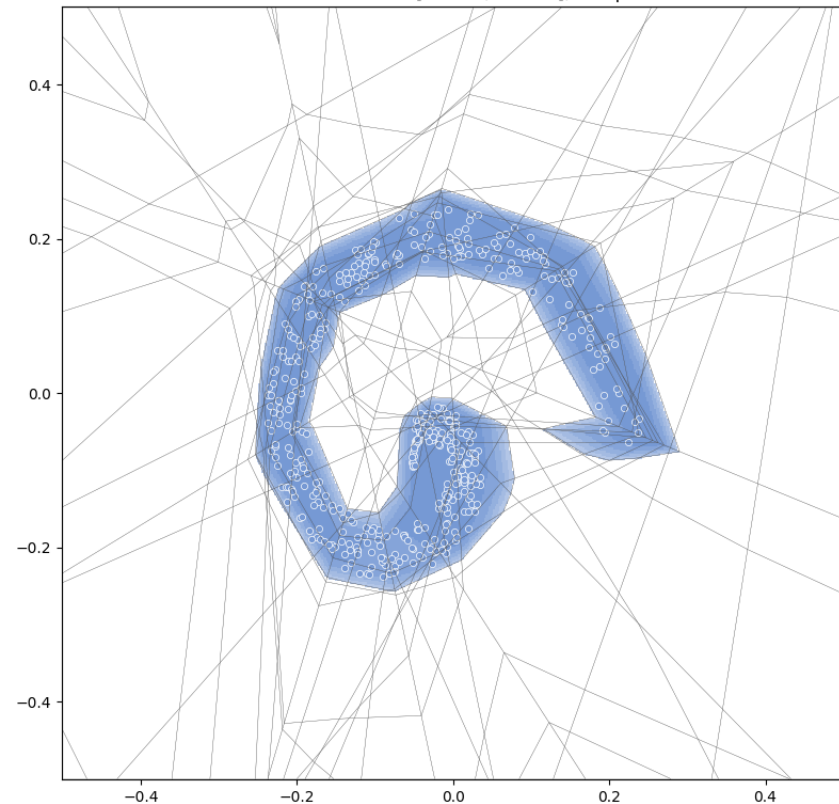
Значения $c_n^*(x)$



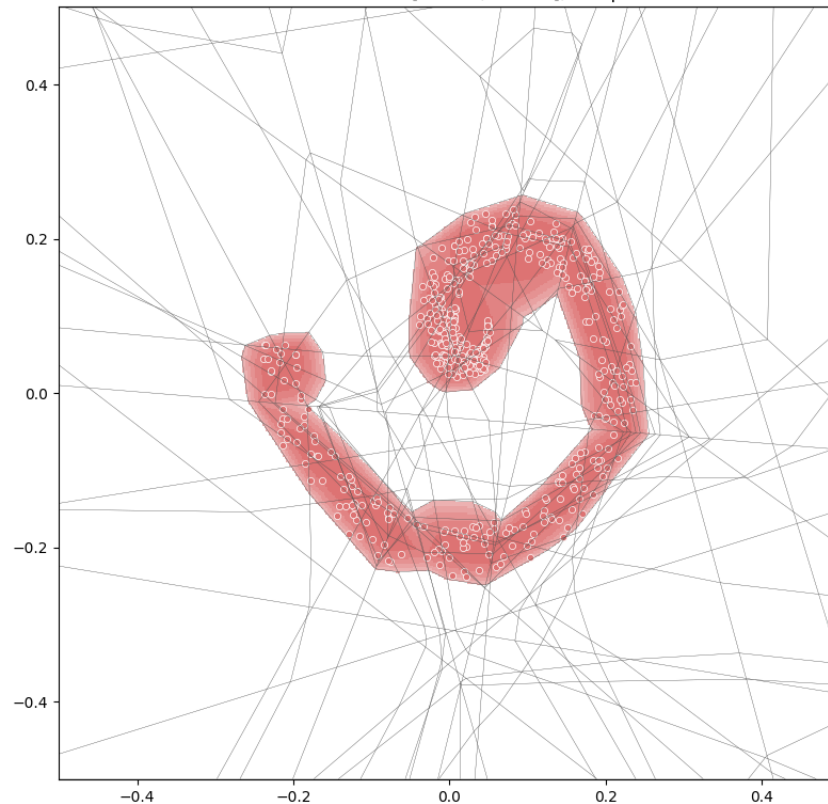
Значения $h_n^*(x)$

Унарная классификация: пример 1

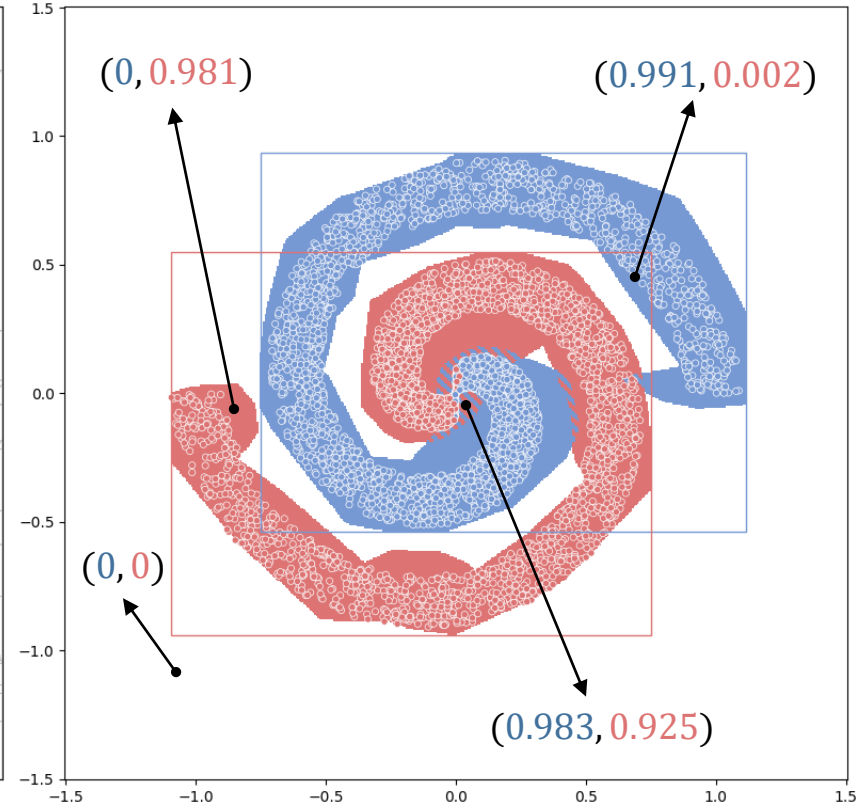
Model for class "0": [0.0000, 0.9912], 500 points



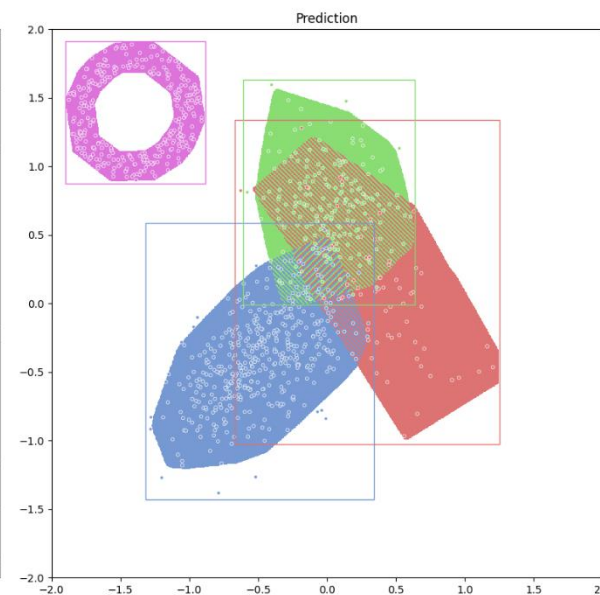
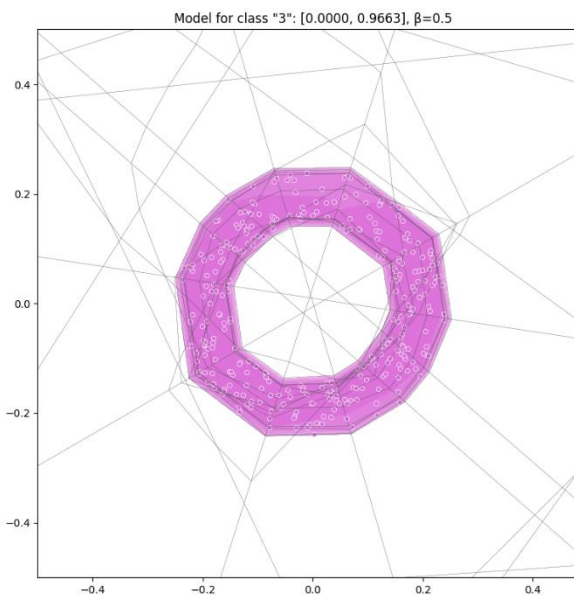
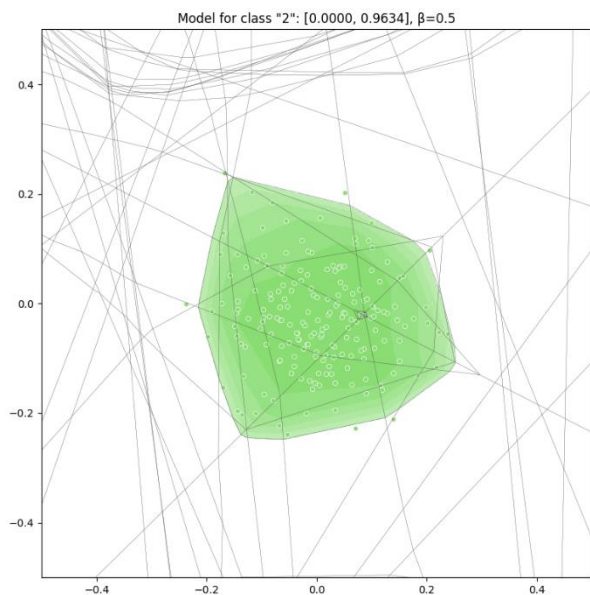
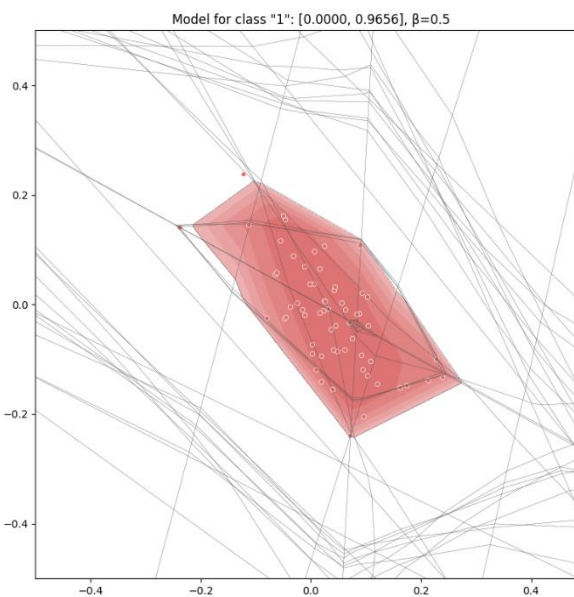
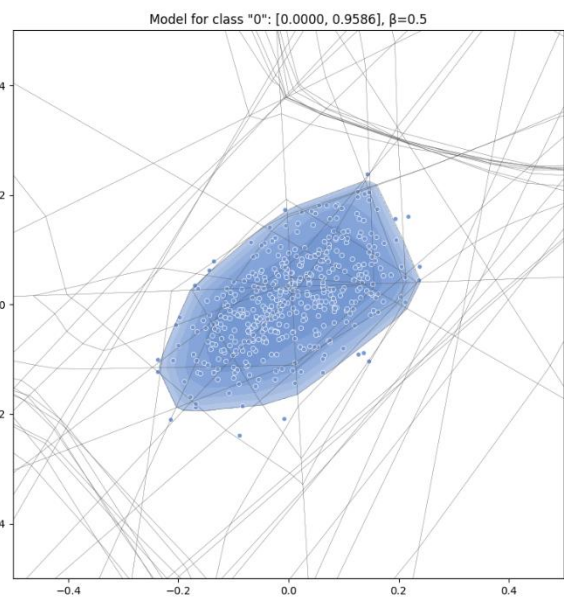
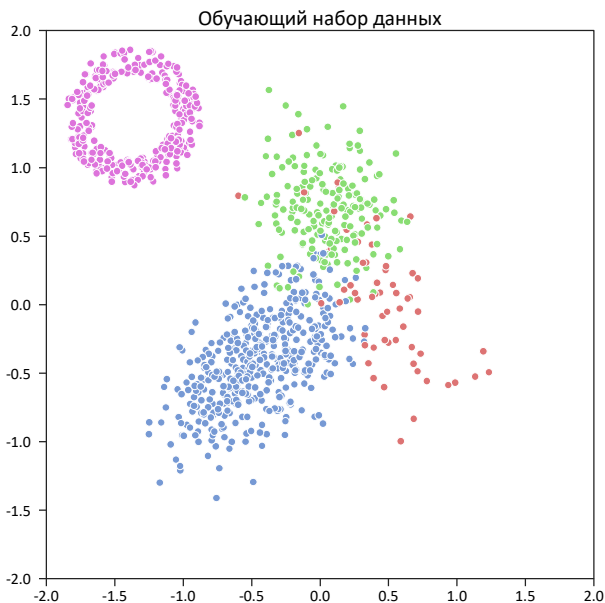
Model for class "1": [0.0000, 0.9895], 500 points



