Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшегообразования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчет по лабораторной работе №4

по дисциплине «Прикладная математика»

Авторы: Власов Роман, Высоцкая Валерия, Тихомиров Дмитрий

Факультет: Информационных технологий и программирования

Группа: M33021

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург, 2022

Машины опорных векторов (support vector machine, SVM) один из крайне популярных алгоритмов машинного обучения. Данное семейство алгоритмов может применяться как для решения задач классификации, так и для задач регрессии. С одной стороны, он относится к классу линейных моделей. И несмотря на свою простоту может давать уверенные результаты. С другой стороны, алгоритм допускает решение задач классификации в случае, если выборка не является линейно разделимой. Данный подход (kernel trick) существенно расширяет возможности алгоритма, позволяя ему быть (даже буквально, геометрически) более гибким, чем другие линейные модели классификации.

Ход выполнения работы

1. Реализовать генератор входных данных, которые будут использоваться для обучения алгоритма и анализа качества обучения с помощью метрик после его обучения. Требования:

(a) Признаки: $(x, y) \in [-1, 1] \times [-1, 1]$. Иными словами, пространство признаков - квадрат в плоскости R^2.

(b) Граница разделения классов: $x^2 + y^2 = 1/4$. Объекты одного класса лежат внутри окружности $R = 1/2$, объекты другого класса лежат вне окружности.

(c) Входной параметр генератора: размер выборки.

2. Реализовать функции метрик качества: accurace, precision, recall, F-мера. Входные данные: истинные метки классов, предсказанные метки классов. Выходные данные: значение метрики

3. Обучить ансамбль моделей NuSVC с различными условиями:

(a) Выбор ядра SVM (линейное, полиномиальное, гауссово (rbf), сигмоид). Построить графически классы с разными метками, а также разделяющую гиперповерхность для каждого из ядер. Объем обучающей выборки произвольный, но одинаковый для сравнения построенной поверхности для различных ядер. Сравнить метрики качества в зависимости от выбора ядра.

(b) Объем обучающей выборки. Исследовать зависимость метрик качества от объема обучающей выборки.

(c) Параметр ν - нижняя граница доли опорных векторов. Сравнить графически и на основе метрик качества. В пунктах (b) и (c) ядро допускается выбрать фиксированным, например rbf

