Санкт-Петербургский Политехнический Университет им. Петра Великого

Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №3 по дисциплине “Теория принятия экономических решений”

**Классификация объектов с помощью алгоритма SVM**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Павлова Людмила Владимировна

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 3](#_Toc60105927)

[1.1 Постановка задачи обучения 3](#_Toc60105928)

[1.2 Сепарабельный случай 3](#_Toc60105929)

[1.3 Несепарабельный случай 4](#_Toc60105930)

[2 Численные эксперименты 4](#_Toc60105931)

[2.1 Линейно разделимые данные 4](#_Toc60105932)

[2.2 Линейно разделимые данные с выбросами 5](#_Toc60105933)

[2.3 Линейно неразделимые данные с выбросами 8](#_Toc60105934)

[2.4 Датасет German 9](#_Toc60105935)

[2.4.1 На полном наборе признаков 10](#_Toc60105936)

[2.4.2 По трём главных компонентам 10](#_Toc60105937)

[3 Реализация 10](#_Toc60105938)

[4 Выводы 11](#_Toc60105939)

[5 Литература 11](#_Toc60105940)

[6 Приложения 11](#_Toc60105941)

# 1 Постановка задачи

## 1.1 Постановка задачи обучения

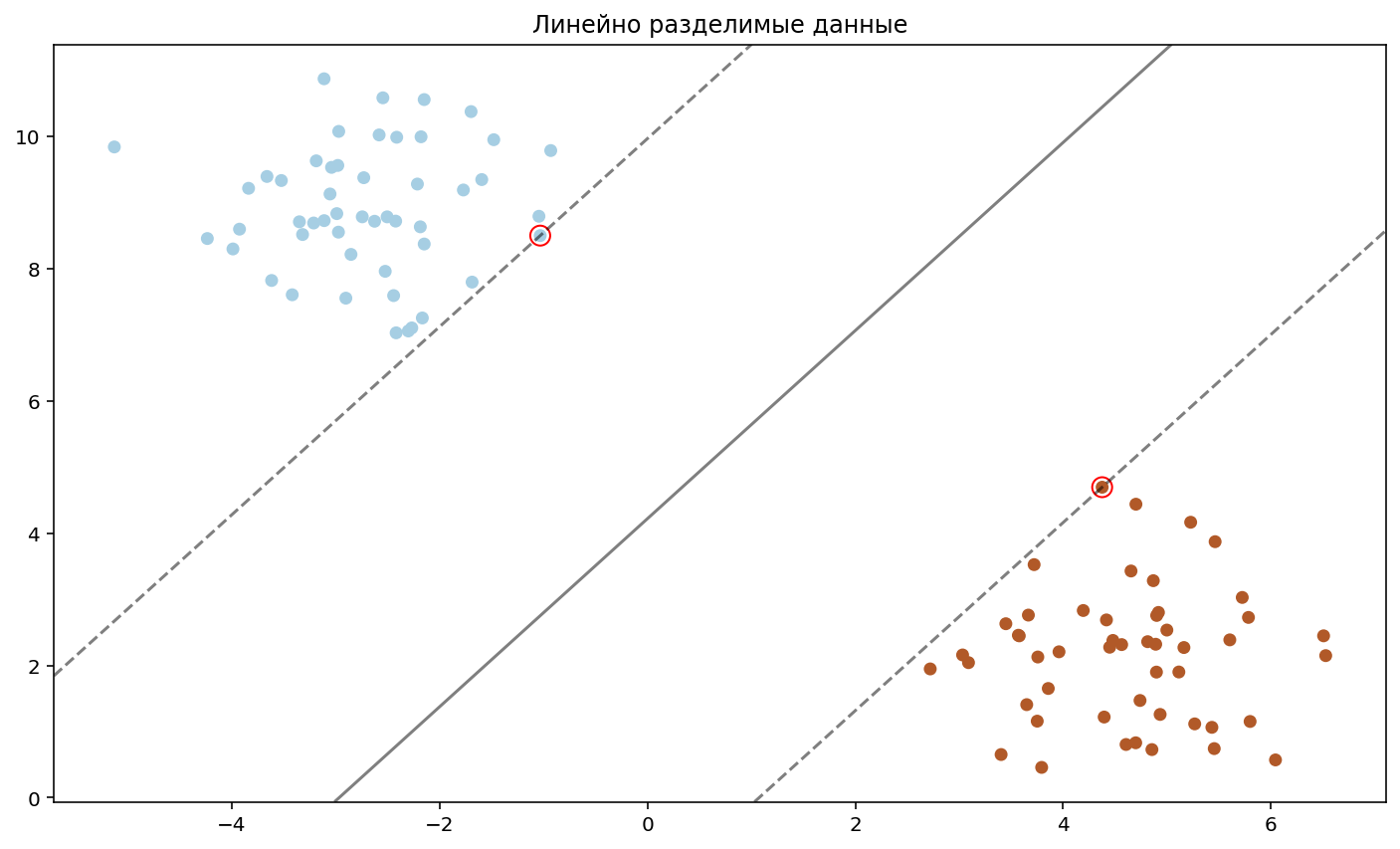
## 1.2 Сепарабельный случай

## 1.3 Несепарабельный случай

# **2 Численные эксперименты**

Будем экспериментировать над данными размера 100, делящимися на 2 класса (метки: 0 и 1). Красным цветом выделяются опорные вектора, для которых коэффициенты Лагранжа .