Classifier

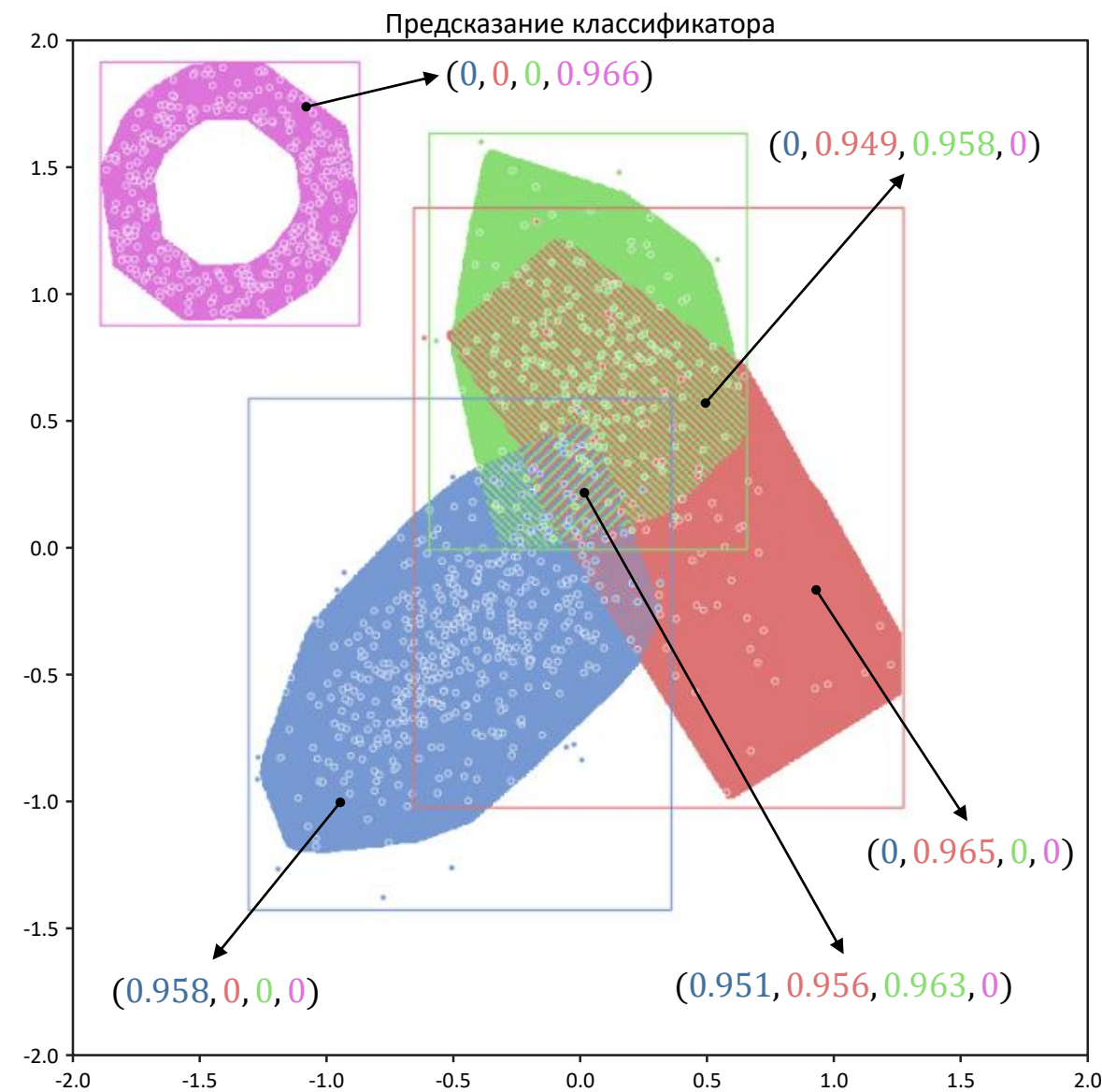
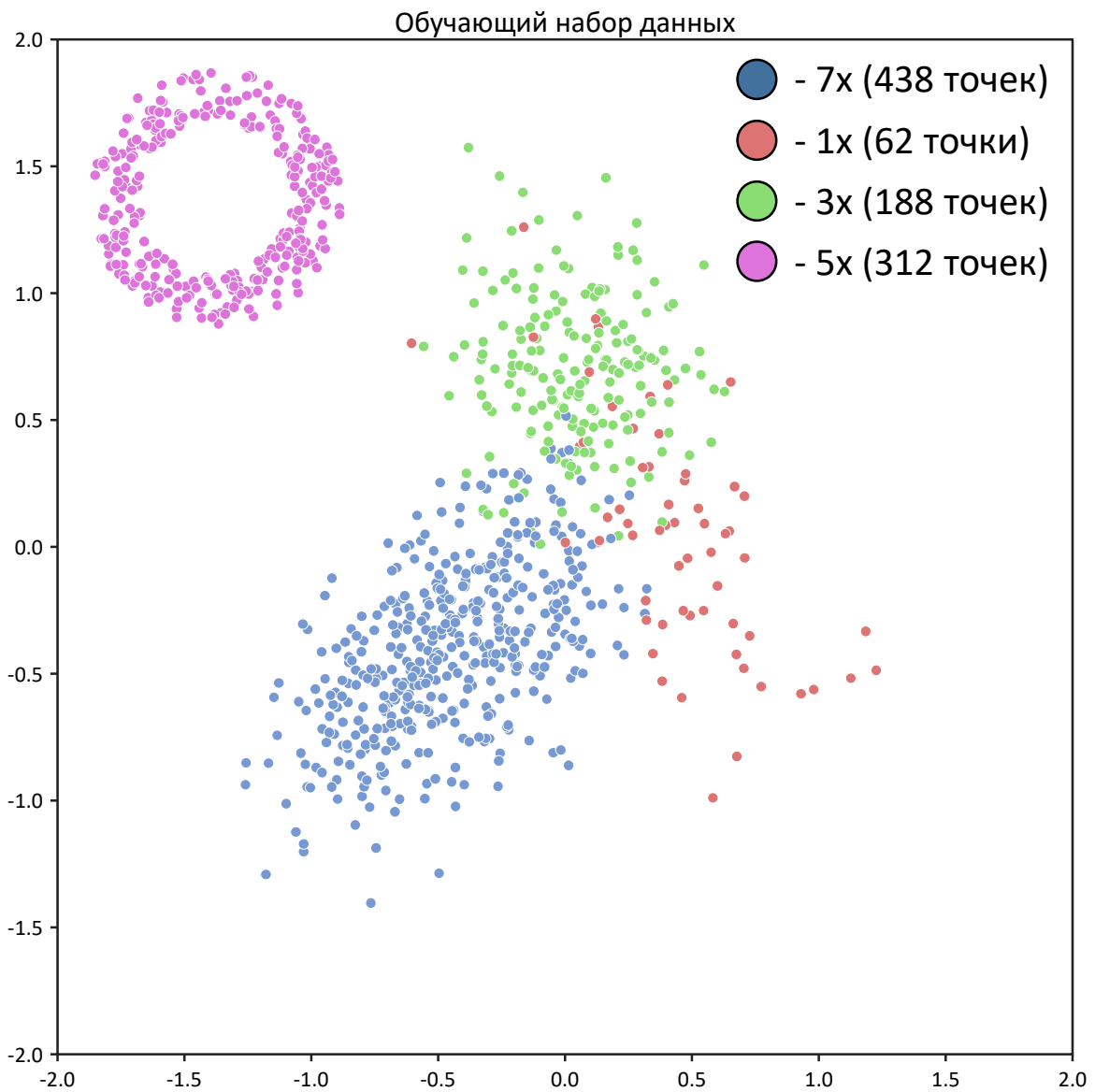


Унарная классификация: пример 2



- - 7x (438 точек)
- - 1x (62 точки)
- - 3x (188 точек)
- - 5x (312 точек)