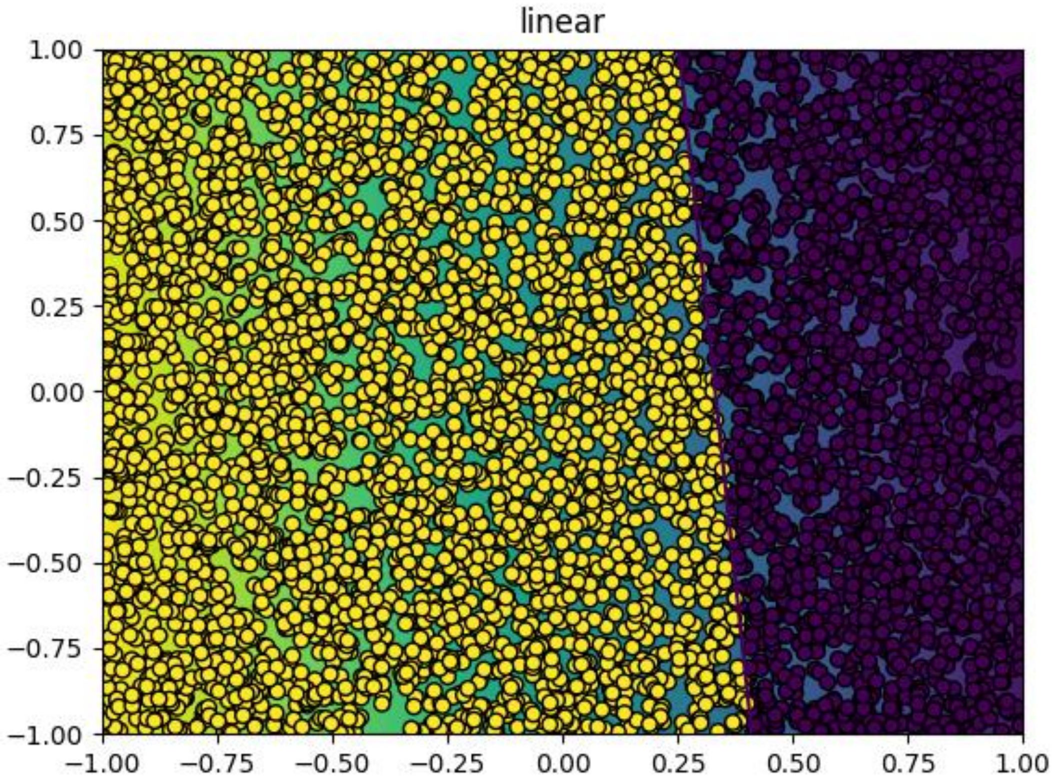
Теория

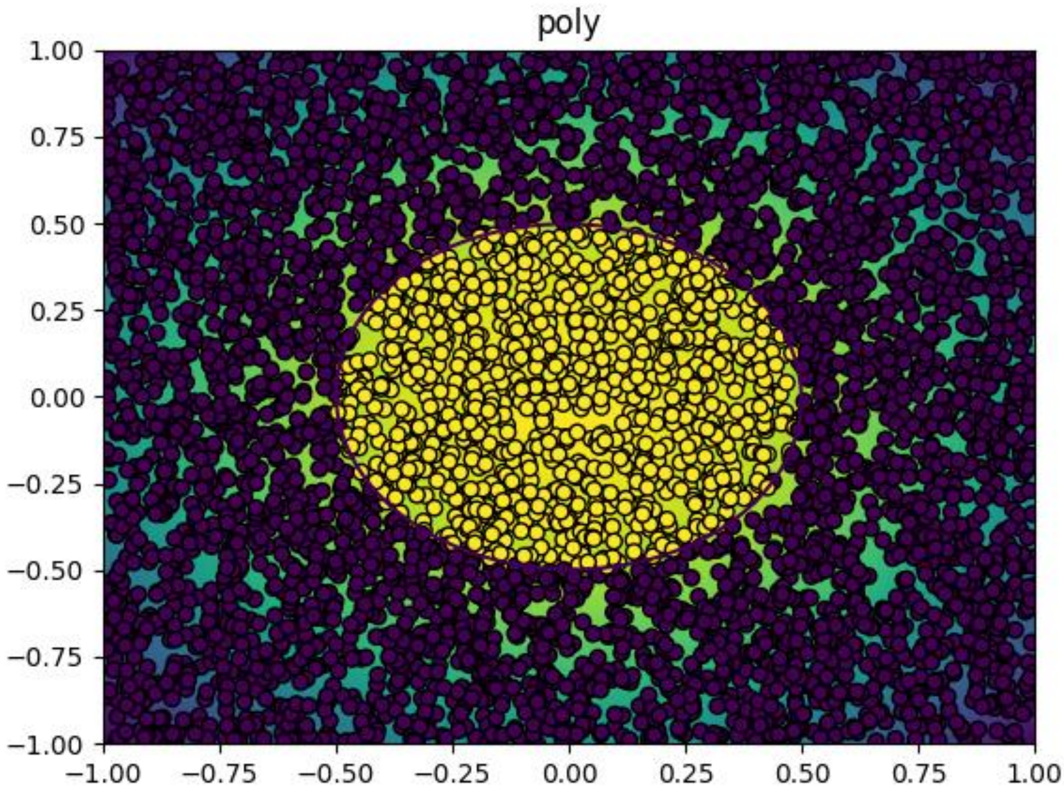
Основная идея метода заключается в отображение векторов пространства признаков, представляющих классифицируемый объекты, в пространство более высокой размерности. Это связано с тем, что в пространстве большей размерности линейная разделимость множества оказывается выше, чем в пространстве меньшей размерности. После перевода в пространство большей размерности, в нём строится разделяющая гиперплоскость. При этом все векторы, расположенные с одной «стороны» гиперплоскости, относятся к одному классу, а расположенные с другой — ко второму. Также, по обе стороны основной разделяющей гиперплоскости, параллельно ей и на равном расстоянии от неё строятся две вспомогательные гиперплоскости, расстояние между которыми называют зазор.

