Последовательный метод опорных векторов и визуализация классификаторов

Притыковская Наталья Николаевна, группа 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., асс. Коробейников А. И.
Рецензент: к.ф.-м.н., доц. Алексеева Н. П.

Санкт-Петербург
2012г.

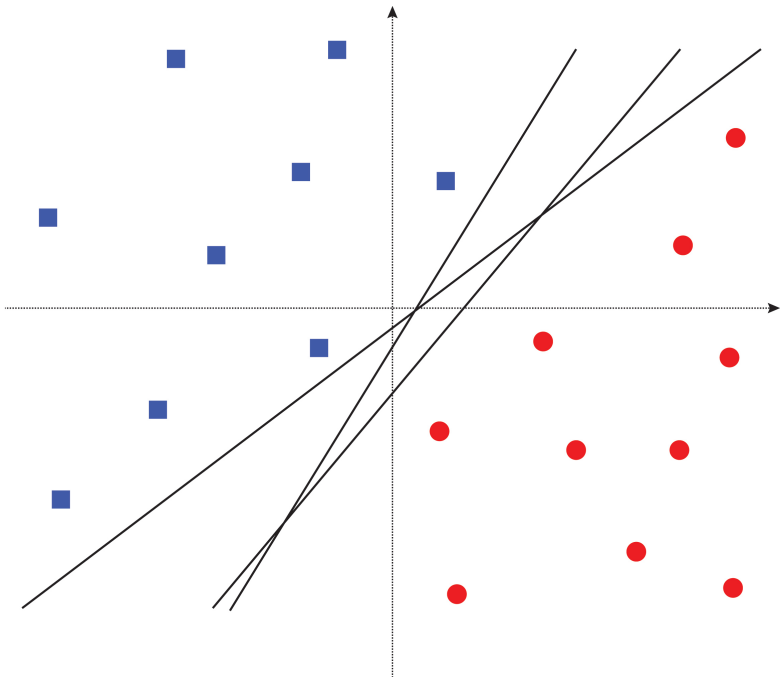
Постановка задачи классического SVM

Рассматривается задача классификации на два непересекающихся класса.

$X = \mathbb{R}^n$ — пространство объектов
$Y = \{-1, +1\}$ — метки
$f : X \to Y$
обучающая выборка $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$, $y_i = f(\mathbf{x}_i)$

Требуется аппроксимировать целевую зависимость f на всем пространстве X.