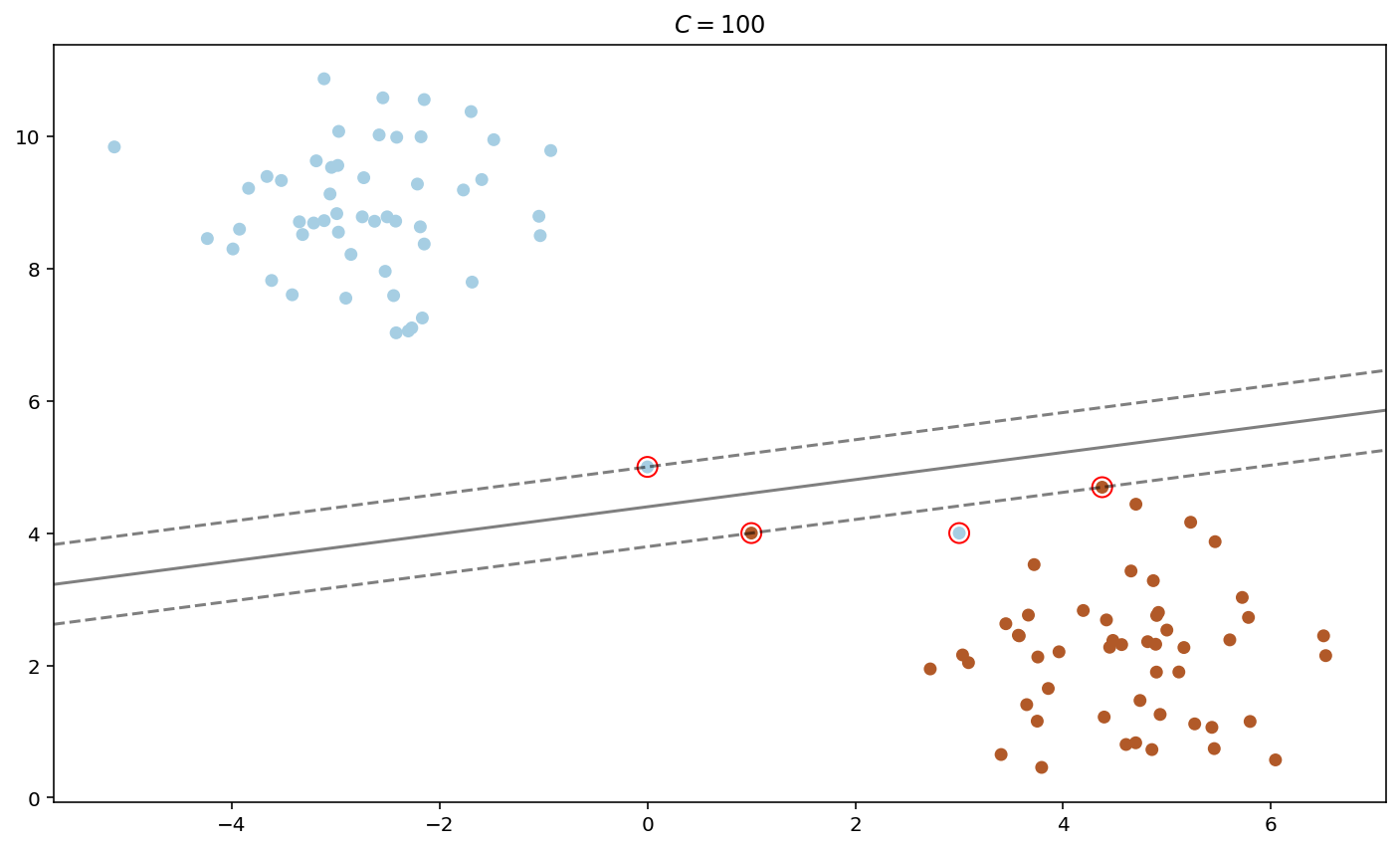
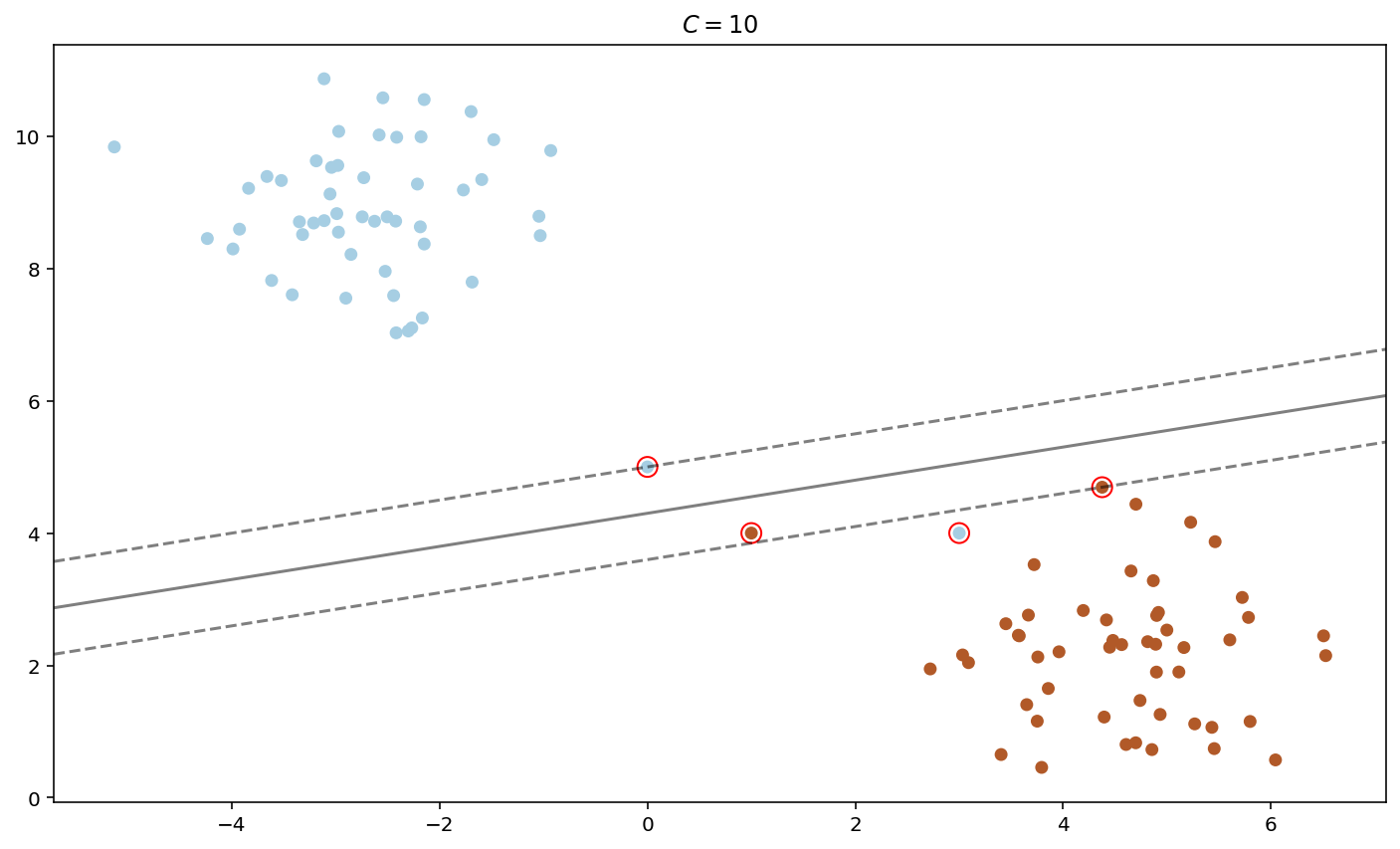