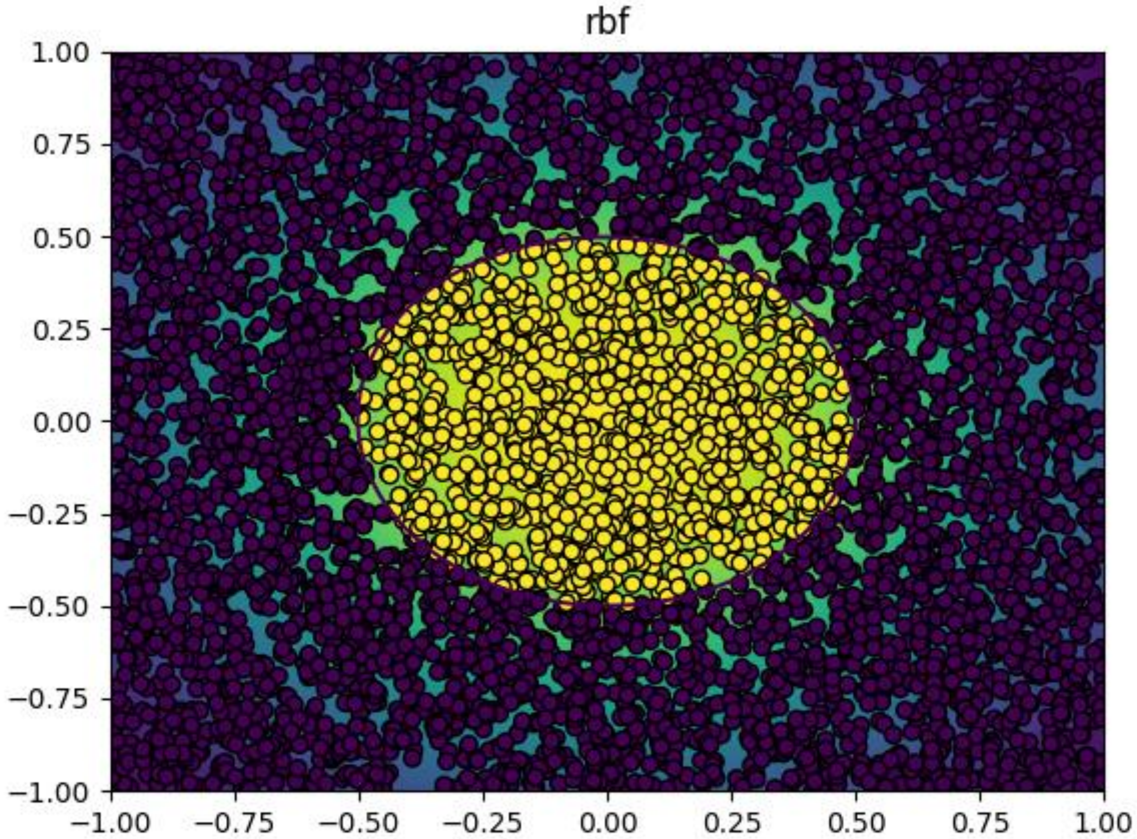
Задача заключается в построении разделяющей гиперплоскость таким образом, чтобы максимизировать зазор — область пространства признаков между вспомогательными гиперплоскостями, в которой не должно быть векторов. Предполагается, что разделяющая гиперплоскость, построенная по данному правилу, обеспечит наиболее уверенное разделение классов и минимизирует среднюю ошибку распознавания.