Линейный пороговый классификатор

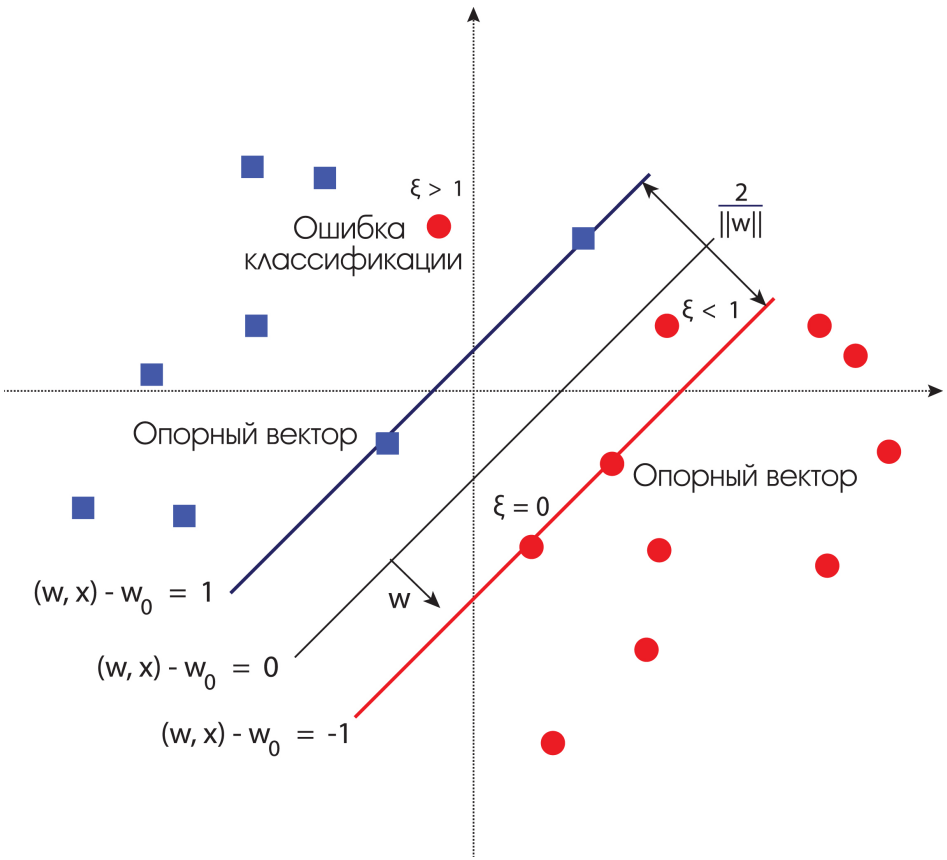


$$a(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - w_0)$$

$\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}} \in \mathbb{R}^n$,
$w_0 \in \mathbb{R}$ — параметры.

$\langle \mathbf{w}, \mathbf{x} \rangle = w_0$ —
гиперплоскость,
разделяющая классы в
пространстве $\mathbb{R}^n$.

Линейный SVM



Параметры $\mathbf{w}$ и w_0 получаются из решения:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \to \min_{\mathbf{w},w_0,\xi}$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle - w_0) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, l\ .$$

Kernel trick

Отображение в расширенное пространство

$$\phi(\mathbf{x}) \ : \ \mathbb{R}^n \to H$$

- H — пространство со скалярным произведением
- ядро $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_H$
- $a(\mathbf{x}) = \mathrm{sign}(\mathrm{K}(\mathbf{w}, \phi(\mathbf{x})) - w_0)$

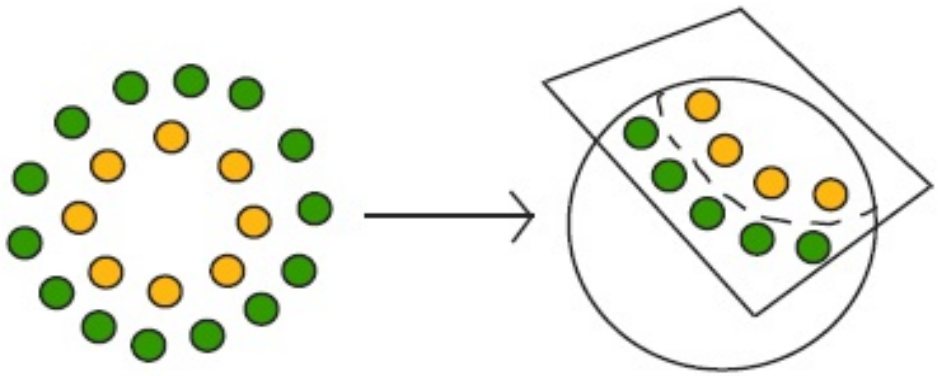


Рис. : $\phi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2)$

Далее все скалярные произведения рассматриваются в H.

Поставленные и достигнутые цели

Проблема:

- мотивация решений, принимаемых классификатором, остается за рамками классического метода.

Способы решения:

- Построить методом SVM несколько классификаторов, ортогональных друг другу.
- Ввести новые признаки — расстояния до классификаторов.
- Проинтерпретировать новые признаки.
- Дополнить имеющиеся данные новыми признаками и посмотреть, как это повлияет на качество классификации.

Формальная постановка задачи

- дано $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$ обучающая выборка
- $\mathbf{v}_1, \ldots, \mathbf{v}_n$ направляющие вектора первых n классификаторов, попарно ортогональных
- требуется найти $\mathbf{w}$ и w_0 - параметры $n+1$ классификатора

$$\begin{cases} \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^l \xi_i \to \min_{\mathbf{w}, w_0, \xi} \\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - w_0) \geq 1 - \xi_i, \quad i = 1, \ldots, l \\ \xi_i \geq 0, \quad i = 1, \ldots, l \\ \langle \mathbf{w}, \mathbf{v}_i \rangle = 0, \quad i = 1, \ldots, n \end{cases}$$

