НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Обзор приложений копул к задачам Байесовской классификации при машинном обучении

Кузнецов Никита Алексеевич

Научный руководитель
Пеникас Генрих Иозович

Цель работы

Основная цель:

Сравнение стандартных походов к классификации с байесовской классификацией с использованием копул, в частности архимедовых копул.

Используемые классификаторы:

Классификация по двум классам.

- Квадратичный дискриминант.
- Байесовский классификатор с копулой Клейтона для первого класса.
- Байесовский классификатор с копулой Гумбеля для второго класса.
- Logit регрессия.

Данные:

Искусственно сгенерированные:

- Гауссовские
- Не гауссовские
  - Копула Клейтона для первого класса
  - Копула Гумбеля для второго класса

Применение

Возможные задачи классификации:

- Кредитный скоринг
- Моделирование вероятностей дефолтов
- Классификация изображений *[Крылов 2010], [Крылов 2013]*
- Распознавание речи
- Медицинские исследования *[Han, Zhao, Liu 2013]*

Байесовский классификатор

- $$a(x) = argmax\left(P_y * p_y(x)\right) \qquad y \in Y$$

$a(x)$ – Классификатор.

P_y – вероятность класса y.

$p_y(x) = p(x|y)$ - функция правдоподобия/вероятности для класса y.

С использованием функции потерь:

$$A(x) = argmax\left(\lambda_y P_y * p_y(x)\right)$$

λ_y – штраф за ошибку на классе y.

Квадратичный дискриминант

- $$a_{quadratic}(x) = argmax\left(P_y * \frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} e^{-\frac{1}{2}(x-u)^T\Sigma_y^{-1}(x-u)}\right)$$

$p_y(x) = p(x|y)$ - Подставляется формула плотности нормального распределения - $\frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} e^{-\frac{1}{2}(x-u)^T\Sigma_y^{-1}(x-u)}$

Логарифмируя и убирая константы:

$$a_{quadratic}(x) = argmax\left(\ln(P_y) - \frac{1}{2}(x-u)^T\Sigma^{-1}(x-u) - \frac{1}{2}\ln(|\Sigma_y|)\right)$$

Интеграция копул в байесовский классификатор

Предложена *Sathe в 2006.*

Производя замену:

$$p_y(x) = f^y(x_1, \dots, x_d)$$

По теореме Скляра:

$$f^y(x_1, \dots, x_d) = c^y\left(F_1{}^y(x_1|\theta_1^y), \dots, F_d{}^y(x_d|\theta_d^y)\right) * \prod_{i=1}^{d} f_i^y(x_i|\theta_i^y)$$

Подставляя в классификатор:

$$a(x) = argmax\left(P_y * c^y\left(F_1{}^y(x_1|\theta_1^y), \dots, F_d{}^y(x_d|\theta_d^y)\right) * \prod_{i=1}^{d} f_i^y(x_i|\theta_i^y)\right)$$

Байесовский классификатор и архимедовы копулы

Архимедова копула:

$$C = \varphi^{-1}(\varphi(F(x_1)) + \cdots + \varphi(F(x_d)))$$

Тогда плотность архимедовой копулы:

$$c = \varphi^{-1[d]}\left(\sum_{i=1}^{d}\varphi(F_i(x_i))\right) * \prod_{j=1}^{d}\varphi'(F_i(x_i))$$

Подставляя плотность копулы в классификатор получается:

$$a(x) = argmax\left(P_y * \varphi^{-1[d]^y}\left(\sum_{i=1}^{d}\varphi^y\left({F_i}^y(x_i|\theta_i^y)\right)\right) * \prod_{j=1}^{d}\varphi'^y\left({F_i}^y(x_i|\theta_i^y)\right) * \prod_{i=1}^{d}{f_i}^y(x_i|\theta_i^y)\right)$$

Происходила генерация двух типов данных:

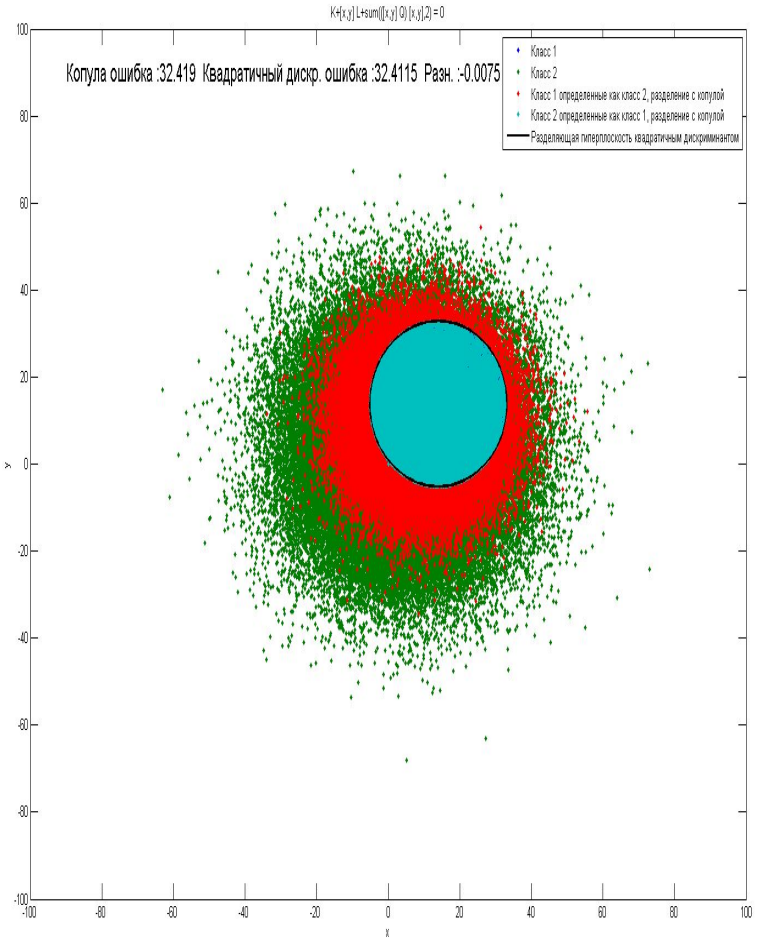
- Гауссовские
- Не гауссовские
  - Копула Клейтона для первого класса.
  - Копула Гумбеля для второго класса.