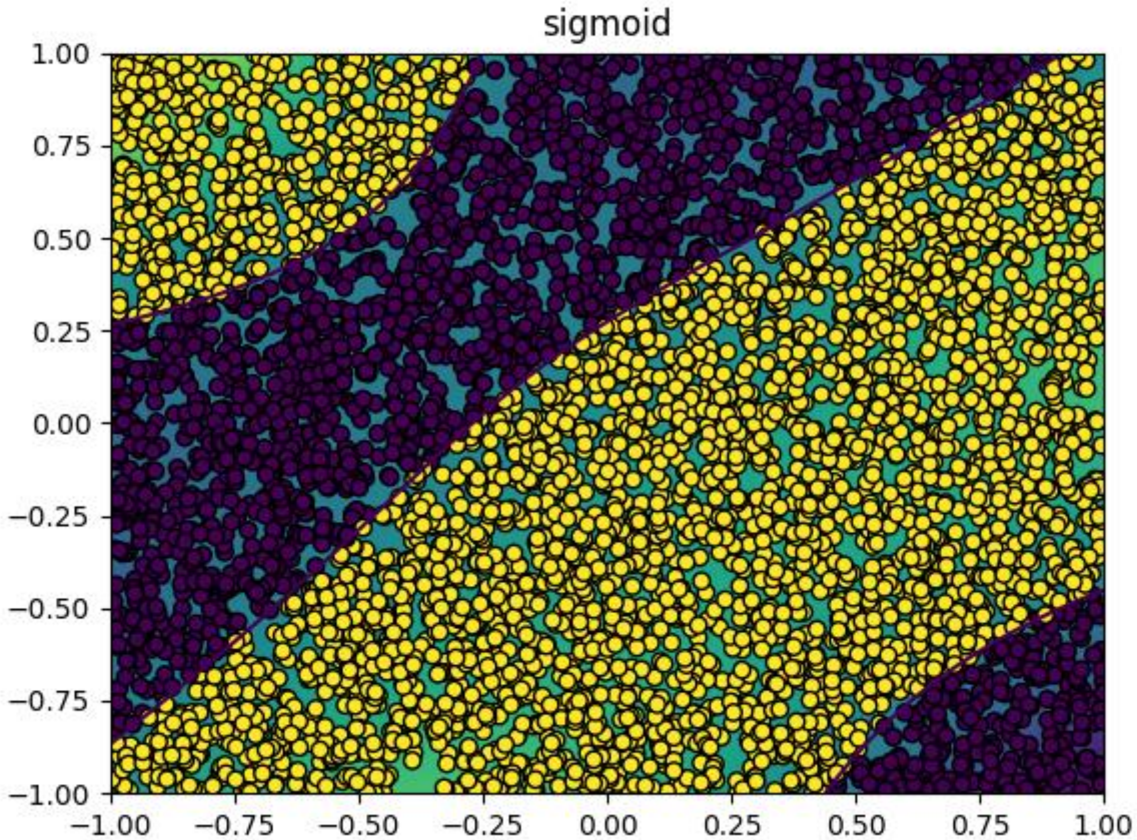
(a) Выбор ядра SVM (n=20000)

	accuracy	precision	recall	F1 score
linear	0.487000	0.264591	0.891129	0.408031
poly(deg=2)	0.998000	0.996936	0.992879	0.994903
rbf	0.998200	0.995964	0.994960	0.995461
sigmoid	0.444600	0.228476	0.757056	0.351017