Двойственная задача

- $\mathbf{v}_1, \ldots, \mathbf{v}_n$ — направляющие вектора n классификаторов
- $(\lambda_1, \ldots, \lambda_l)$ - вектор двойственных переменных

$$-\frac{1}{2}\langle\sum_{i=1}^{l}\lambda_i y_i \mathbf{x}_i, \sum_{j=1}^{l}\lambda_j y_j \mathbf{x}_j\rangle - \frac{1}{2}\sum_{i=1}^{n} t_i^2\langle \mathbf{v}_i, \mathbf{v}_i\rangle - \sum_{i=1}^{n} t_i\langle \mathbf{v}_i, \sum_{j=1}^{l}\lambda_j y_j \mathbf{x}_j\rangle -$$

$$-\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} t_i t_j\langle \mathbf{v}_i, \mathbf{v}_j\rangle + \sum_{i=1}^{l}\lambda_i \to max_{\mathbf{t},\lambda}$$

$$\sum_{i=1}^{l}\lambda_i y_i = 0$$

$$0 \le \lambda_i \le C, \quad i = 1, \ldots, l$$

- $\mathbf{w} = \sum_{i=1}^{l}\lambda_i y_i \mathbf{x}_i + \sum_{i=1}^{n} t_i \mathbf{v}_i$

Сложности, возникающие при решении

Компактно двойственную задачу можно записать в виде

$$\begin{cases} \frac{1}{2}x^T Gx + d^T \to min_x \\ Ax \geq b \end{cases},$$

где G - матрица Грамма, образованная множеством векторов $(\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{v}_1, \ldots, \mathbf{v}_n)$

- при условии, что в выборке есть повторяющиеся или линейно зависимые $\mathbf{x}_i$ матрица G, становится из положительно определенной положительно полуопределенной
- размерность матрицы G равна $n+1$, в случае больших выборок вычисления становятся очень долгими

Алгоритм SMO [Джон Платт, 1999] используется для решения двойственной задачи классического метода SVM:

$$\begin{cases} -\sum_{i=1}^{n} \lambda_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \to \min_{\lambda} \\ \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, n \end{cases}.$$

Идея:

- Выбираются 2 двойственные переменные
- Одна выражается через другую с помощью линейного ограничения
- Аналитически решается экстремальная задача

Двойственная задача для второго классификатора

Пусть $\mathbf{v}_1$ - направляющий вектор первого классификатора.

$$\begin{cases} \frac{1}{2}\langle \mathbf{v}_1, \mathbf{v}_1 \rangle t^2 + t \sum_{i=1}^{n} \lambda_i y_i \langle \mathbf{v}_1, \mathbf{x}_i \rangle + \\ + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{n} \lambda_i \to \min_{\lambda, \mathbf{t}} \\ \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \le \lambda \le C, \quad i = 1, \ldots, n \end{cases}$$

Решение:

- **Шаг 1:** находим λ_i и λ_j при помощи SMO
- **Шаг 2:** находим минимум квадратично функционала относительно t