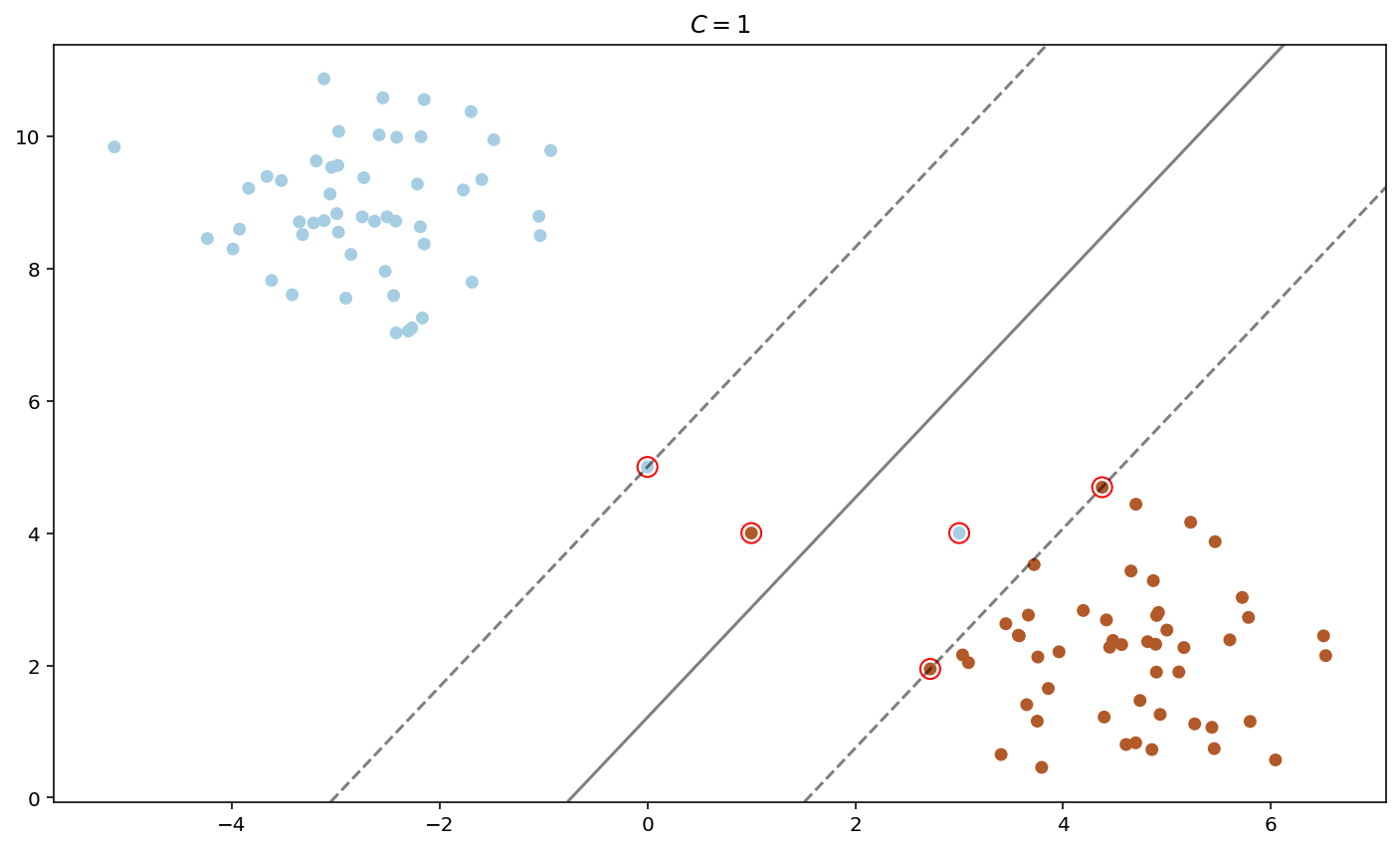