Унарная классификация: пример 2



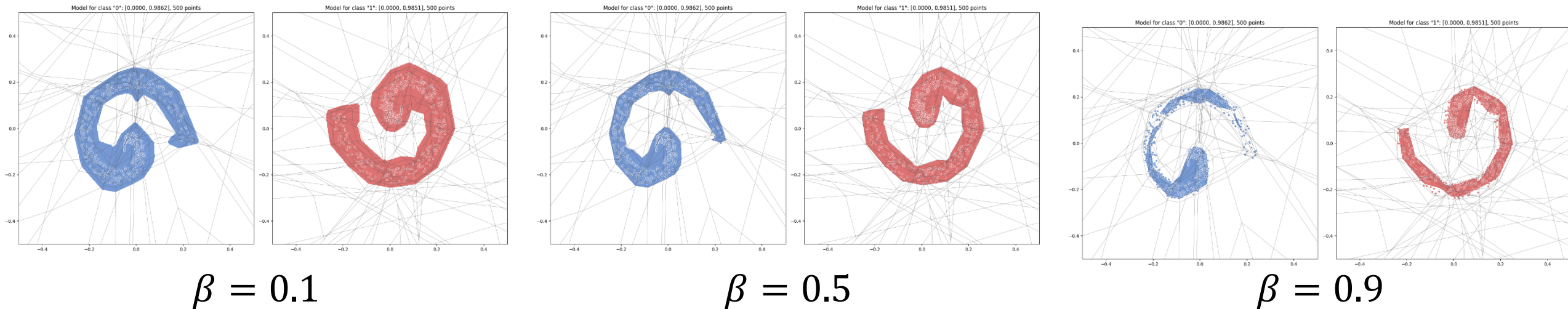
Порог доверия

Порог $\beta \in [0, 1)$: определяет области, в которых выход классификатора ненадёжен;

Непрерывность: MLP не может мгновенно прижаться к нулевой плоскости вне носителя;

Отказ от классификации: не принимаем решение, если $c_n^*(x) < \beta$;

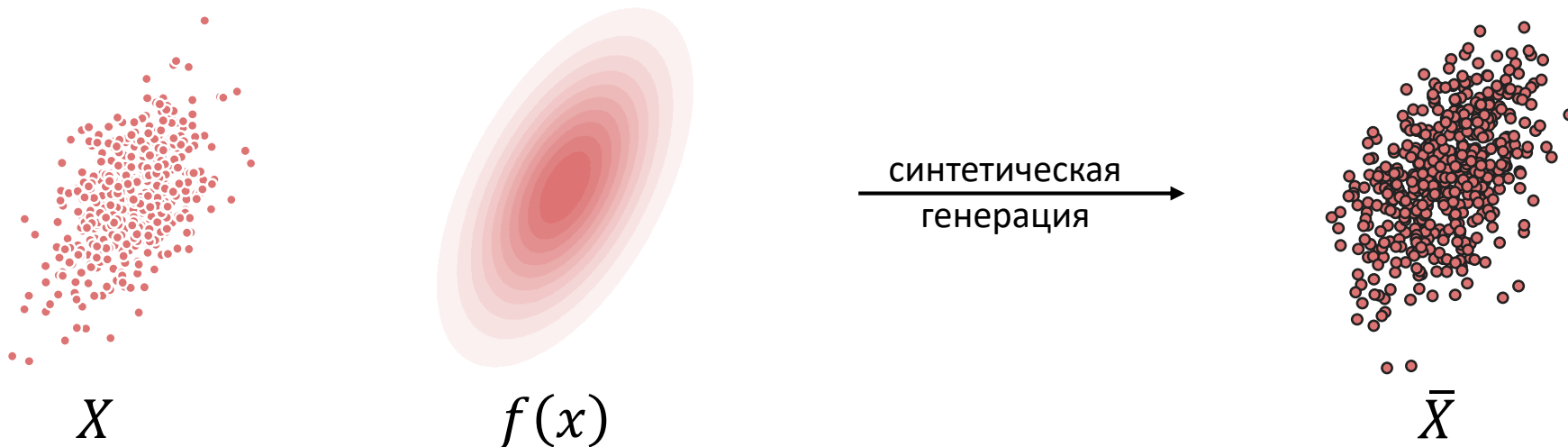
Влияние: чем выше порог, тем больше отказ, но и выше доверие к решению.



Генерация синтетических данных

- $X = \{x_1, x_2, \dots, x_n\}, x_i \in [0,1]^d$ – набор данных из неизвестного распределения с равномерно непрерывной плотностью $f(x)$;
- $\bar{X} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m\}$ – синтетический набор данных.

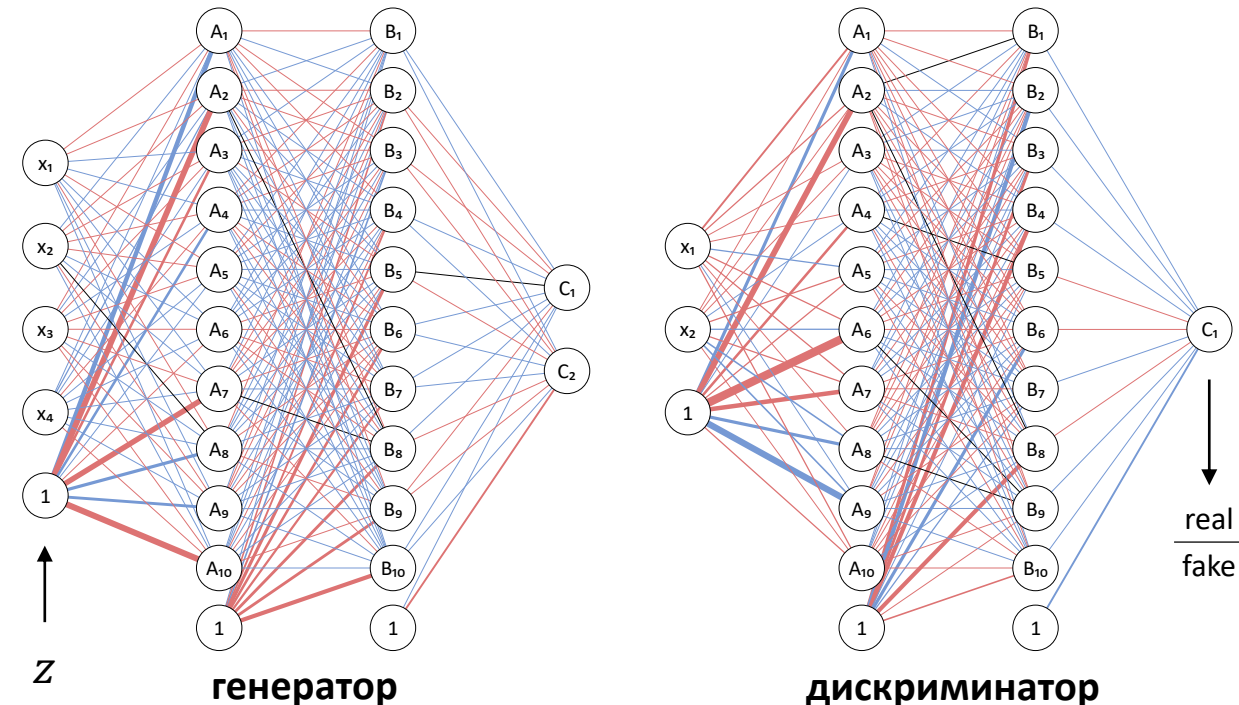
Цель: получить новый набор данных $\bar{X}$, который сохраняет статистические и структурные свойства X .



Создание синтетических данных: подходы

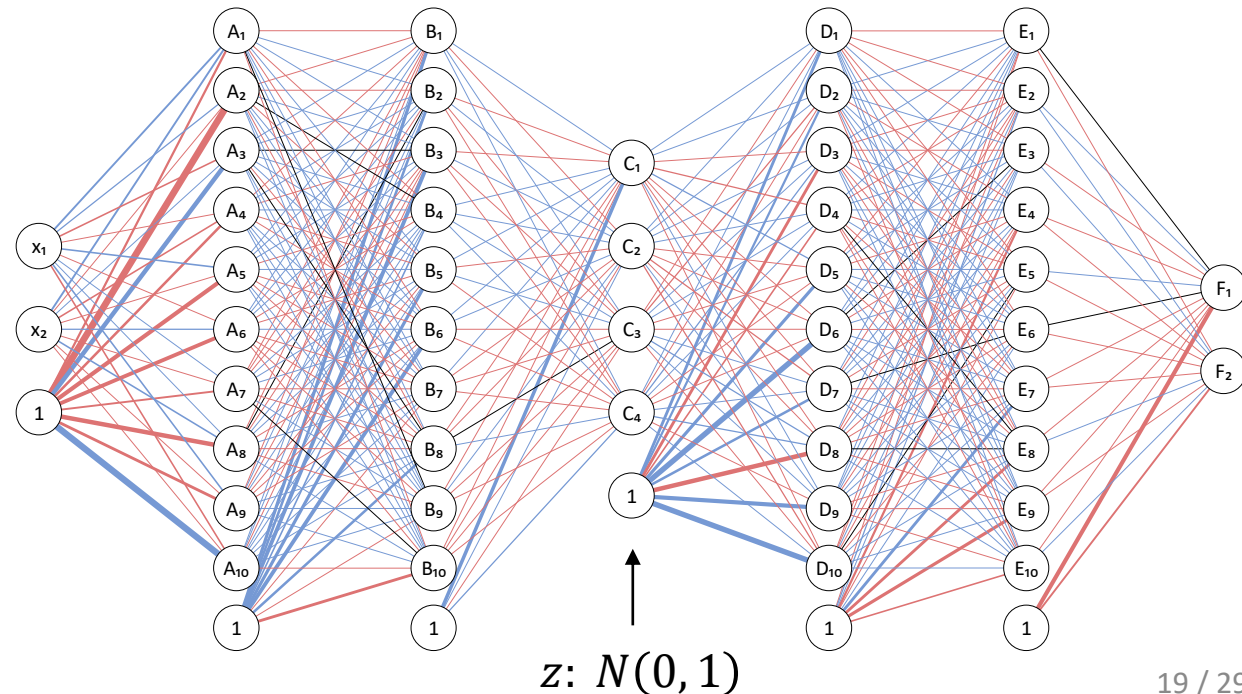
Генеративно-состязательная сеть

- **Генератор:** учится переводить входной нормальный шум в реальные данные;
- **Дискриминатор:** определяет, сгенерированы ли данные (учится отличать реальные данные от синтетических, используя бинарную классификацию).