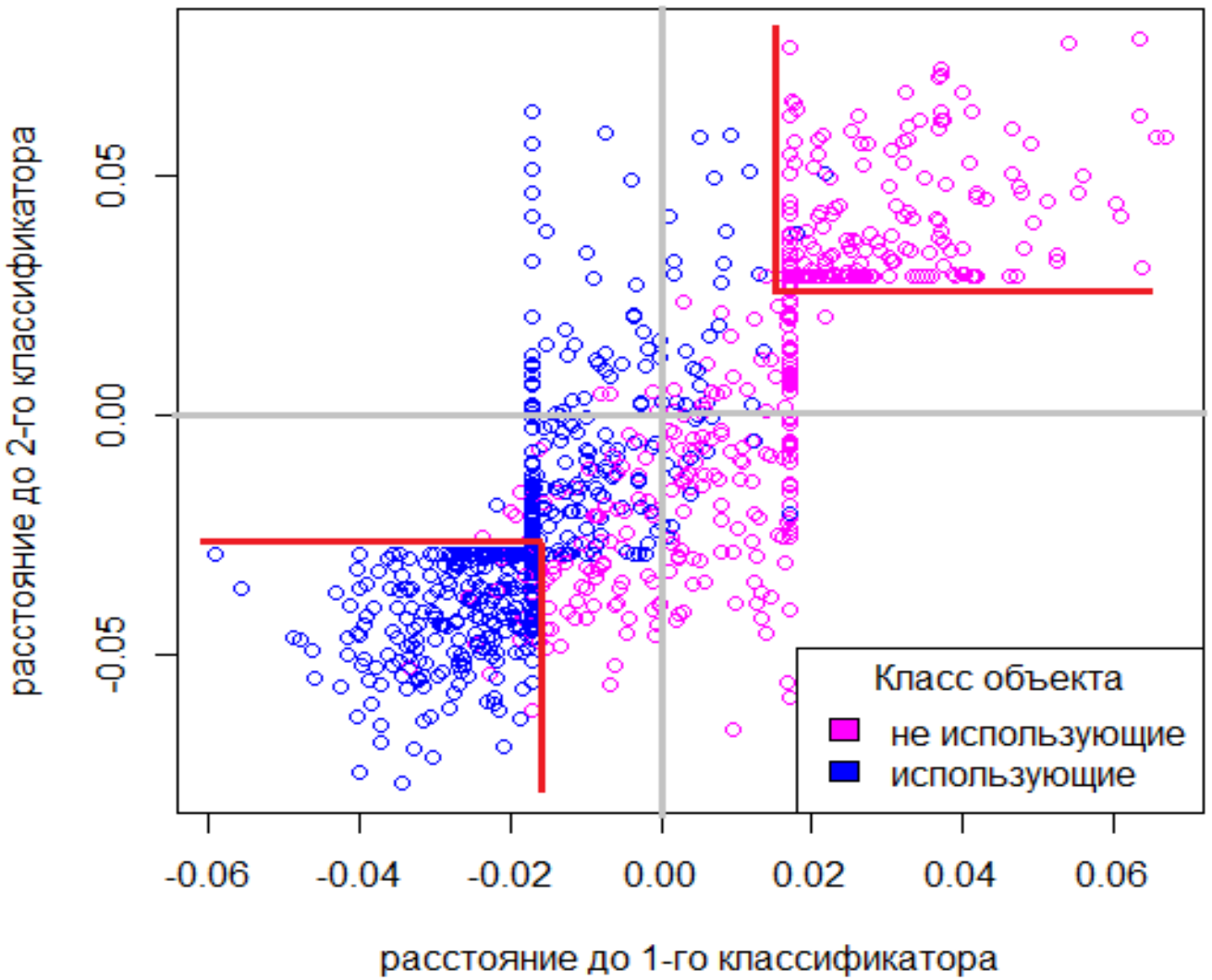
Модифицированный метод SMO был реализован на языке R.

Наборы данных

Все наборы данных были взяты из UC Irvine Machine Learning Repository.

- Данные об использовании контрацептивов в Индонезии, 1473 наблюдения, 9 признаков.
- Haberman's survival - данные о выживаемости женщин, перенесших рак груди в 1958-1970 годах, 306 наблюдений, 3 признака.
- Сведения о домах в предместье Бостона, 506 наблюдений, 14 признаков.