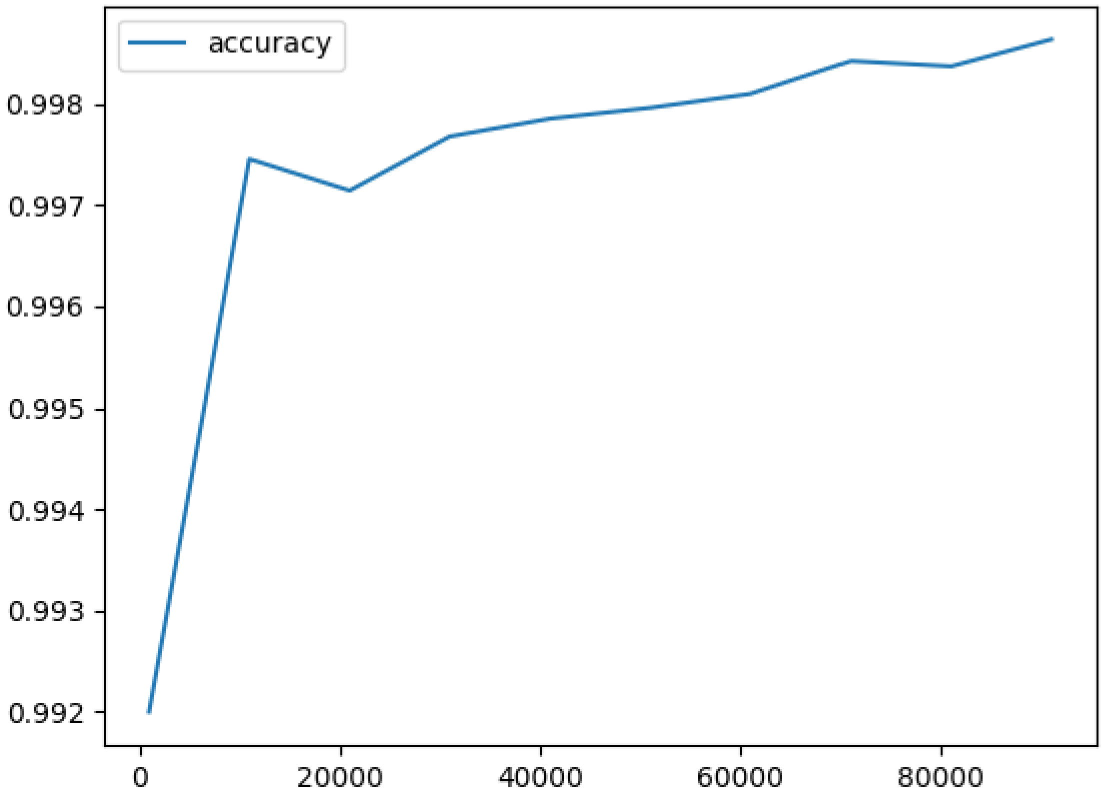


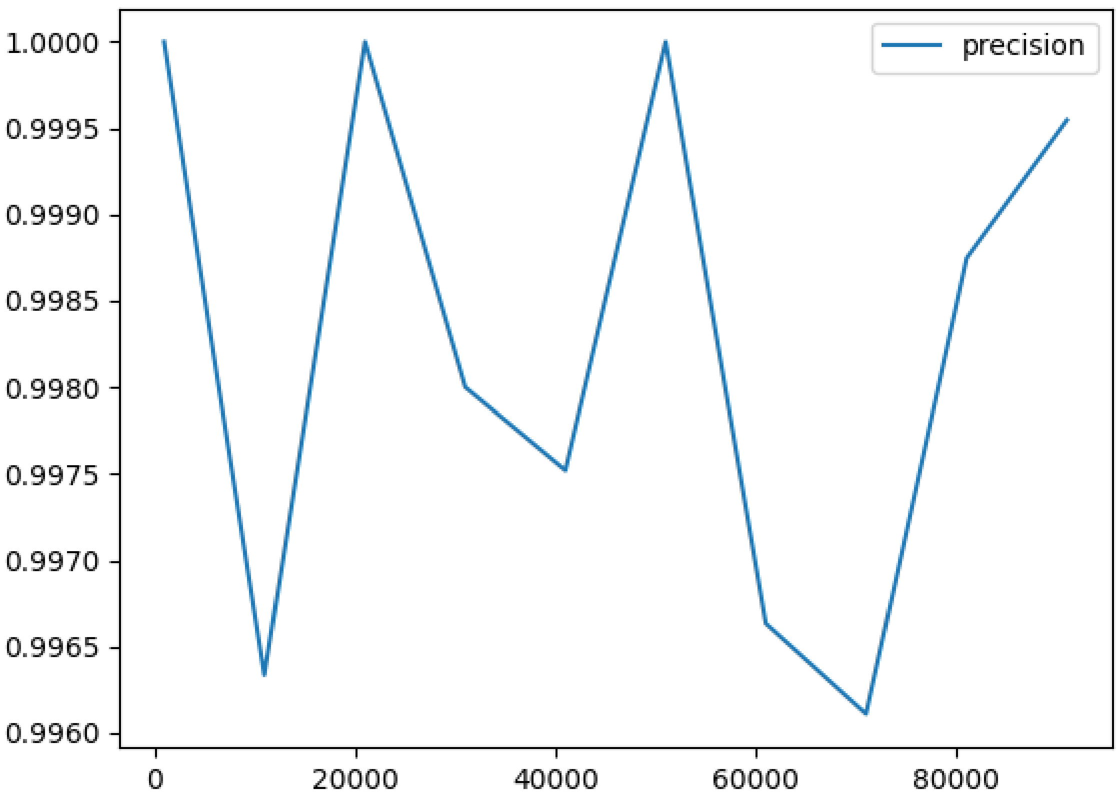