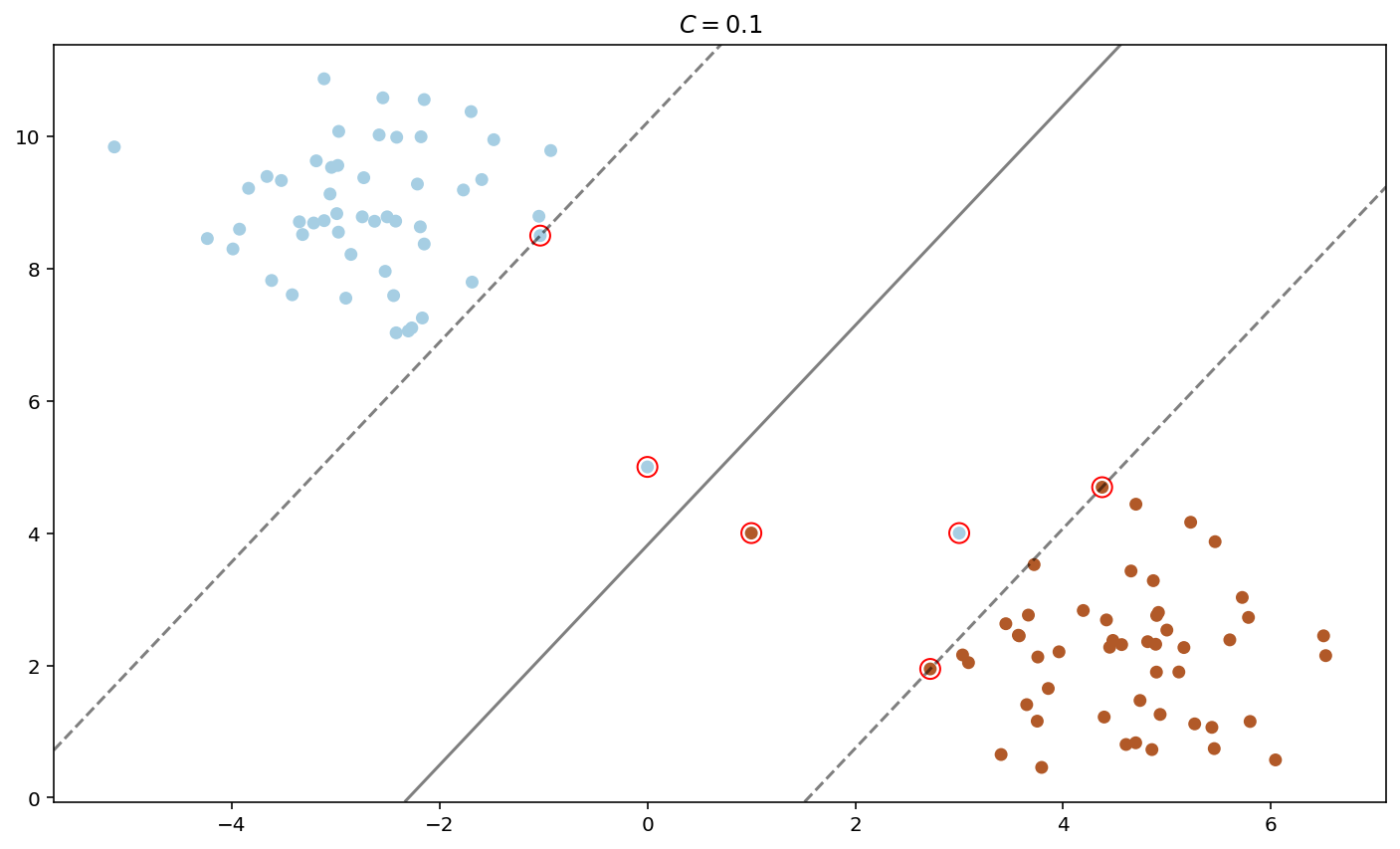
## 2.1 Линейно разделимые данные