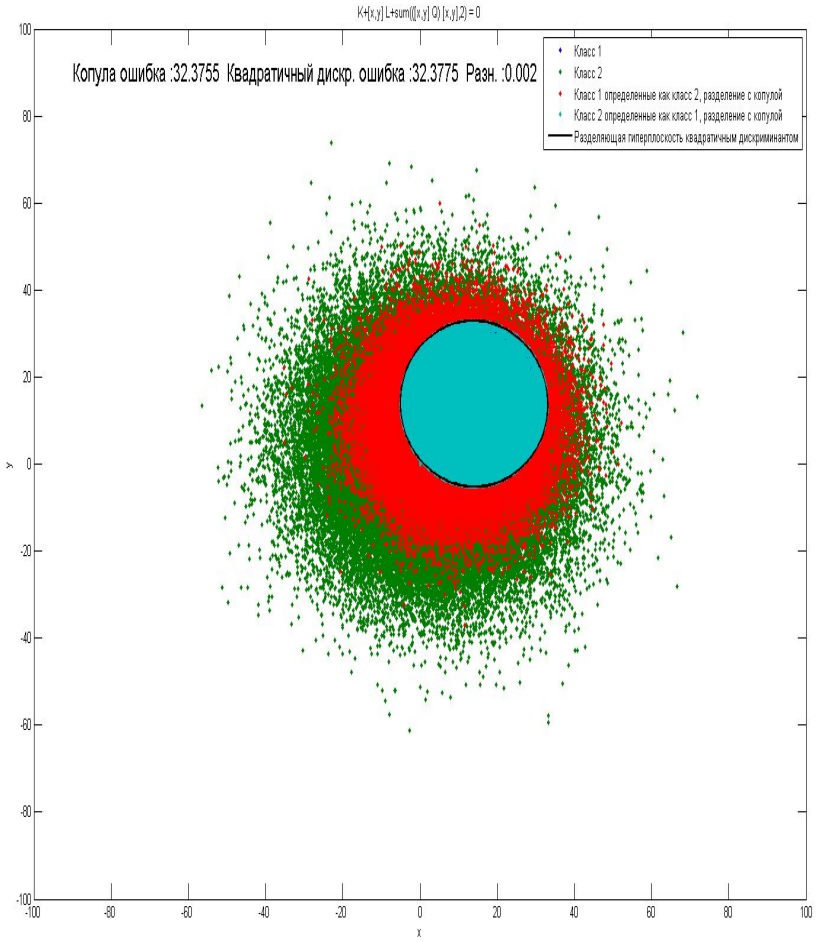
В обоих случаях маржинальные распределения:

Класс 1		Класс 2	
Среднее X	10	Среднее X	5
Среднее Y	10	Среднее Y	5
Ст. откл. X	10	Ст. откл. X	15
Ст. откл. Y	10	Ст. откл. Y	15