Использование контрацептивов



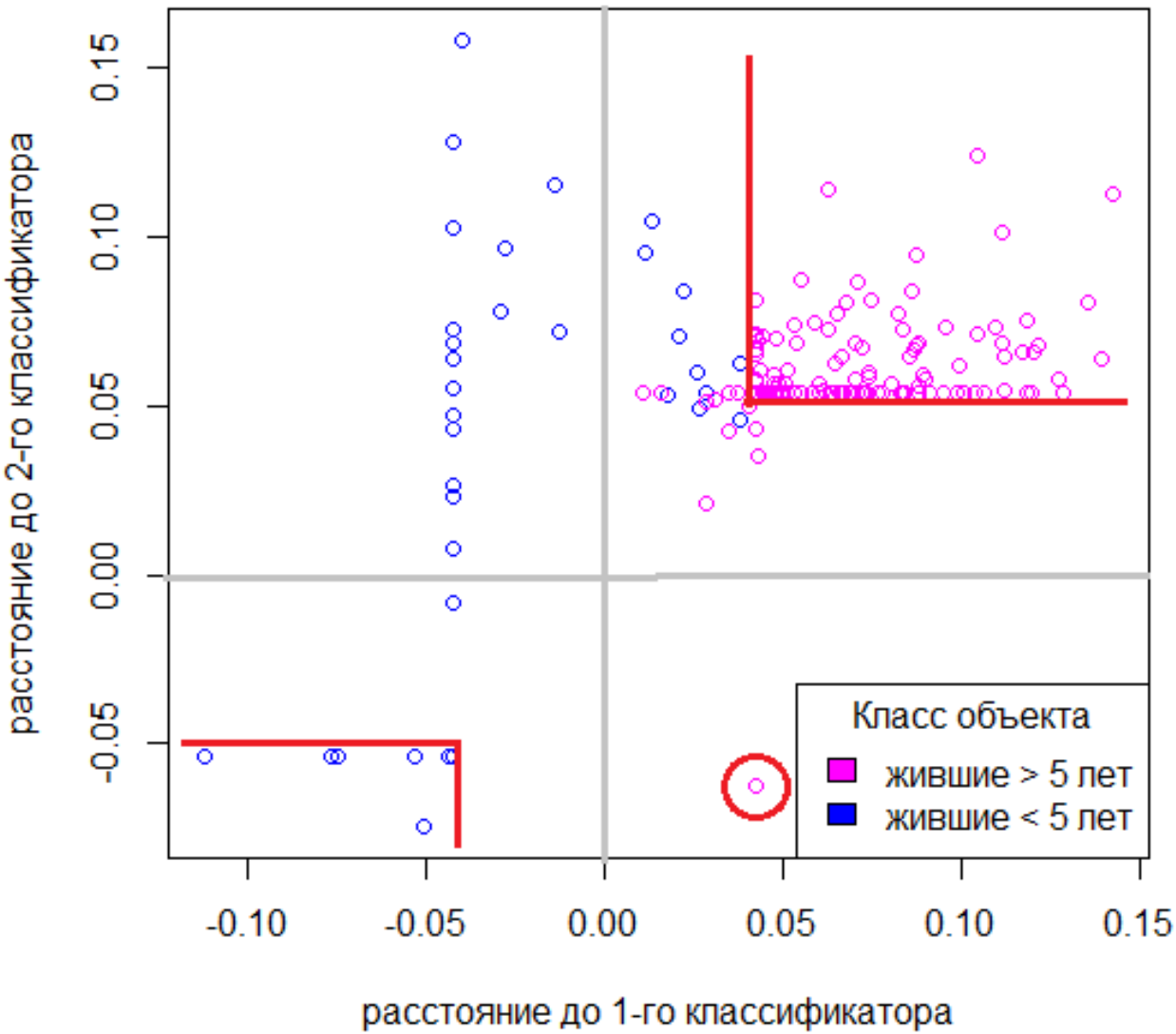
Использование контрацептивов

Значимо различающиеся средние значения в уверенно классифицированных группах:

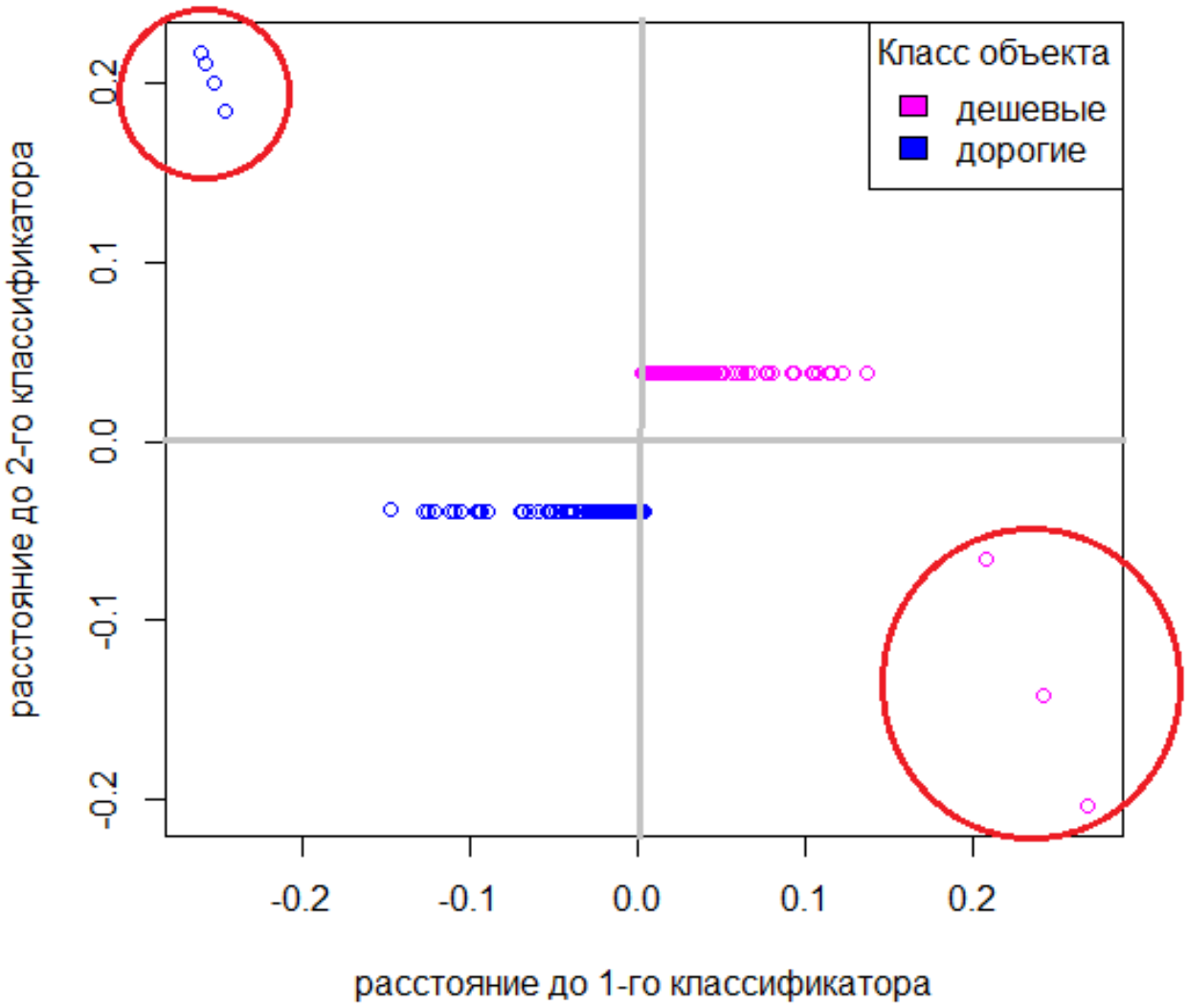
	use	not use
уровень образования женщины	3.3	2.3
количество детей	3.8	2.6
уровень жизни	3.3	2.7
% исламских семей	0.78	0.91
% работающих женщин	0.76	0.68