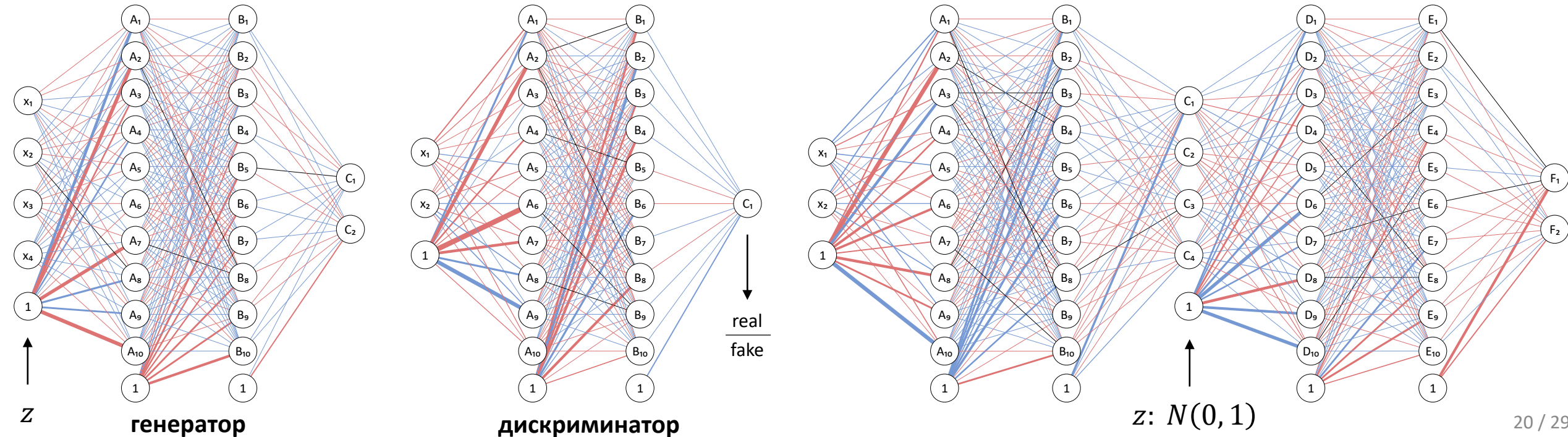
Вариационный кодировщик

- **Архитектура:** учится восстанавливать на выходе входные данные, превращая скрытое пространство в нормальное;
- **Генерация:** после обучения остается только часть, переводящая нормальное распределение в целевое распределение.



Создание синтетических данных: проблемы

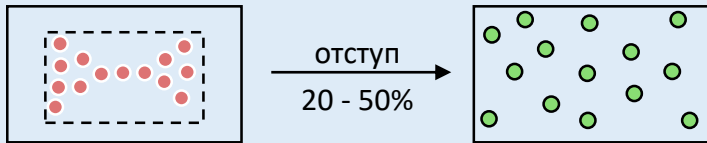
- **Обучение:** низкая скорость сходимости GAN, смещение к среднему у VAE
- **Вероятностные оценки:** отсутствуют
- **Статистические свойства:** сохранение не гарантируется
- **Структурные свойства:** сохранение не гарантируется



Получение репродуцированных данных

Генерация фоновых точек

- **Компакт:** $K \subset [0,1]^d$ – выровненный по осям гиперкуб, выходящий на 20-50% за пределы реальных данных в каждом измерении);
- **Отступ:** обеспечивает достаточное количество отрицательных образцов для классификатора, чтобы распознать носитель и помогает прижать функцию принятия решения к нулевой плоскости.
- **Набор точек:** $B = \{b_1, b_2, \dots, b_n\}$ генерируются равномерно на компакте K , $|B| = |X|$ ($\alpha = 0.5$)



Обучение классификатора

- **Классификатор:** $c_n(x): [0,1]^d \rightarrow [0,1]$;
- **Набор данных:** $X \cup B$ с бинарными метками:

$$c_n(x) = \begin{cases} 1, & x \in X \\ 0, & x \in B \end{cases}$$

- **Функция потерь:** MSE

$$L = \sum_{x \in X} (1 - c_n(x))^2 + \sum_{b \in B} (0 - c_n(b))^2$$

- **Динамический фон:** каждую эпоху точки B генерируются заново;

Почему MSE?

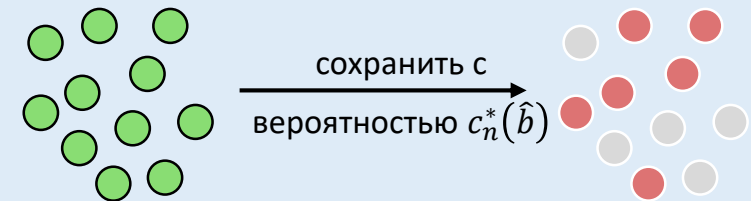
- Более плавная аппроксимация апостериорных вероятностей без необходимости сигмоиды;
- Поддерживает интерпретацию на основе гистограммы.

Репродуцирование

- **Точки-кандидаты:** генерируем $\hat{B} = \{\hat{b}_1, \hat{b}_2, \dots\}$ равномерно на K .
- **Предсказание:** получаем значение $c_n^*(\hat{b})$ для каждой точки $\hat{b} \in \hat{B}$
- **Правило сохранения:** точка $\hat{b}$ остаётся с вероятностью $c_n^*(\hat{b})$:

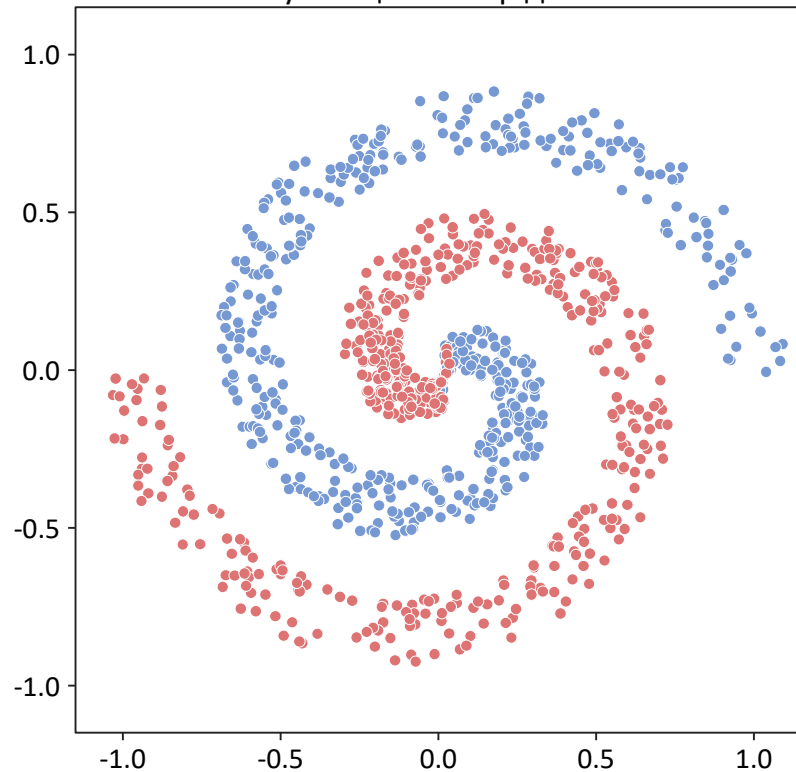
$$\hat{X} = \{\hat{b} \in \hat{B} \mid \xi < c_n^*(\hat{b})\}$$

ξ – случайная переменная из равномерного распределения $U(0,1)$

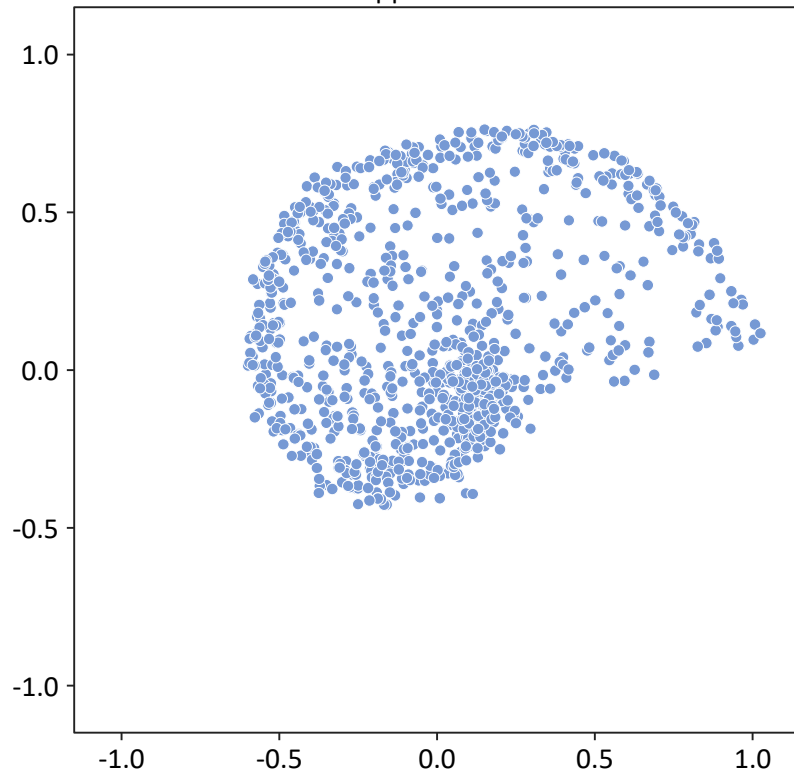


Создание синтетических данных: VAE

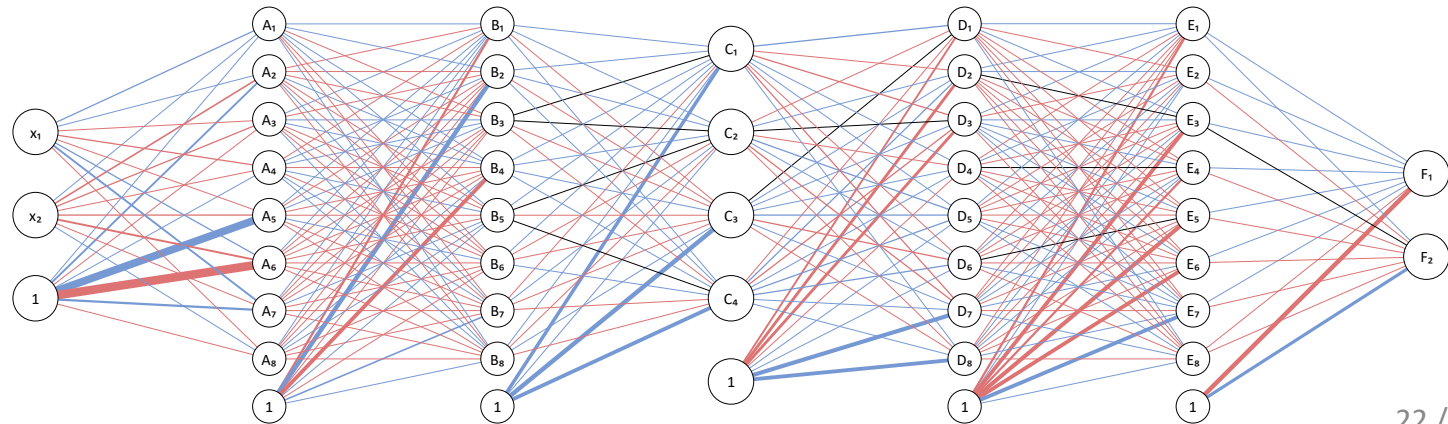
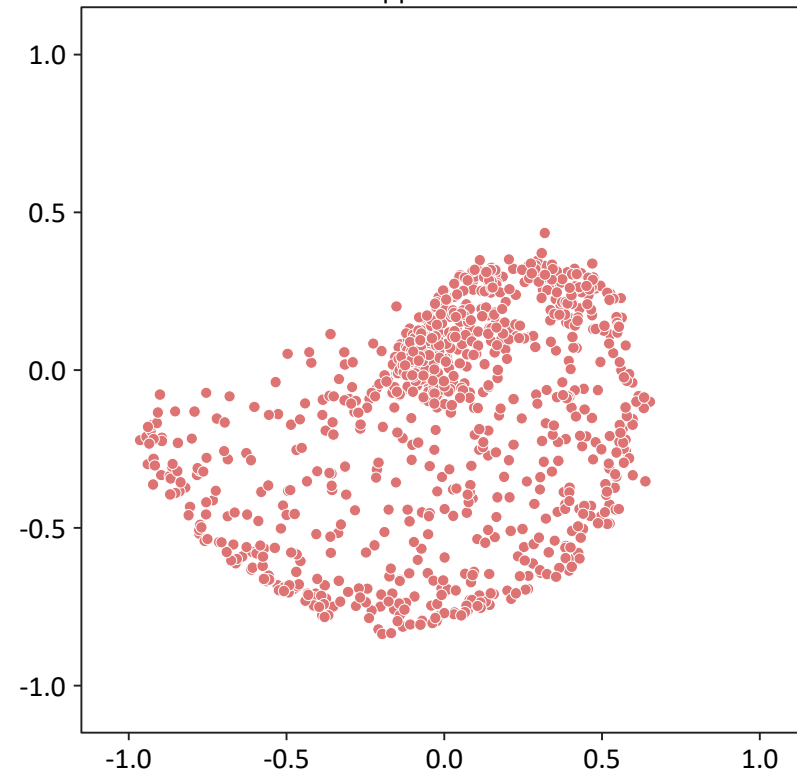
Обучающий набор данных