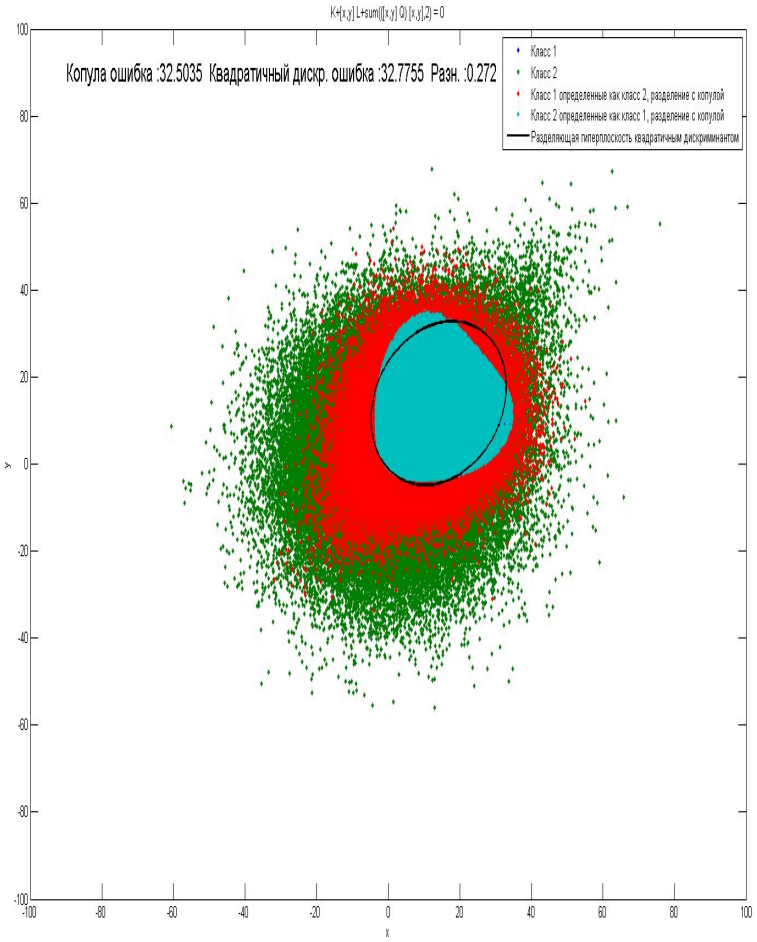
Схема процесса

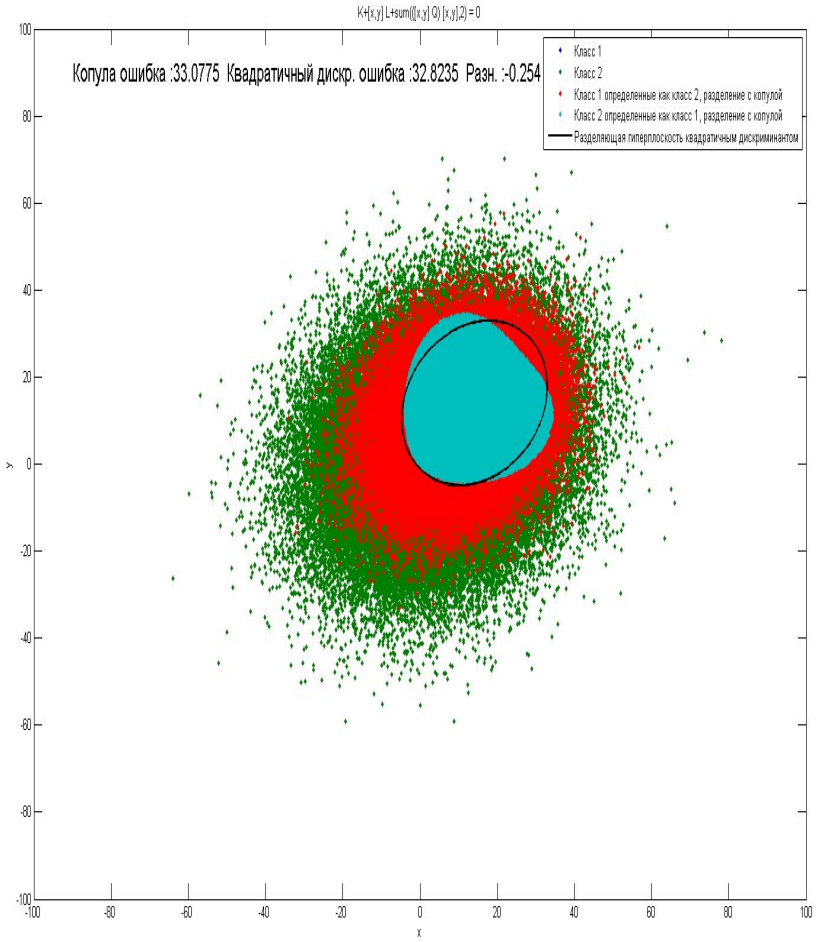
Результаты. Корр. 0



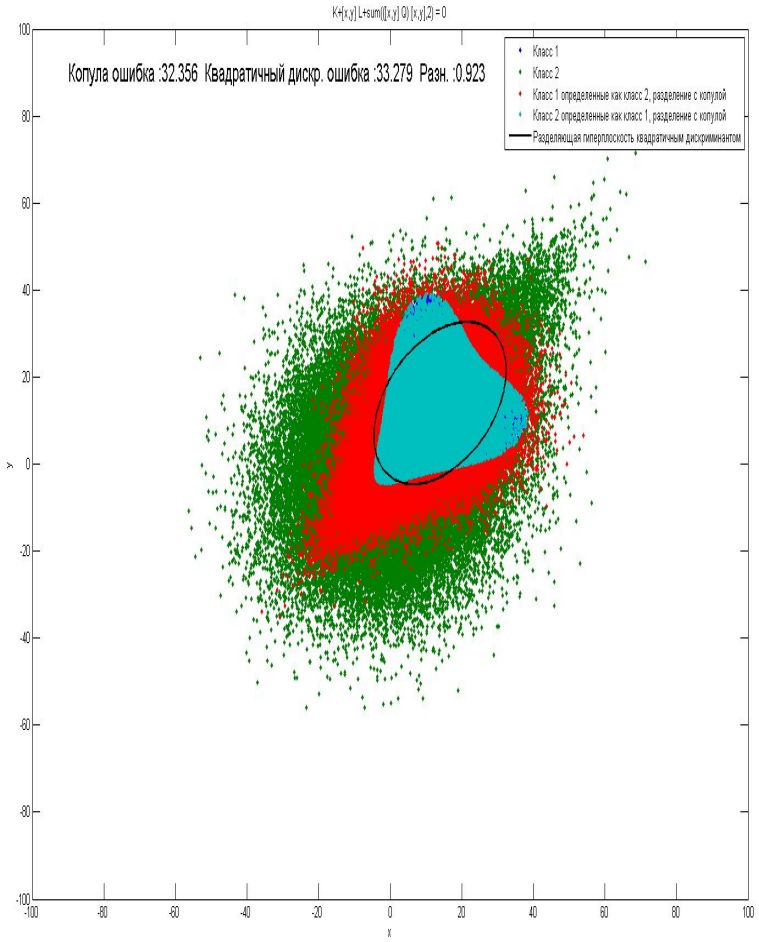


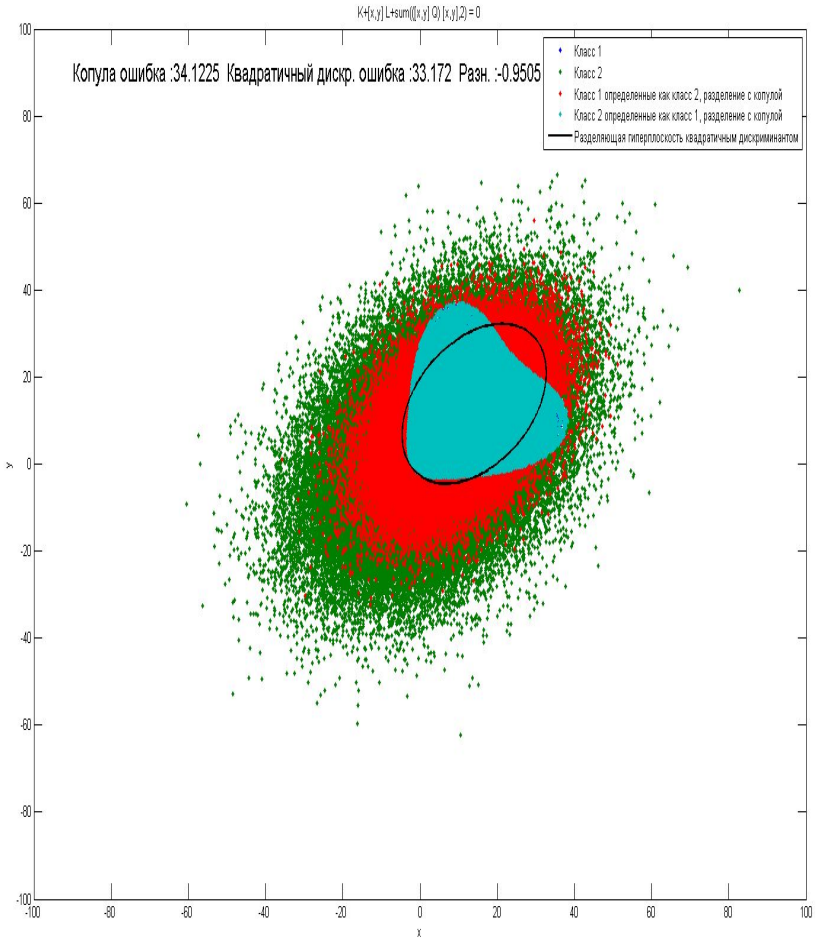
Результаты Корр. 0,2



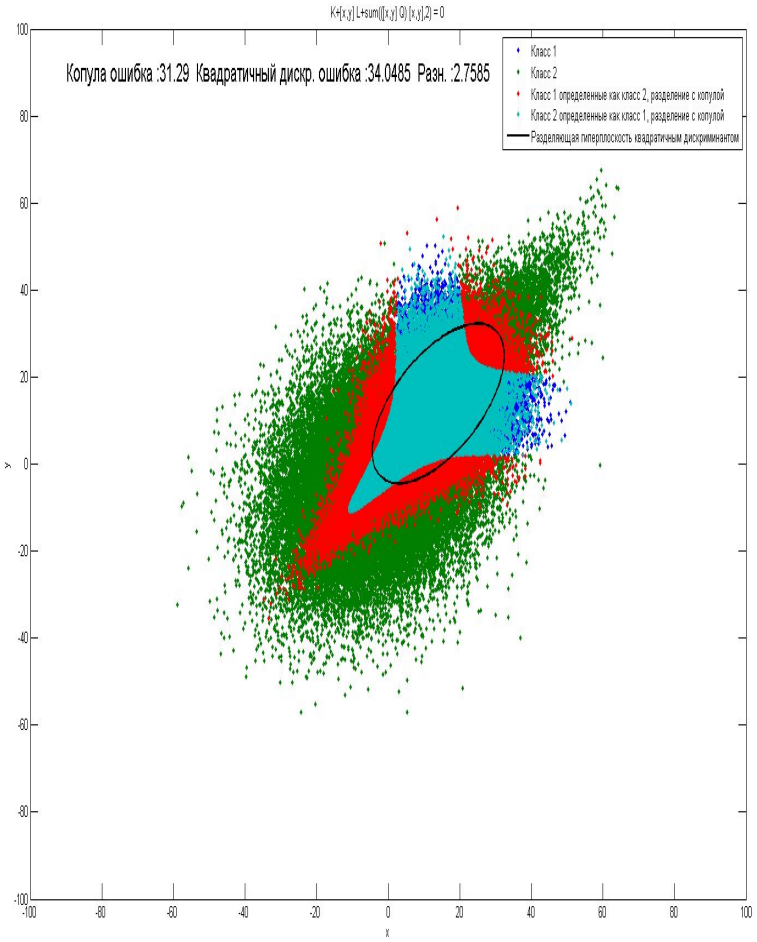


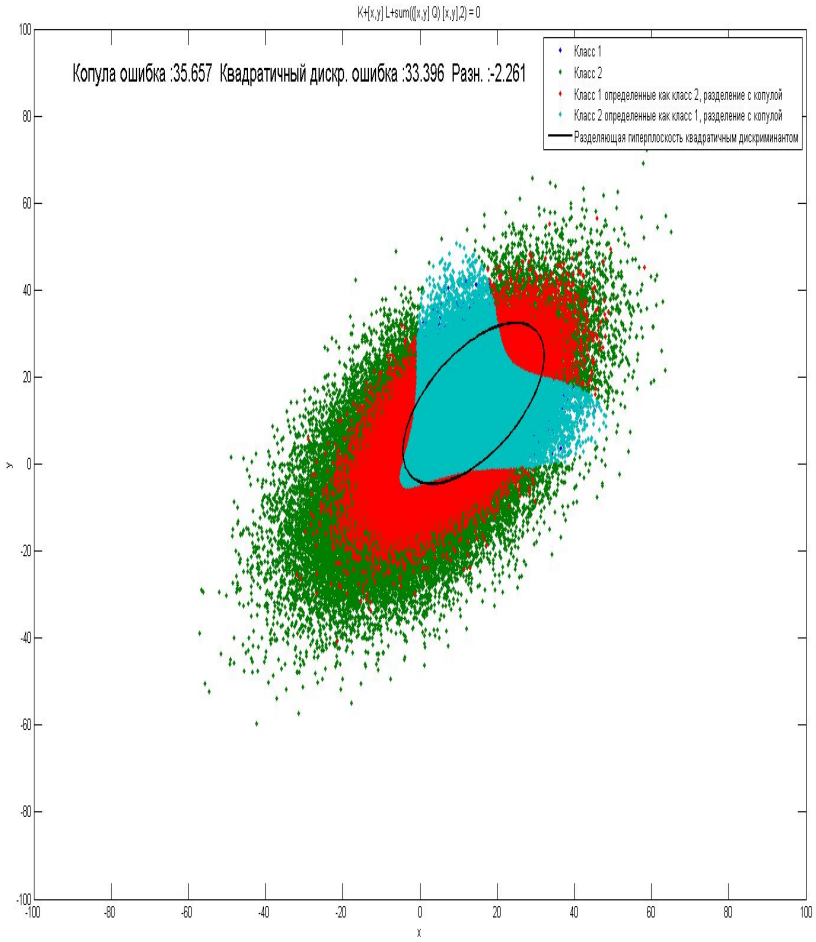
Результаты Корр. 0,4



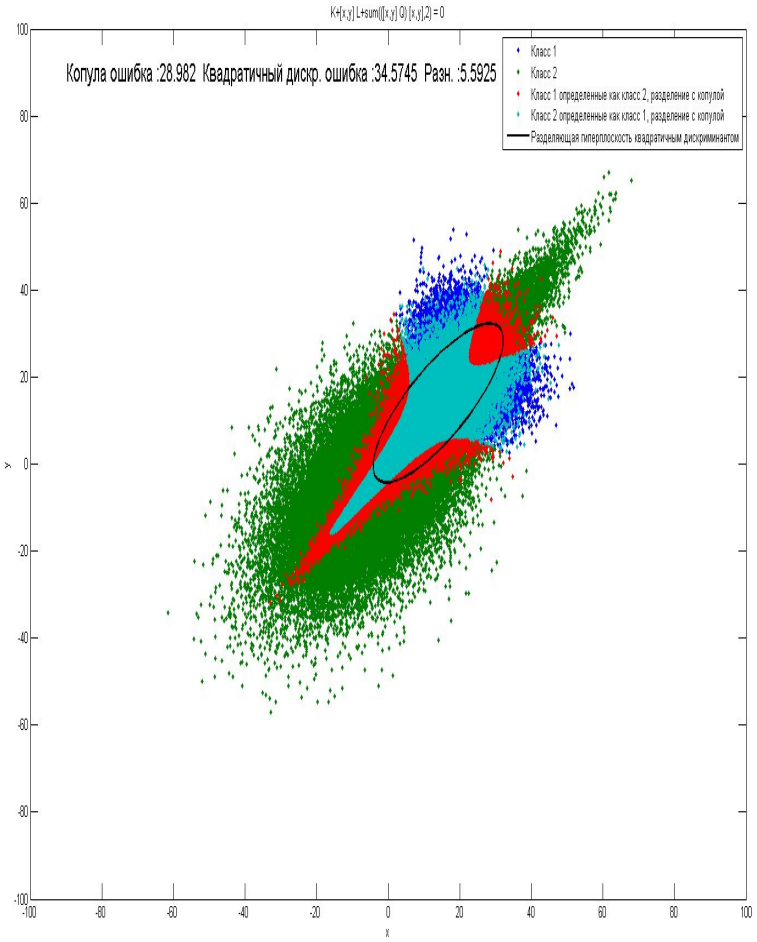