## 2.2 Линейно разделимые данные с выбросами

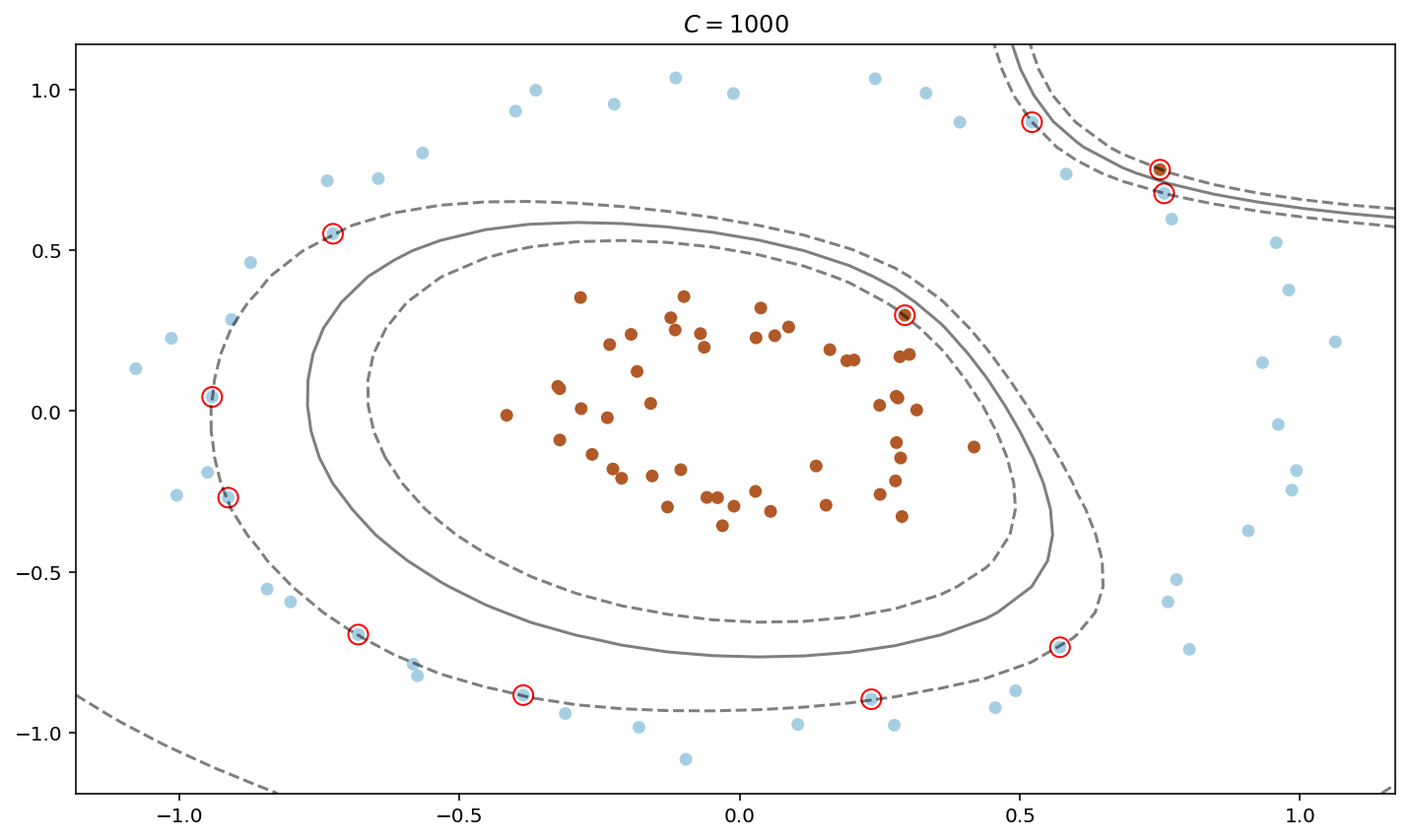