VAE для класса 0



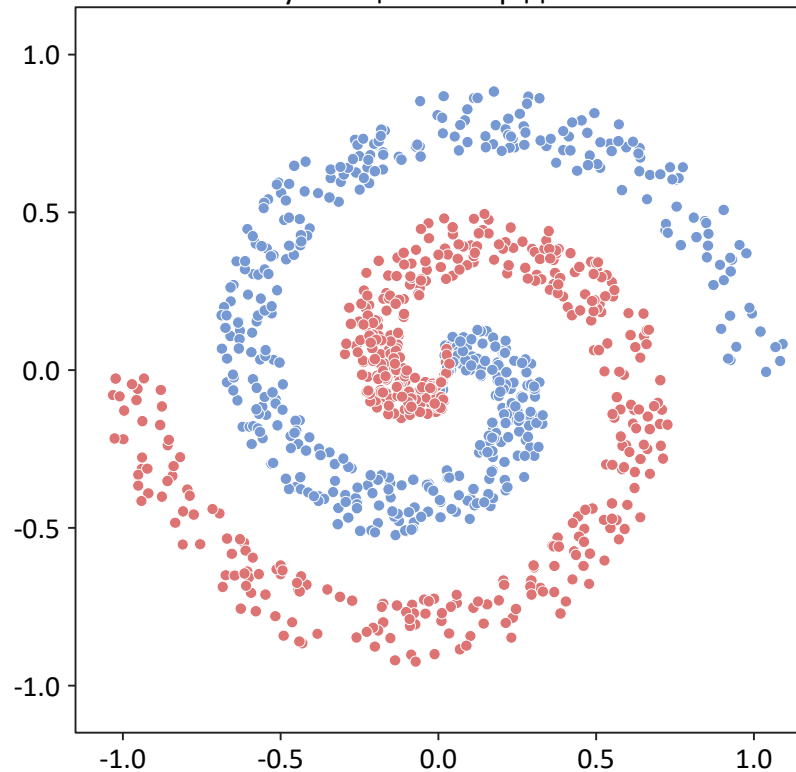
VAE для класса 1



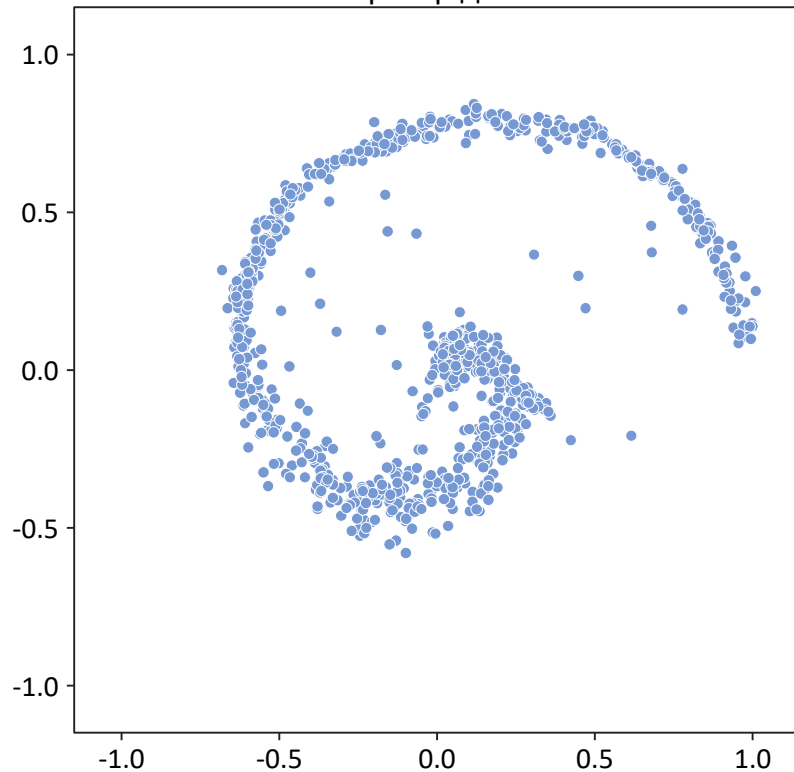
Модель: 2-20-20-4-20-20-2
Генератор: 4-20-20-2

Создание синтетических данных: GAN

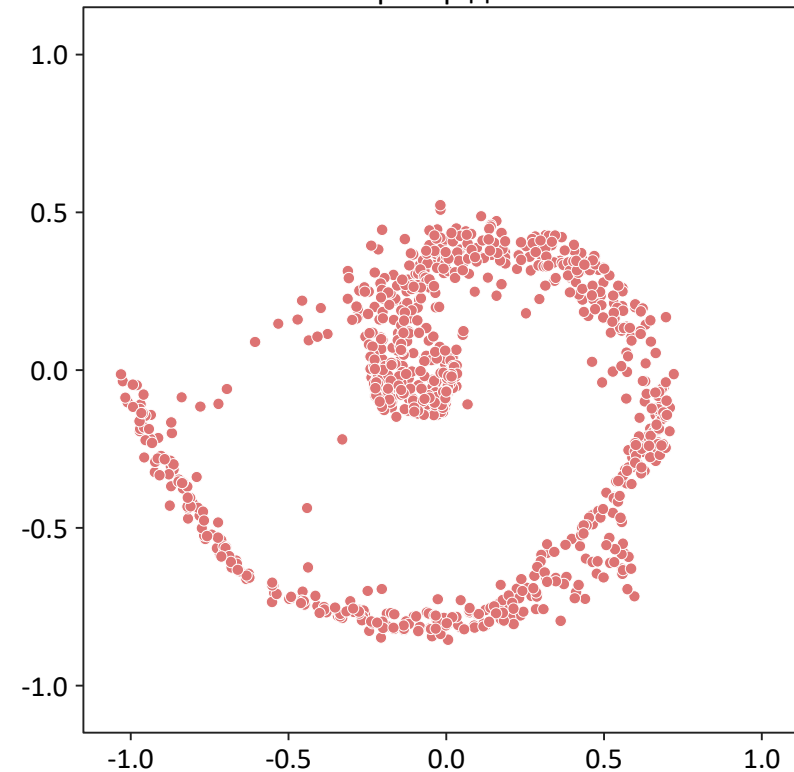
Обучающий набор данных



GAN генератор для класса 0

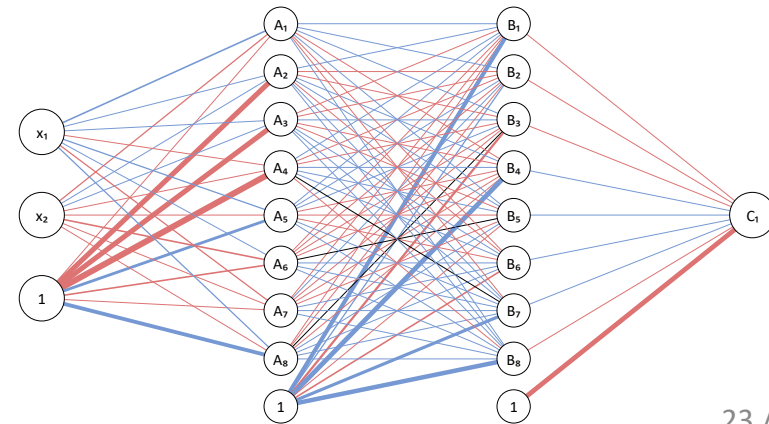
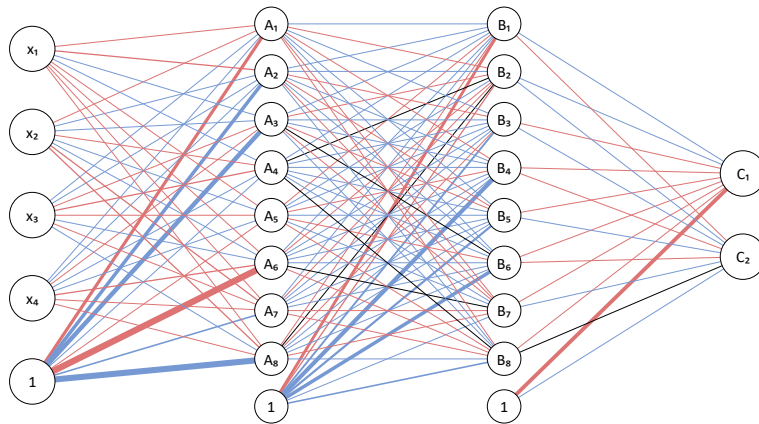