Результаты имеют естественную интерпретацию.

Выживаемость после рака груди



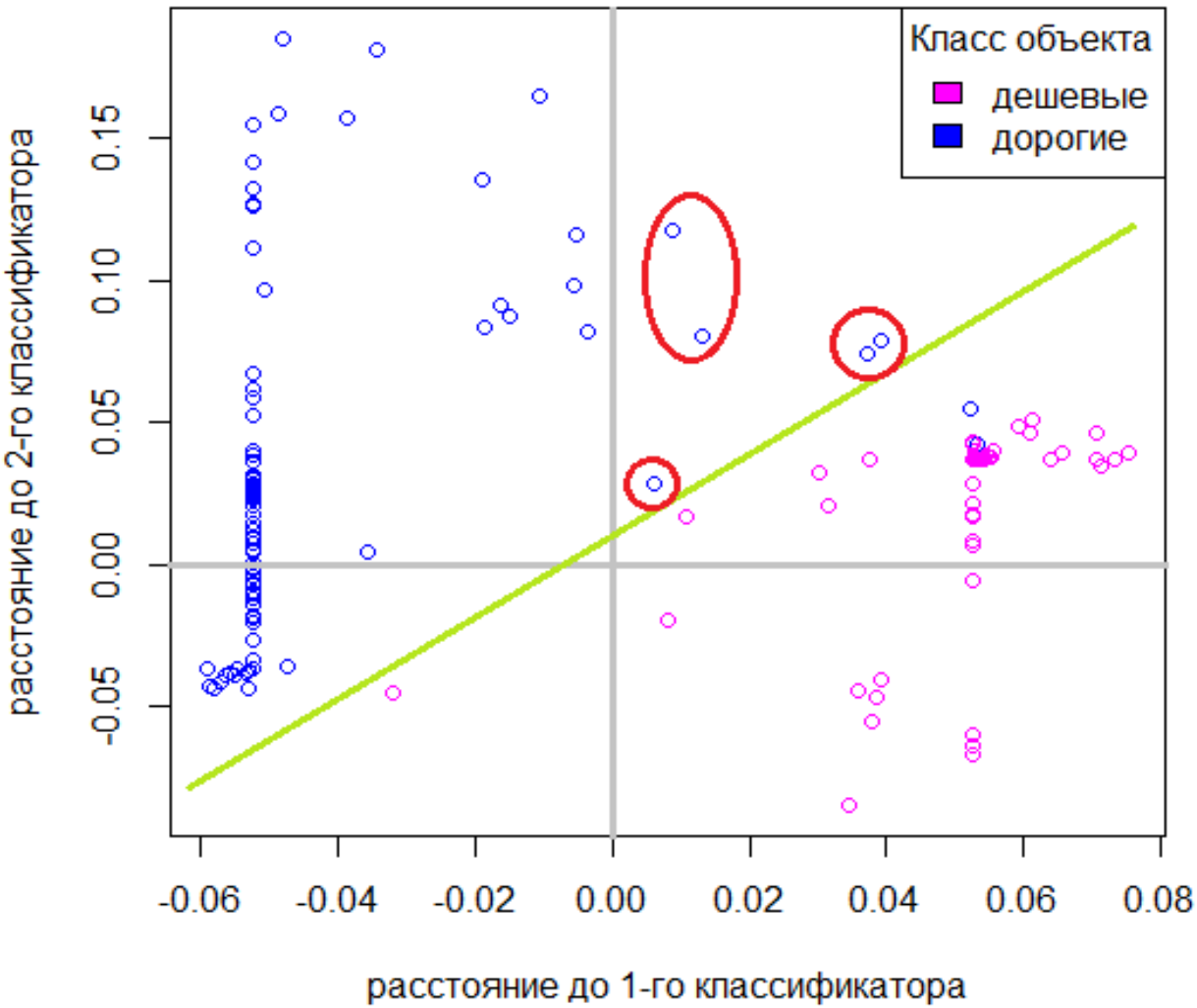
Бостон


Класс объекта
дешевые
дорогие
расстояние до 2-го классификатора
расстояние до 1-го классификатора
-0.2
-0.1
0.0
0.1
0.2

Бостон

crim	indus	nox	rm	age	dis	rd	tax	b	lstat	classes
0.05	3.4	0.49	6.42	66.1	3.09	2	270	392	8.8	2
0.04	3.4	0.49	6.41	73.9	3.09	2	270	394	8.2	1
1.19	21.9	0.62	6.33	97.7	2.27	4	437	397	12.3	1
0.59	21.9	0.62	6.37	97.9	2.32	4	437	386	11.1	2
2.30	19.6	0.61	6.32	96.1	2.10	5	403	297	11.1	2
2.45	19.6	0.61	6.40	95.2	2.26	5	403	330	11.3	1
0.06	11.9	0.57	6.98	91.0	2.17	1	273	397	5.6	2
0.11	11.9	0.57	6.79	89.3	2.39	1	273	393	6.5	1

Бостон


Класс объекта
дешевые
дорогие
расстояние до 2-го классификатора
-0.05
0.00
0.05
0.10
0.15
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
расстояние до 1-го классификатора

Бостон

- Данные были случайным образом разделены на обучающую и тестовую. выборки, так что обучающая составляла 70% от исходных данных, а тестовая — 30%
- На обучающей выборке были построены 2 классификатора.
- Были найдены расстояния до обоих классификаторов, как на тестовой, так и на обучающей выборках.
- На новых координатах обучающей выборки был построен линейный классификатор.

Качество классификации на тестовой выборке улучшилось

	rfb, исходные данные	линейное, новые координаты
test	68%	83%

Итог

- Составлена квадратичная задача для n классификаторов.
- Реализован модифицированный алгоритм SMO на языке R.
- Предложен способ визуализации, позволяющий оценить уверенность классификации для каждого наблюдения.
- Работа предложенного метода проверена и исследована на реальных данных, а именно
  - выявлены признаки важные для классификатора, путем оценки средних значений признаков в уверенно классифицированных группах;
  - на некоторых данных увеличено качество классификации.

Теорема Мерсера и примеры ядер

Определение

Функция $K : X \times X \to \mathbb{R}$ называется ядром, если она представима в виде $K(x, y) = \langle \phi(x), \phi(y) \rangle_H$ при некотором отображении $\phi : X \to H$, где H - пространство со скалярным произведением.

Теорема

Функция $K(x, x')$ является ядром тогда и только тогда, когда она симметрична, $K(x, x') = K(x', x)$, непрерывна по обоим аргументам и неотрицательно определена, т.е. для любой конечной выборки $X^p = (x_1, \ldots, x_p)$ из X матрица $K = \|K(x_i, x_j)\|$ неотрицательно определена.

Стандартные ядра:

- полиномиальное $K(x, x') = (\langle x, x' \rangle + 1)^d$;
- RBF $K(x, x') = \exp(-\beta \|x - x'\|^2)$.