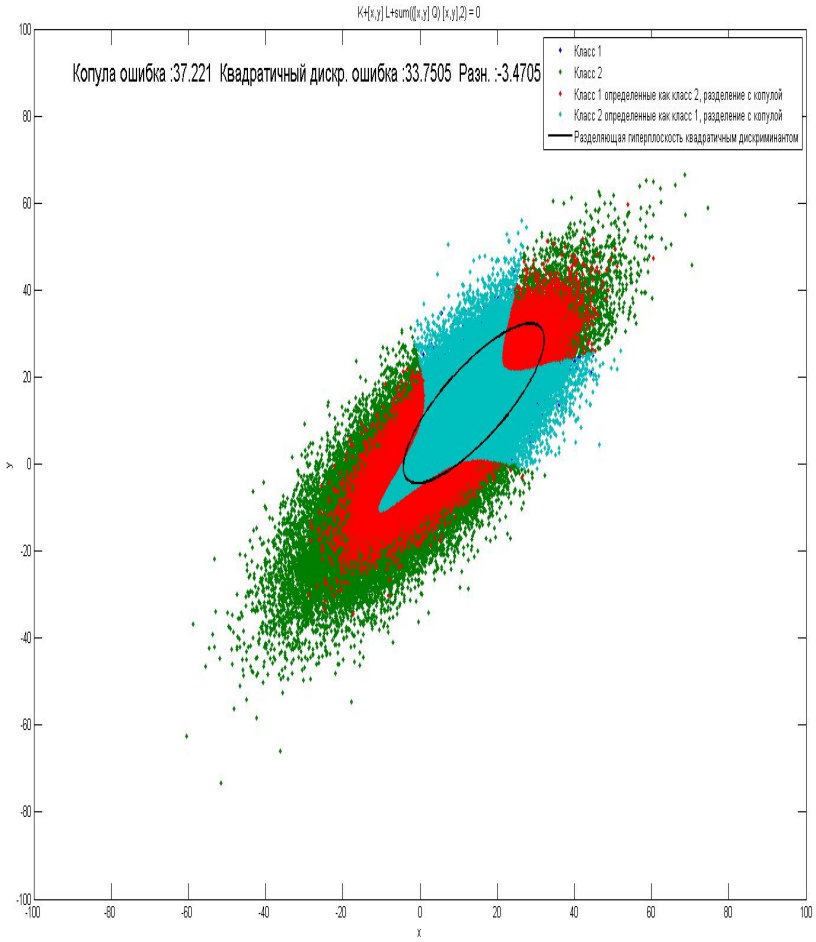
Результаты Корр. 0,6



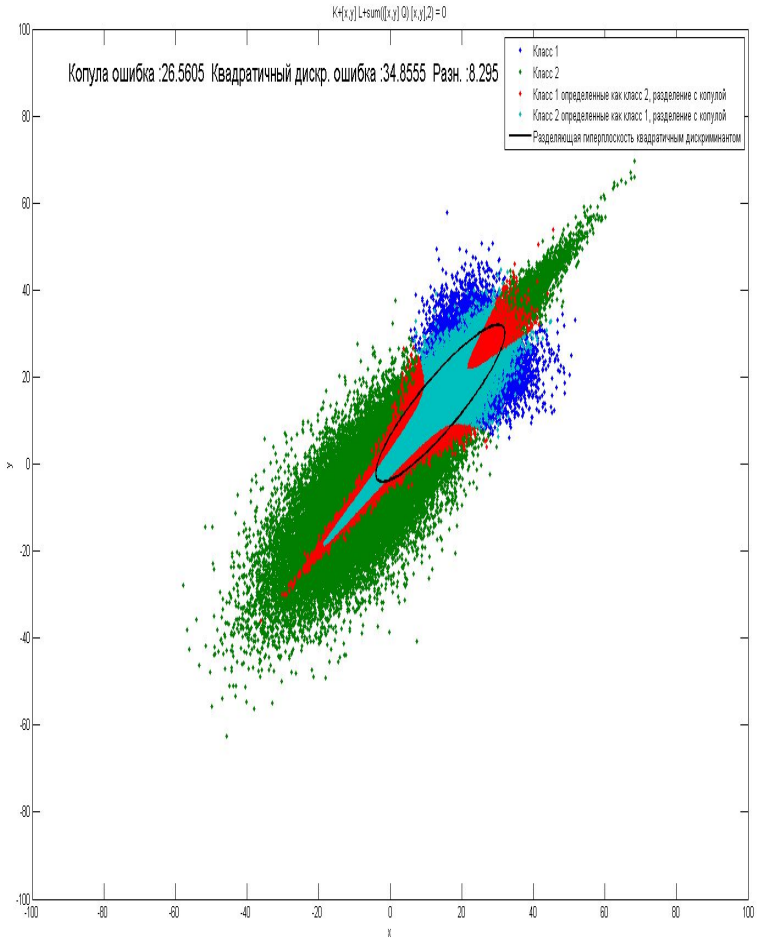


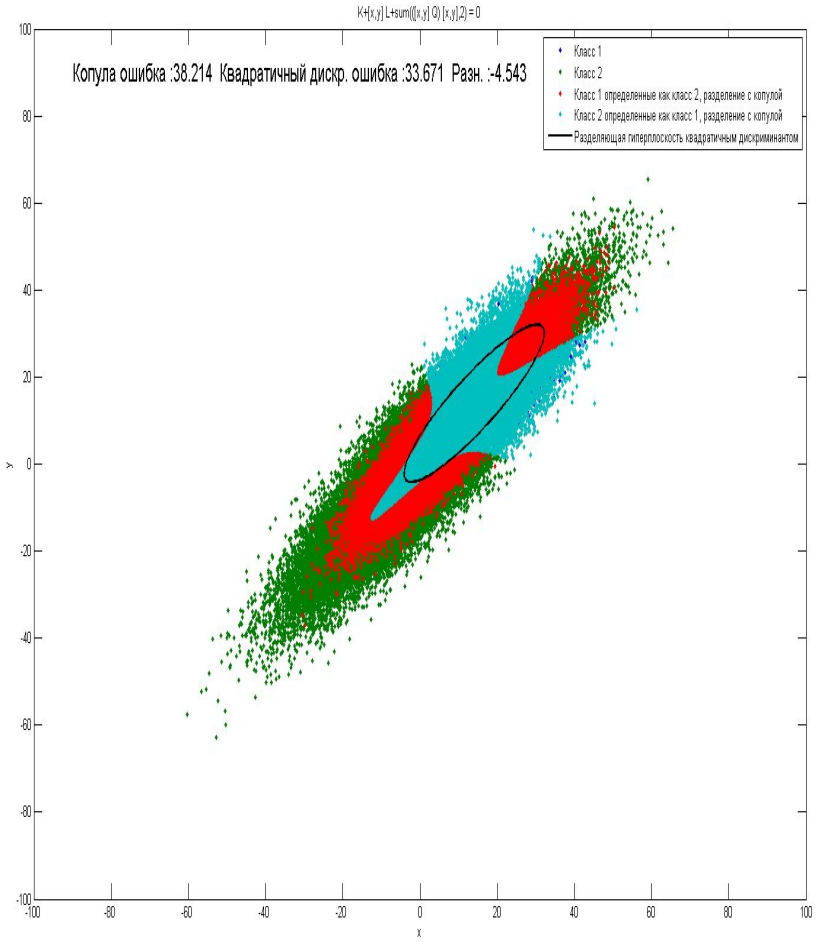
Результаты Корр. 0,8



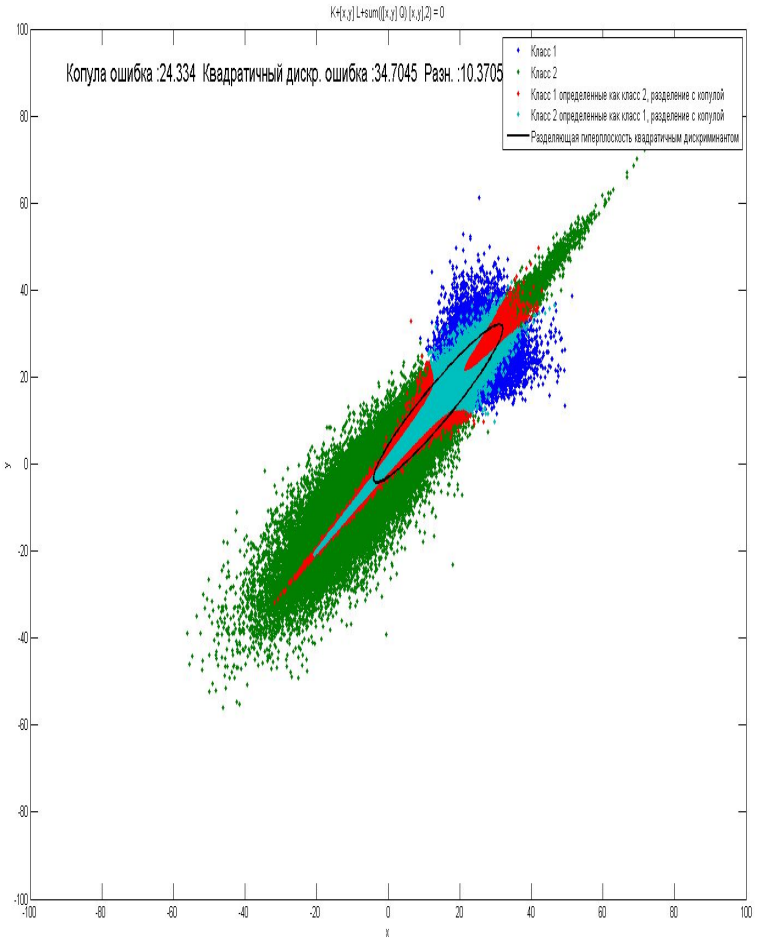


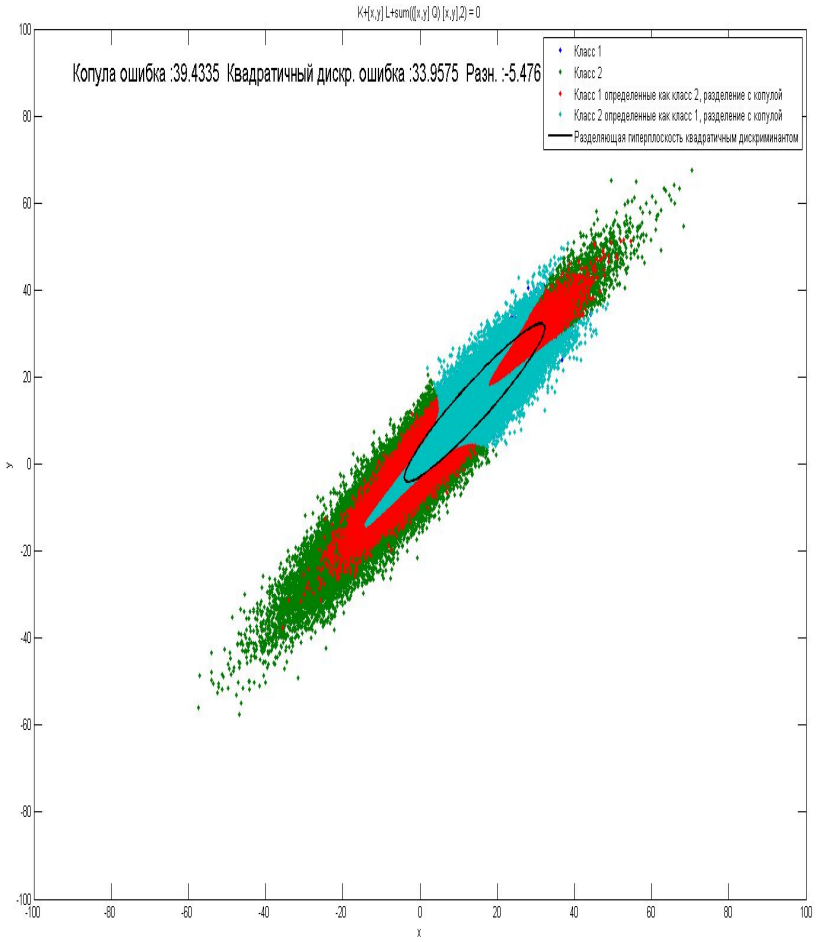
Результаты Корр. 0,9



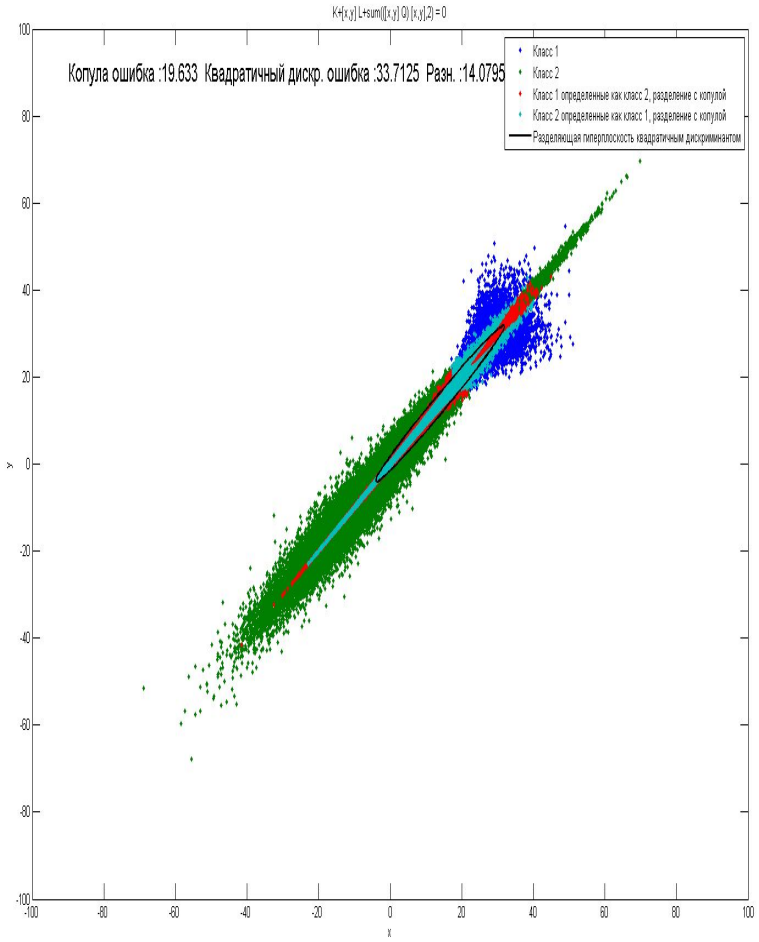


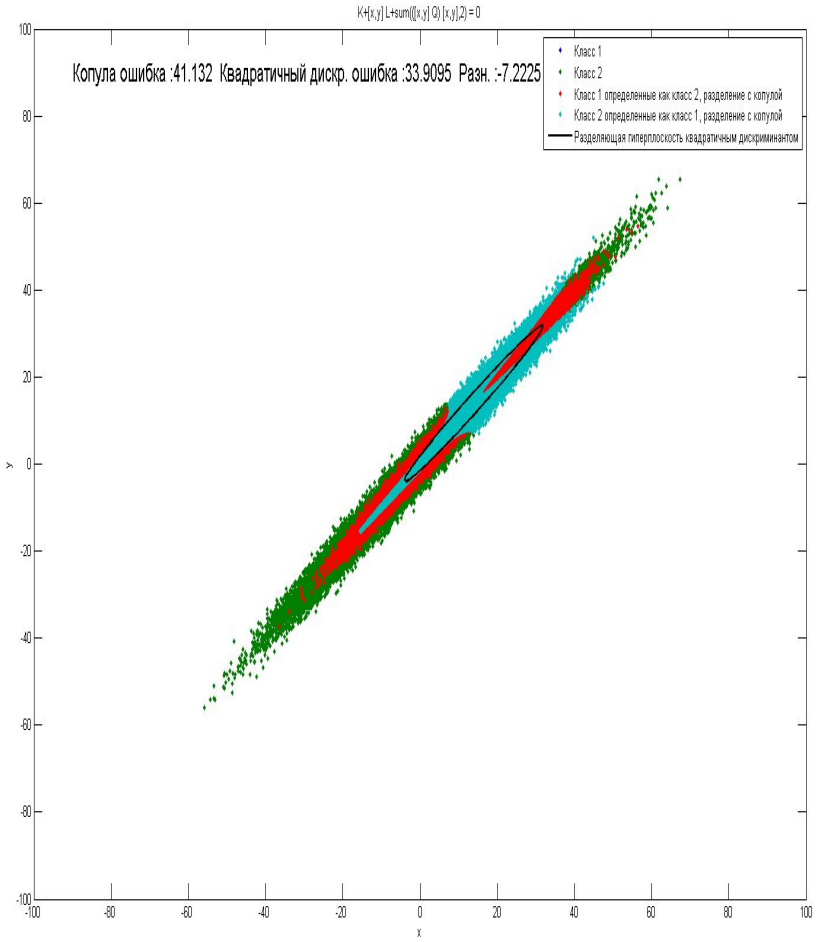
Результаты Корр. 0,95