GAN генератор для класса 1



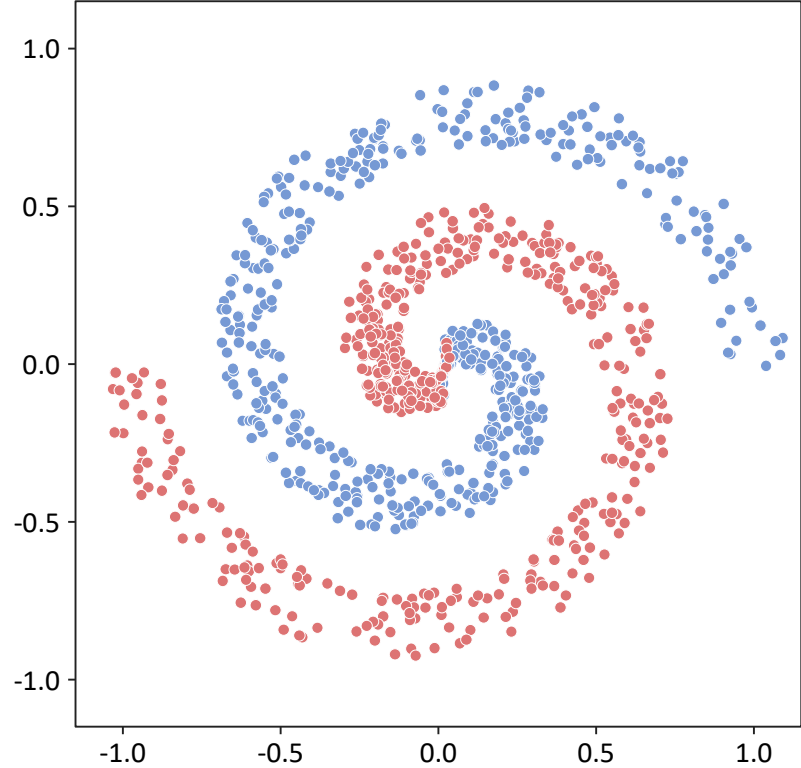
Генератор: 4-20-20-2

Дискриминатор: 2-20-20-1

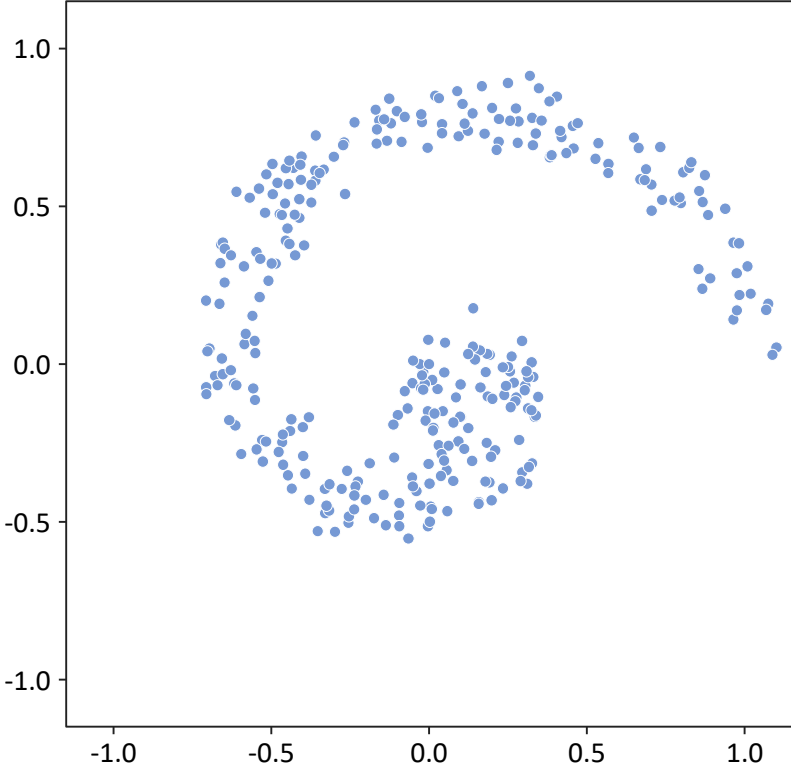


Создание синтетических данных: репродукция

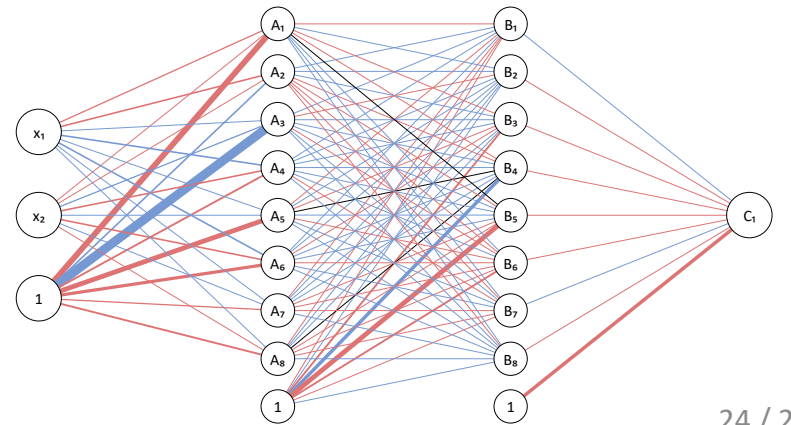
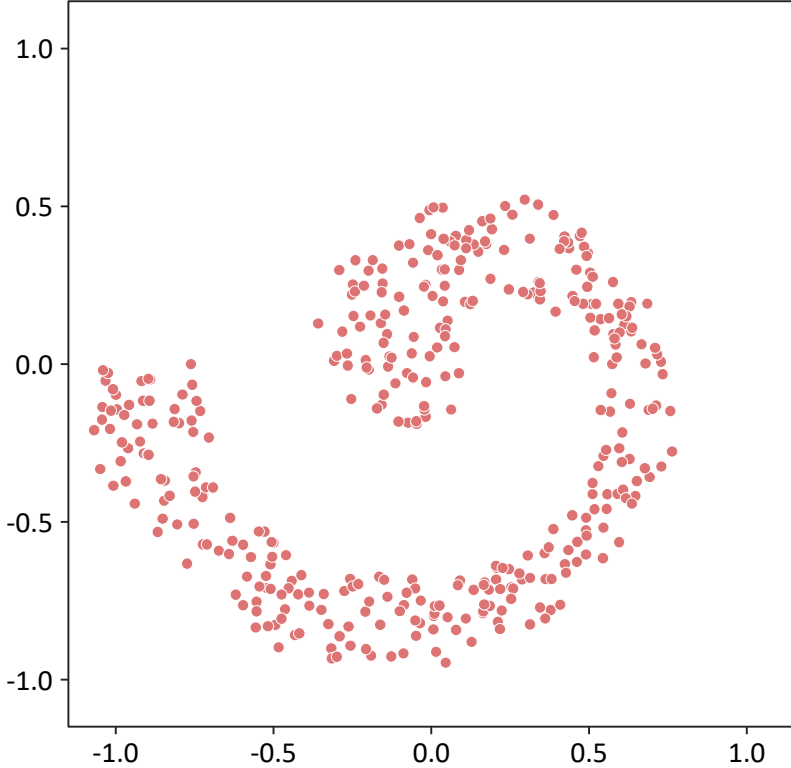
Обучающий набор данных