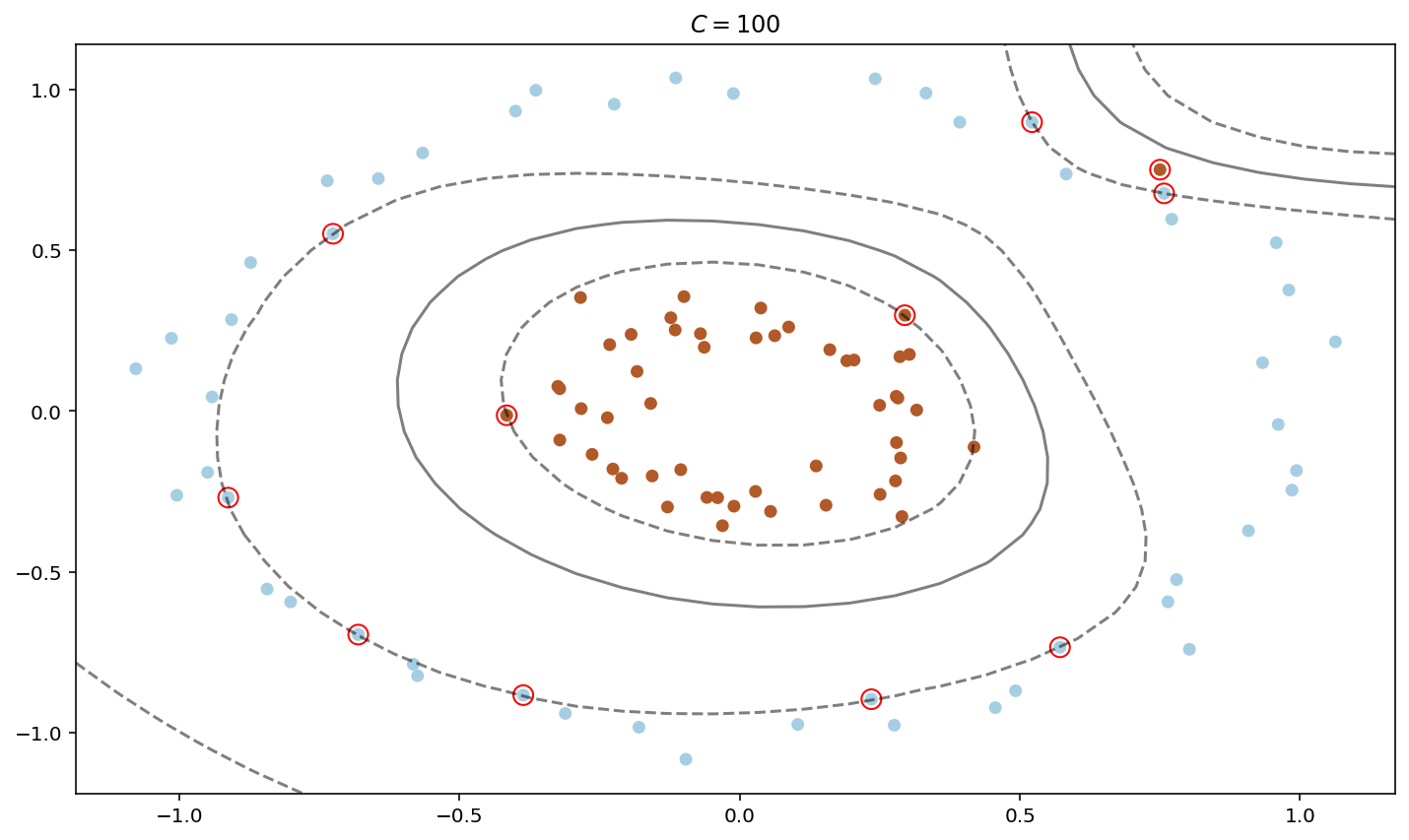
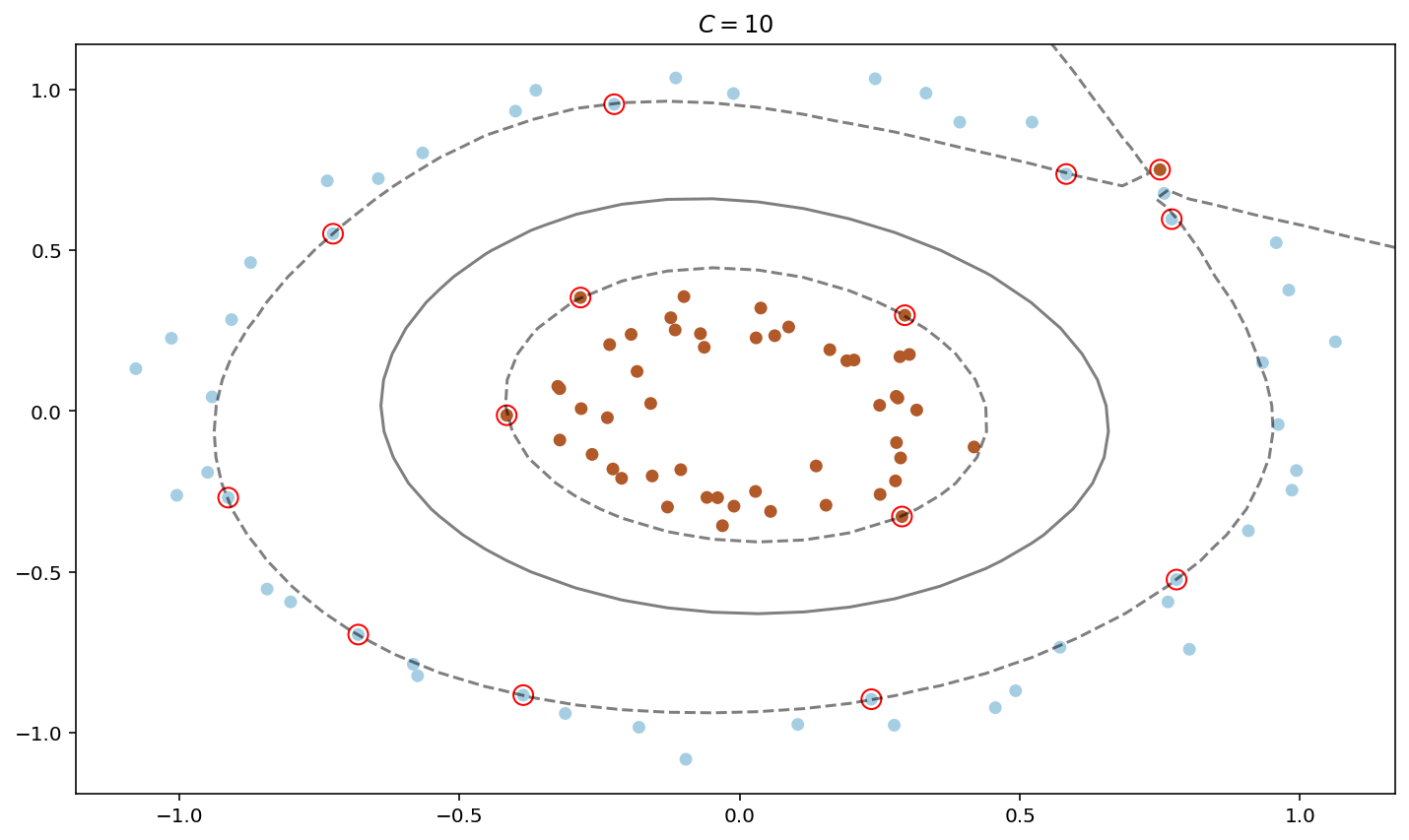
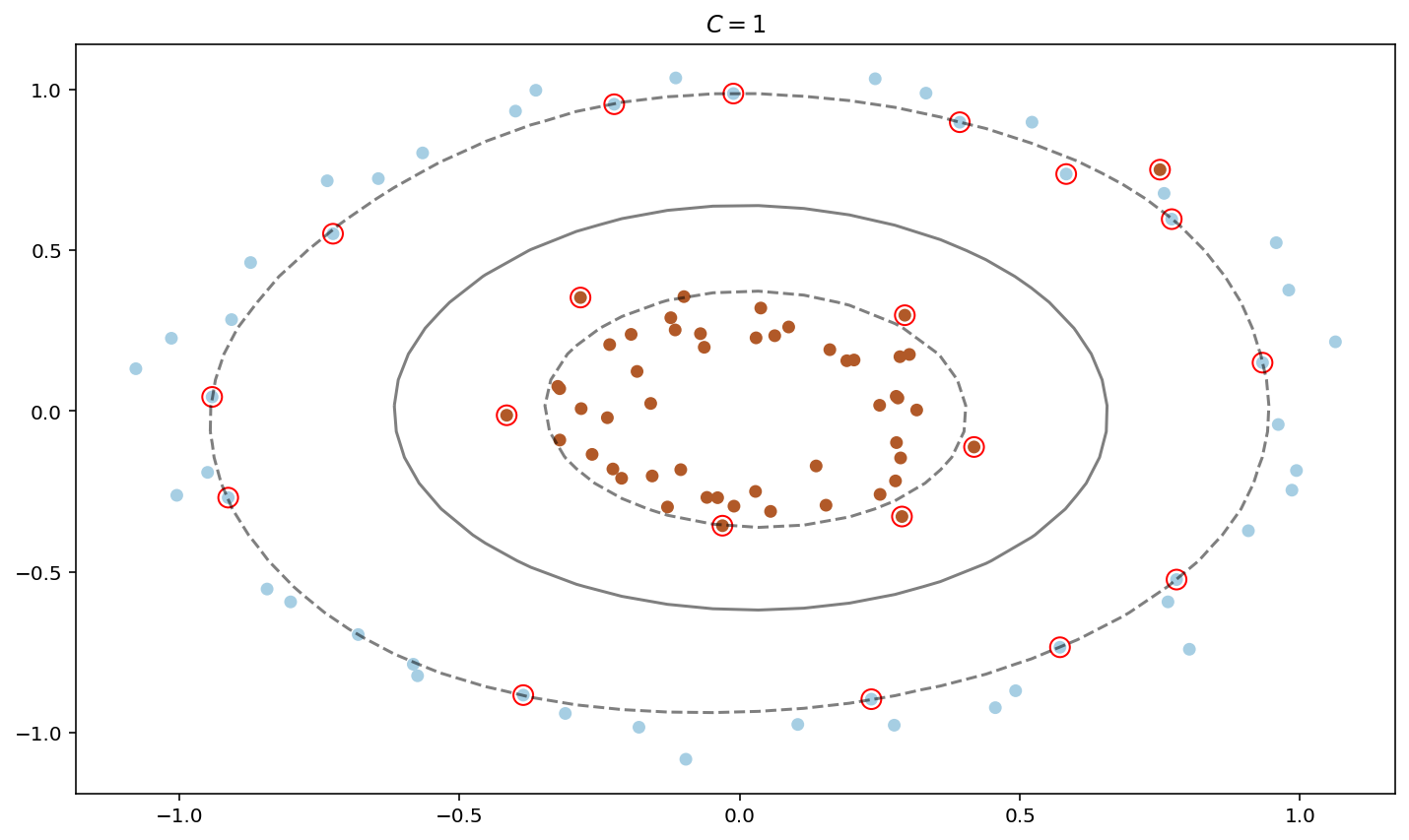
Посмотрим, как значение коэффициента влияет на вид разделяющей гиперплоскости и ширину разделяющей полосы .