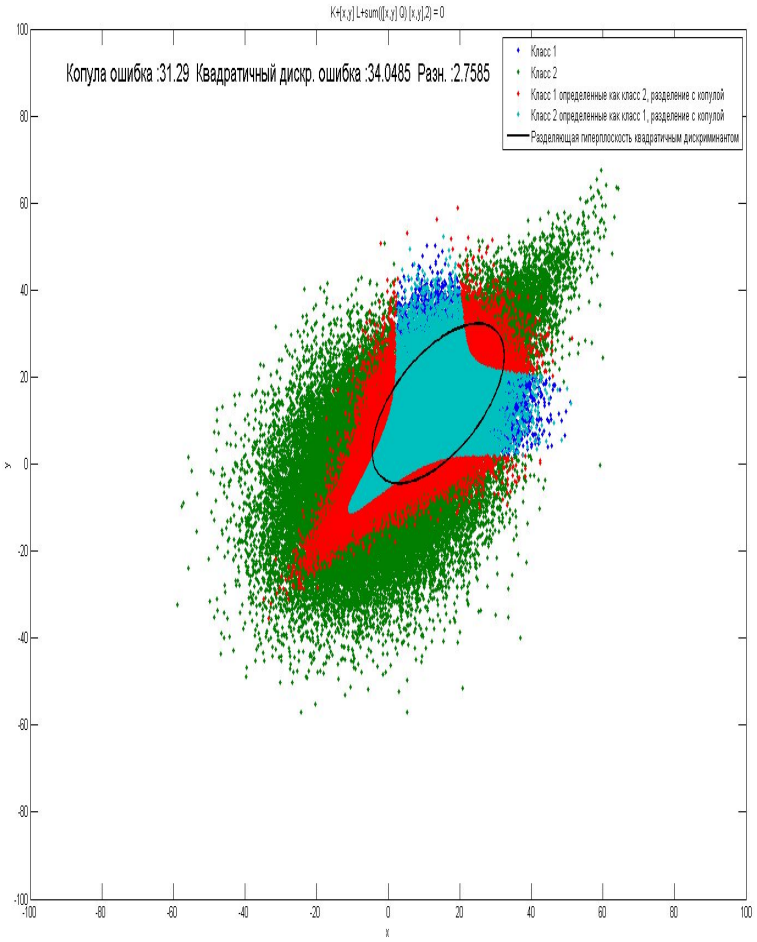


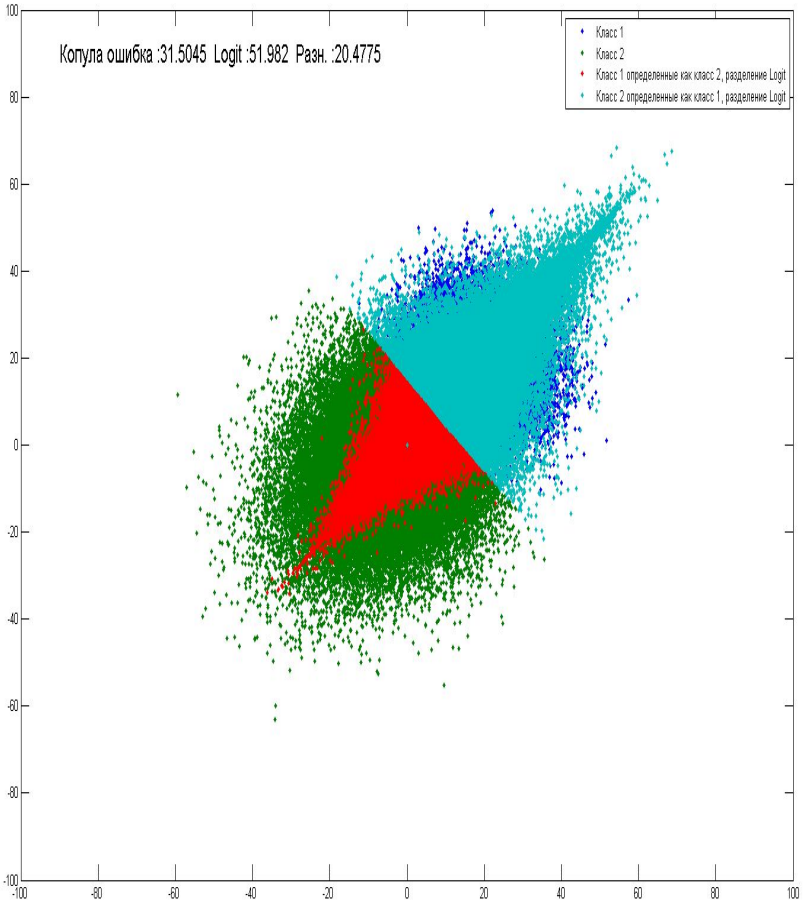
Результаты Корр. 0,99





Результаты Корр. 0,6, Logit





Результаты таблица 1

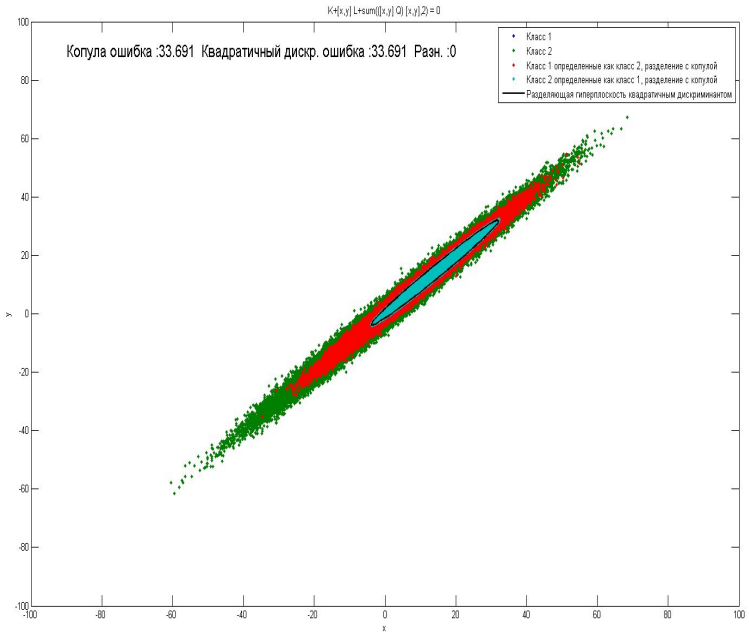
Гауссовские данные			Не Гауссовские данные		
Корреляция	Средняя ошибка в %		Параметр копулы	Средняя ошибка в %	
	Копула	Квадр. Дискр.		Копула	Квадр. Дискр.
0	32.38	32.38	0/1	32.42	32.41