Репродукция для класса 0

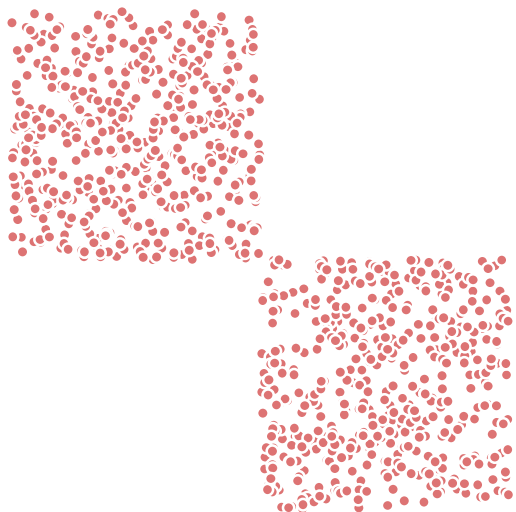


Репродукция для класса 1

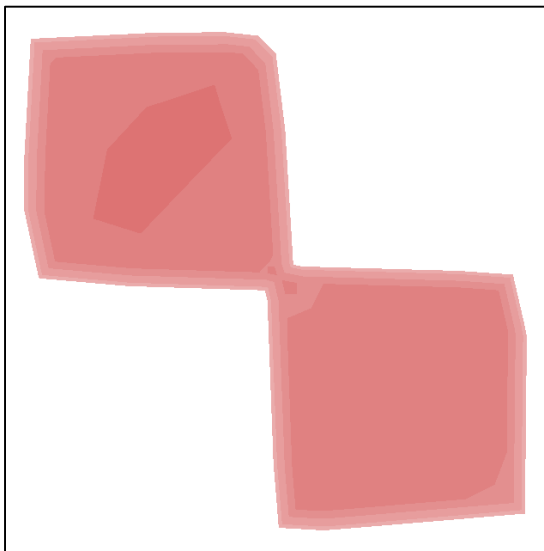


Модель: 2-20-20-1

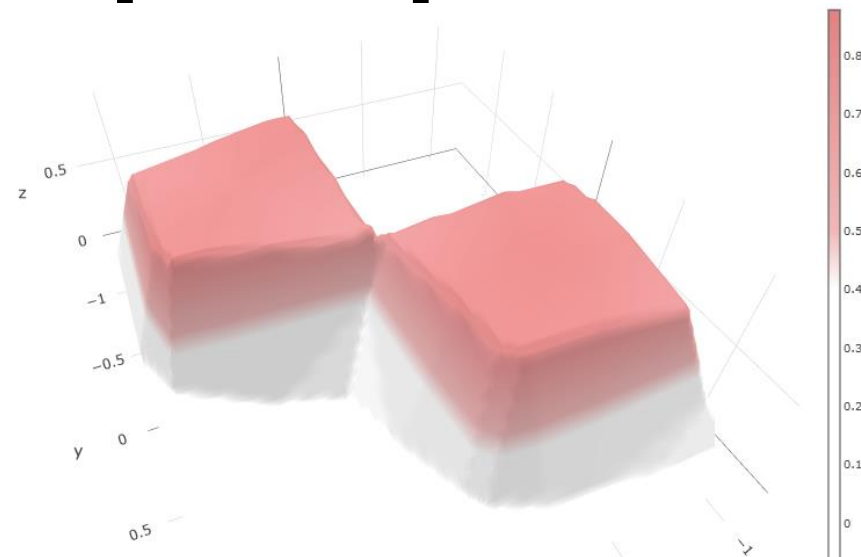
Репродуцированные данные: пример 1



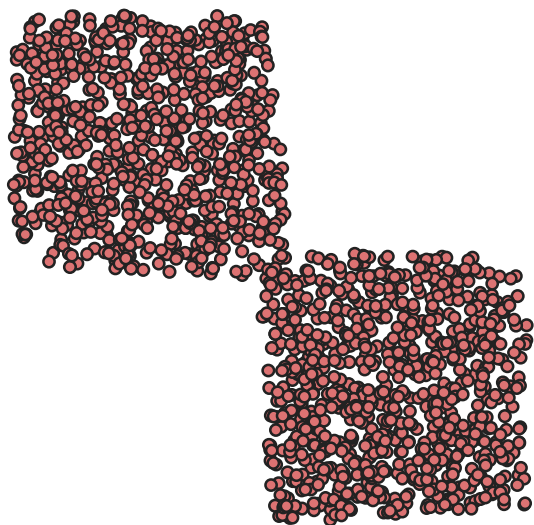
Оригинальный набор данных



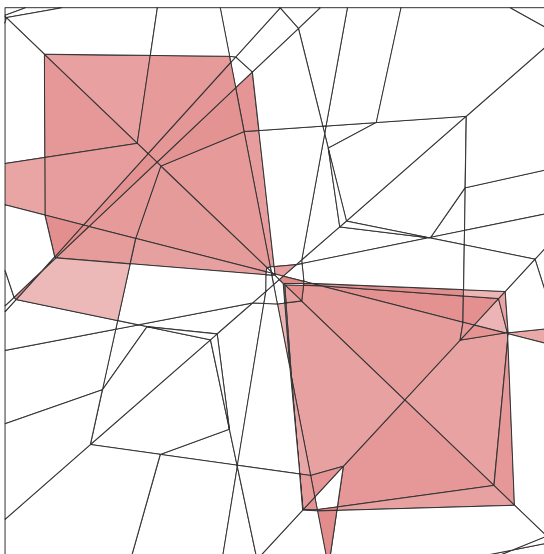
Выход $c_n^*(x)$



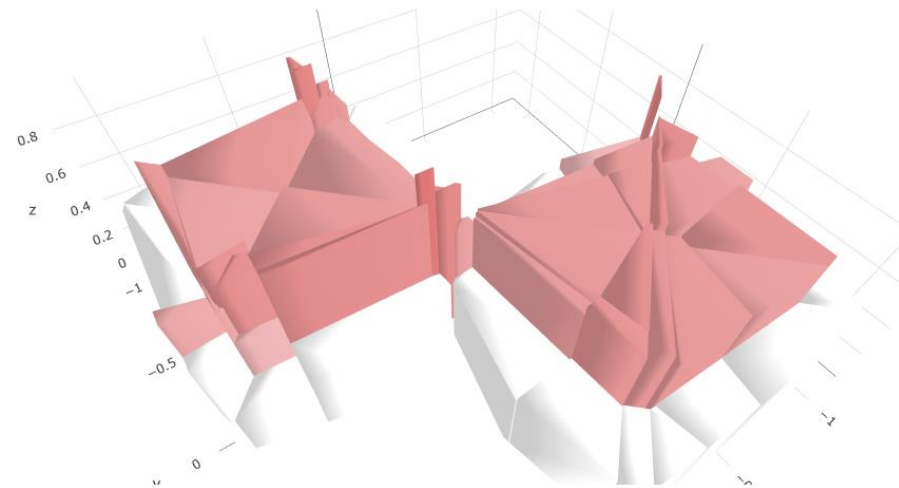
Выход $c_n^*(x)$ в 3D



Репродуцированный набор данных

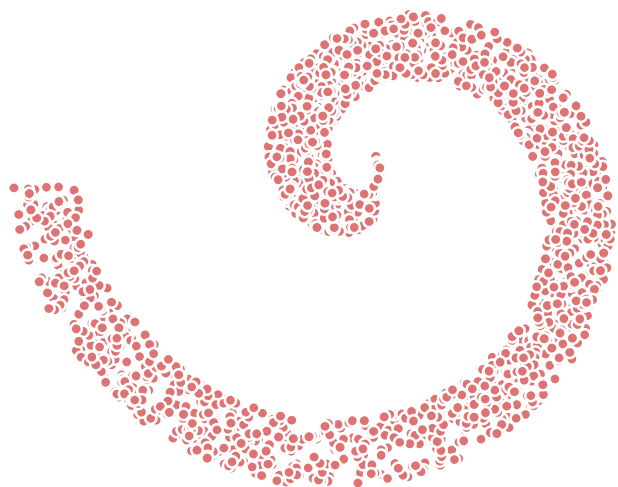


Выход $h_n^*(x)$



Выход $h_n^*(x)$ в 3D

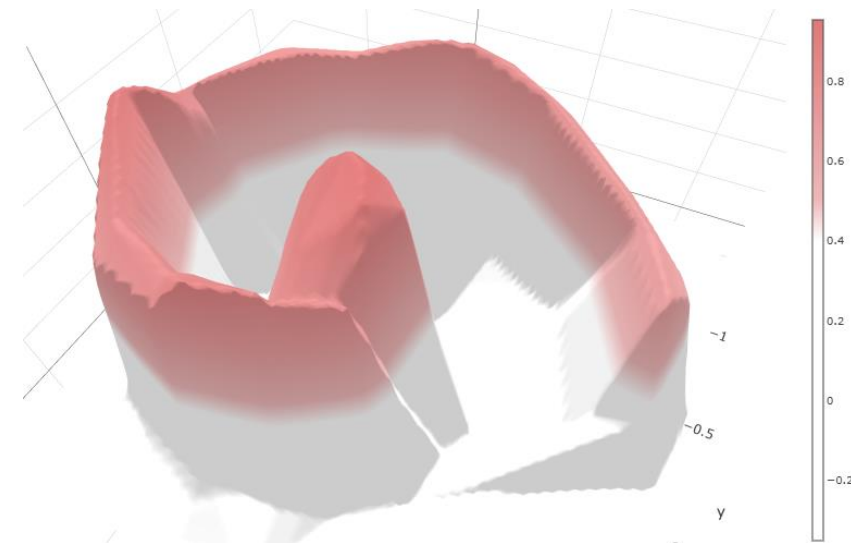
Репродуцированные данные: пример 2



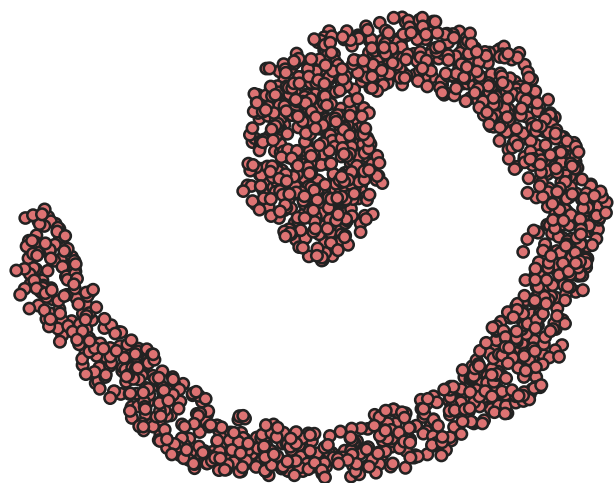
Оригинальный набор данных



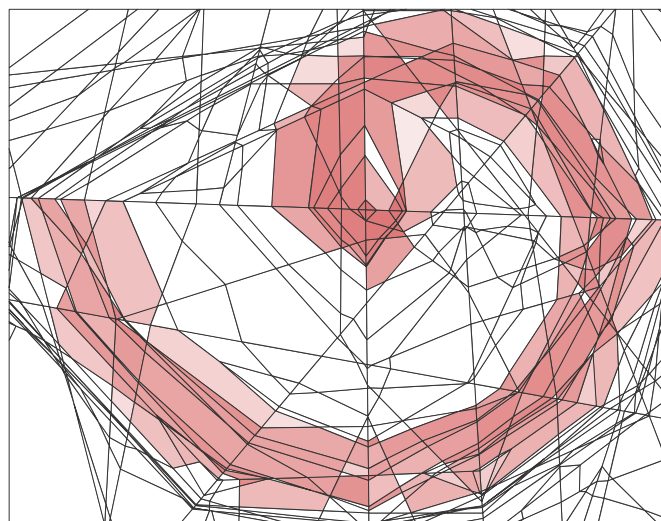
Выход $c_n^*(x)$



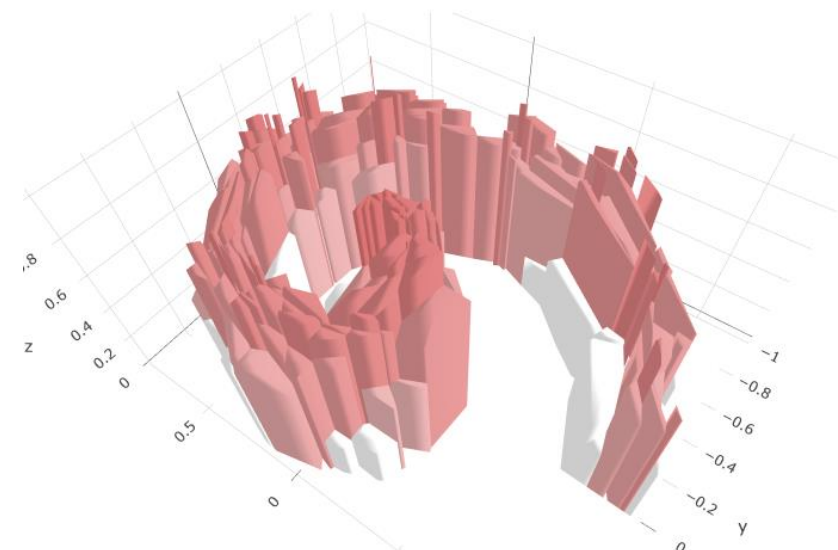
Выход $c_n^*(x)$ в 3D



Репродуцированный набор данных



Выход $h_n^*(x)$

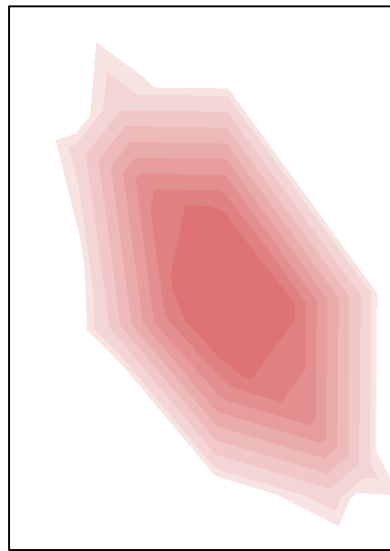


Выход $h_n^*(x)$ в 3D

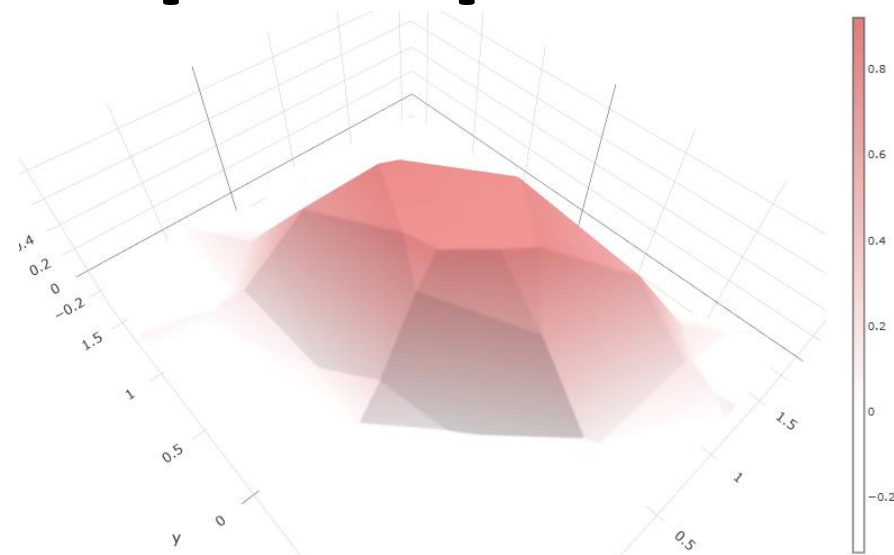
Репродуцированные данные: пример 3



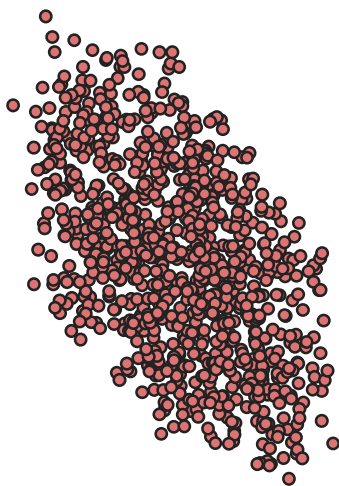
Оригинальный набор данных



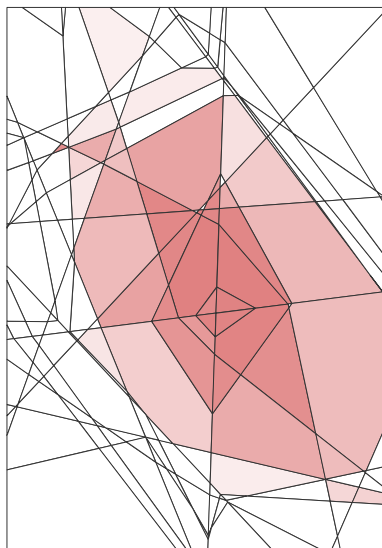
Выход $c_n^*(x)$



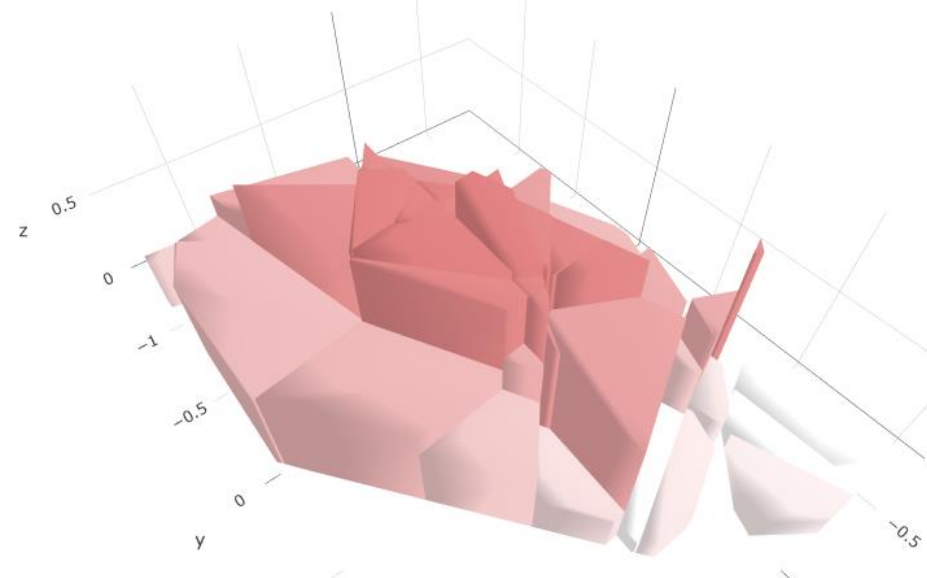