|  |  |
| --- | --- |
|  |  |
| **0.1** | 3.296144 |
| **1** | 1.952715 |
| **10** | 0.679880 |
| **100** | 0.590560 |



Чем больше значение коэффициента регуляризации ошибки , тем сильнее разделяющая гиперплоскость и интервал подстраиваются под граничные выбросы.

## 2.3 Линейно неразделимые данные с выбросами

Будем использовать Гауссово ядро с . 

Здесь та же ситуация, что и с выбросами в линейно разделимых данных: гиперплоскость сильно подстраивается под обучающую выборку с выбросами.

## 2.4 Датасет German

Рассмотрим данные из репозитория (файл *german.data-numeric*). Данные представляют собой 1000 объектов (1000 резюме различных людей), относящихся к одному из двух классов. Каждый объект характеризует набор из 24 численных признаков, являющихся характеристиками материального, финансового, семейного положения и его трудоустройства.

К первому классу относятся объектов, ко второму – . Проведём разделение каждого класса на ОВ и ТВ случайным образом в соотношении 80 % : 20 %.

### 2.4.1 На полном наборе признаков

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **0.1000** | **0.0100** | **0.0010** | **0.0001** |
| **0.1** | 70.125 | 70.125 | 70.125 | 70.125 |
| **1.0** | 99.250 | 83.375 | 72.875 | 71.250 |
| **10.0** | 100.000 | 97.250 | 83.000 | 72.500 |
| **100.0** | 100.000 | 99.875 | 88.125 | 80.000 |

Лучшая точность на обучающей выборке: 100 % при .

Точность на тестовой выборке с данными параметрами: 70 %.

Судя по результатам для ОВ и ТВ, можно сказать, что построенная SVM-машина сильно приспособилась к ОВ, из-за чего разница между точностями составляет 30 %.

### 2.4.2 По трём главных компонентам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **0.1000** | **0.0100** | **0.0010** | **0.0001** |
| **0.1** | 70.125 | 70.125 | 70.125 | 70.125 |
| **1.0** | 88.625 | 77.250 | 72.500 | 71.125 |
| **10.0** | 95.375 | 82.500 | 74.875 | 71.625 |
| **100.0** | 97.875 | 87.375 | 75.625 | 71.750 |

Лучшая точность на обучающей выборке: 97.875 % при .

Точность на тестовой выборке с данными параметрами: 64 %.

Проценты от общей дисперсии первых 3-х главных компонент:

1-я главная компонента: 79.42693195 %

2-я главная компонента: 12.05481689 %

3-я главная компонента: 7.46295661 %

При переходе к трём главным компонентам результаты не сильно изменились: разница 33,875 %.

# 3 Реализация

Была использована среда *IPython Notebook* (язык *Python 3.8.2*): модули *numpy* для работы с массивами и математических расчётов, *pandas* - для хранения данных в таблицах, *matplotlib.pyplot* – для построения графиков; функции *decomposition.PCA, preprocessing.StandartScaler, model\_selection.train\_test\_split, pipeline.make\_pipeline, datasets.make\_blobs, datasets.make\_circles* из модуля *sklearn* для работы с выборками и *display* из модуля *IPython.display* для отображения pandas’ских таблиц в браузере.

# 4 Выводы

SVM обладает преимуществом перед стандартными линейными классификаторами, т. к. строит максимальную ширину разделяющей полосы между классами.

Главный недостаток SVM – он не является робастным, т. е. для данных с выбросами, находящимися внутри разделяющей полосы, разделяющая гиперплоскость всё сильнее и сильнее с увеличением коэффициента регуляризации ошибки подстраивается под граничные выбросы.

Так для данных нужно специально подбирать ядро.

# 5 Литература

[Основы работы с *numpy* (отдельная глава курса)](https://stepik.org/course/401)

Лекции по ТПЭР, “алгоритм SVM”, Павлова Л. В., 2020 г.

# 6 Приложения

[Код лабораторной](https://github.com/MeShootIn/economics-decision-theory/blob/main/lab_3/main.ipynb)