0.1	32.58	32.54	0.14/1.07	32.41	32.51
0.2	33.08	32.82	0.29/1.14	32.50	32.78
0.3	33.54	33.04	0.47/1.24	32.59	33.15
0.4	34.12	33.17	0.7/1.35	32.36	33.28
0.5	34.91	33.24	1/1.5	32.07	33.65
0.6	35.66	33.40	1.45/1.7	31.29	34.05
0.7	36.26	33.45	2.15/2	30.34	34.49
0.8	37.22	33.75	3.4/2.5	28.98	34.57
0.9	38.21	33.67	6.8/3.65	26.56	34.86
0.95	39.43	33.96	12.8/5.25	24.33	34.70
0.99	41.13	33.91	50/12	19.63	33.71

Результаты таблица 2

Параметры	Разница гаусс.	Разница не гаусс.
0/0/1	0.00	- 0.01
0.1/0.14/1.07	- 0.03	0.11
0.2/0.29/1.14	- 0.25	0.27
0.3/0.47/1.24	- 0.50	0.56
0.4/0.7/1.35	- 0.95	0.92
0.5/1/1.5	- 1.67	1.58
0.6/1.45/1.7	- 2.26	2.76
0.7/2.15/2	- 2.82	4.15
0.8/3.4/2.5	- 3.47	5.59
0.9/6.8/3.65	- 4.54	8.30
0.95/12.8/5.25	- 5.48	10.37
0.99/50/12	- 7.22	14.08

Выводы

1) На не гауссовских данных, классификатор с использованием копулы более точное чем стандартный квадратичный дискриминант.
2) На гауссовских данных происходит потеря точности, которую можно компенсировать «умным» использованием копулы.



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!

101000, Россия, Москва, Мясницкая ул., д. 20
Тел.: (495) 621-7983, факс: (495) 628-7931
www.hse.ru