Выход $c_n^*(x)$ в 3D



Репродуцированный набор данных



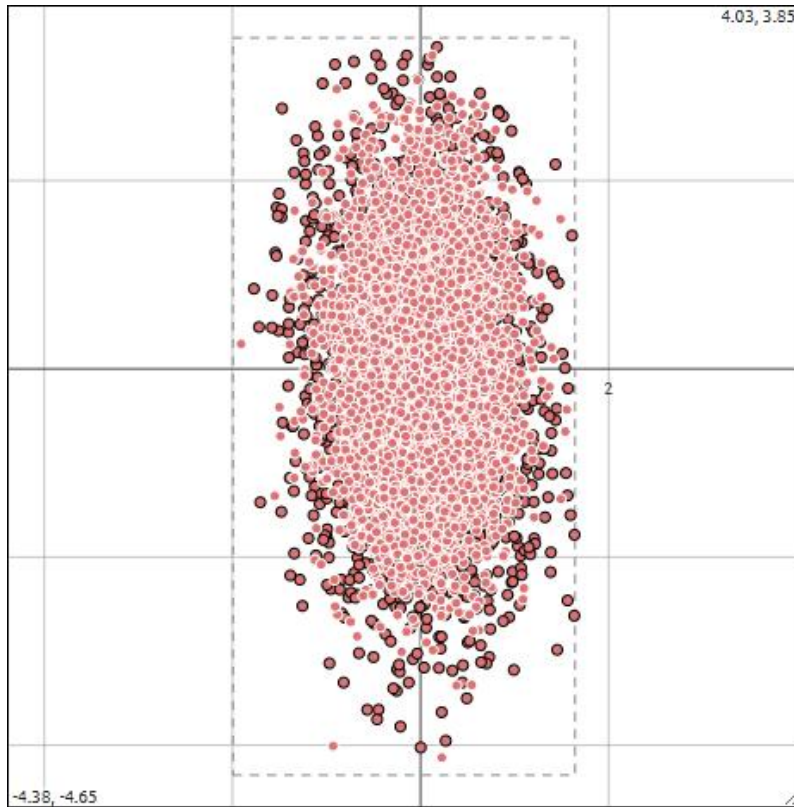
Выход $h_n^*(x)$



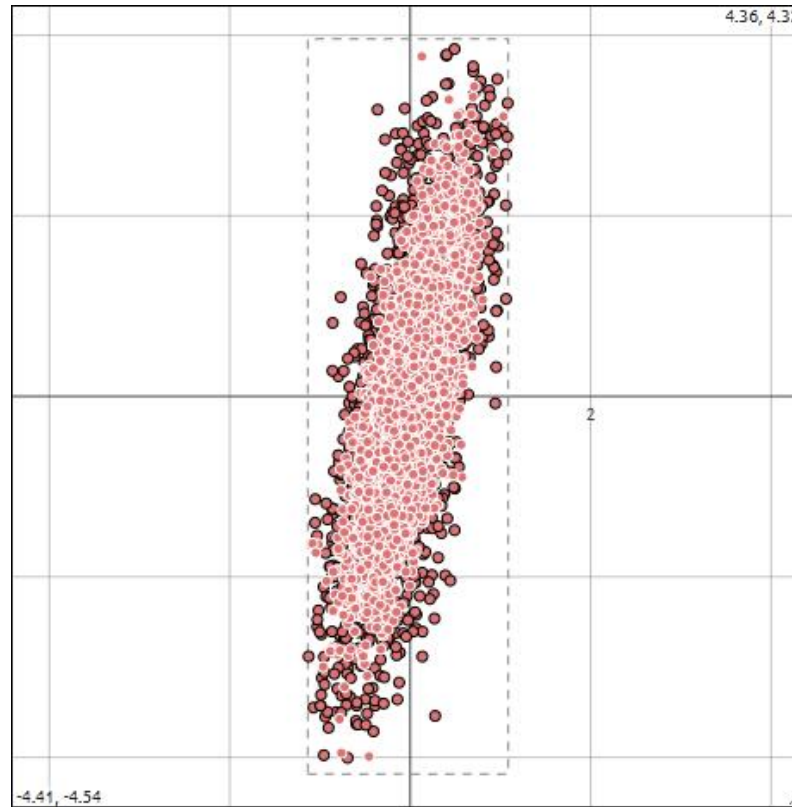
Выход $h_n^*(x)$ в 3D

Репродуцированные данные: пример 4

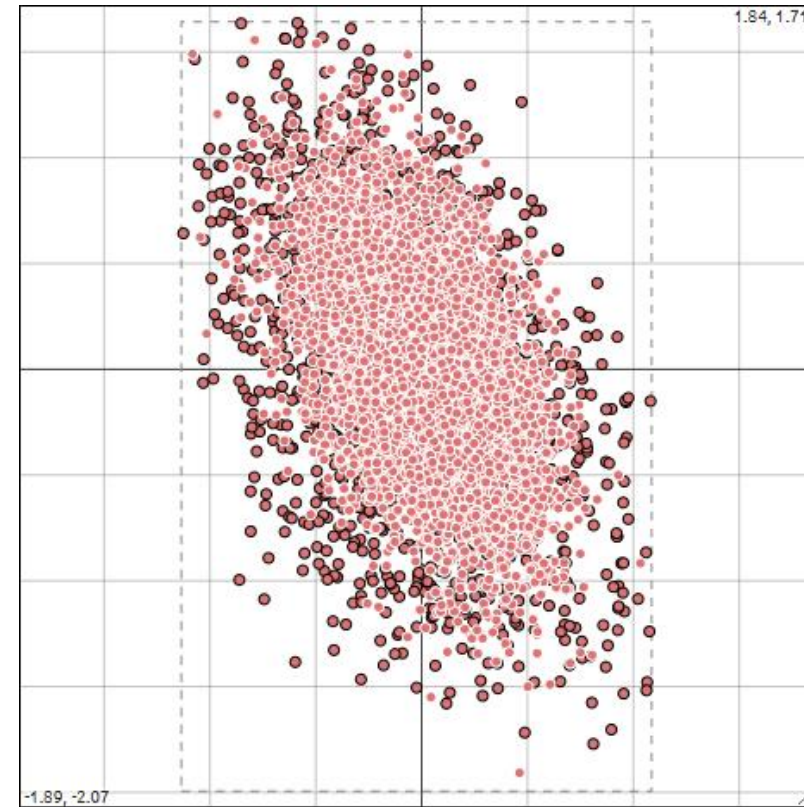
- Реальный X : 10-мерное Гауссово распределение.
- Синтетический набор $\hat{X}$: более высокая дисперсия по сравнению с 2D.



x_1/x_2 проекция



x_1/x_3 проекция



x_2/x_{10} проекция

Заключение

- **Унарная классификация** – теоретически обоснованная (ну почти) альтернатива многоклассовой, особенно в условиях **отсутствия отрицательных примеров** или при **сильном дисбалансе** классов.
- **Простая реализация:** достаточно только одного класса и функции отбора фоновых объектов – никаких разметок или сложных процедур.
- **Восстановление плотности распределения** открывает путь к новым способам анализа и визуализации данных.
- **Генерация синтетических данных** с сохранением структурных и статистических свойств оригинального распределения без латентного пространства.
- **Тонкая оценка доверия к модели:** анализ областей с высокой и низкой уверенностью, улучшение интерпретируемости.



**Интерактивный
визуализатор**