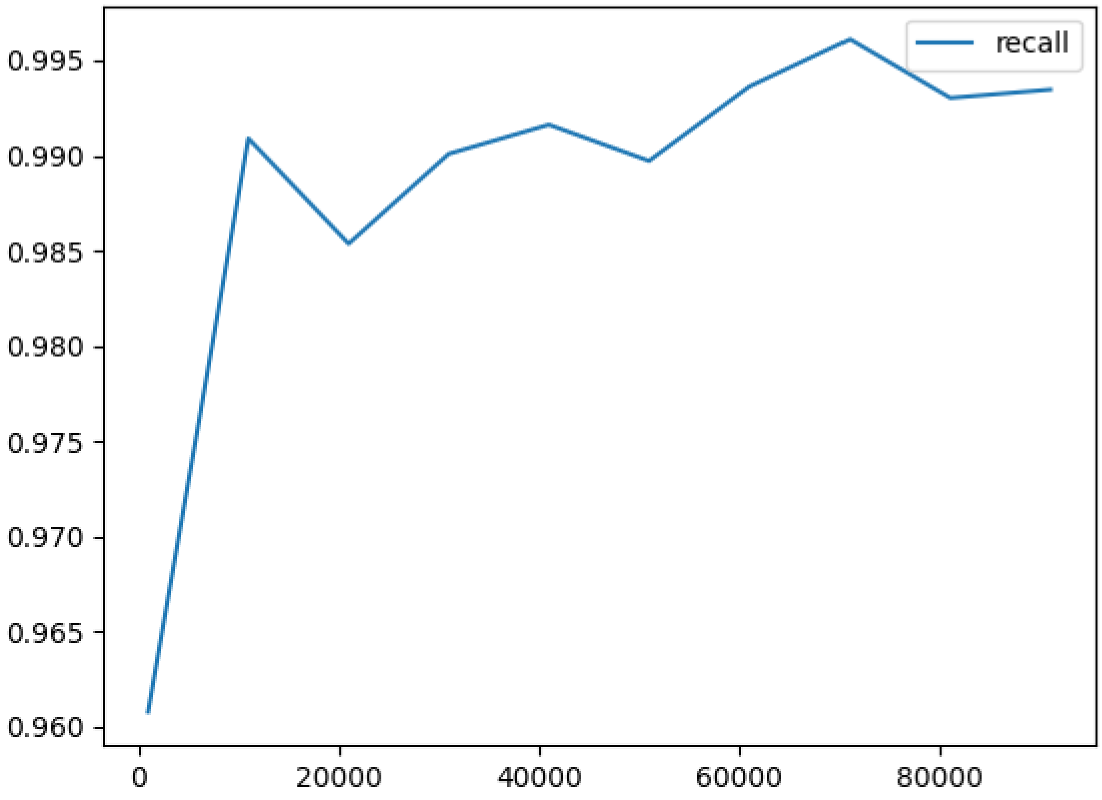

rbf
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00

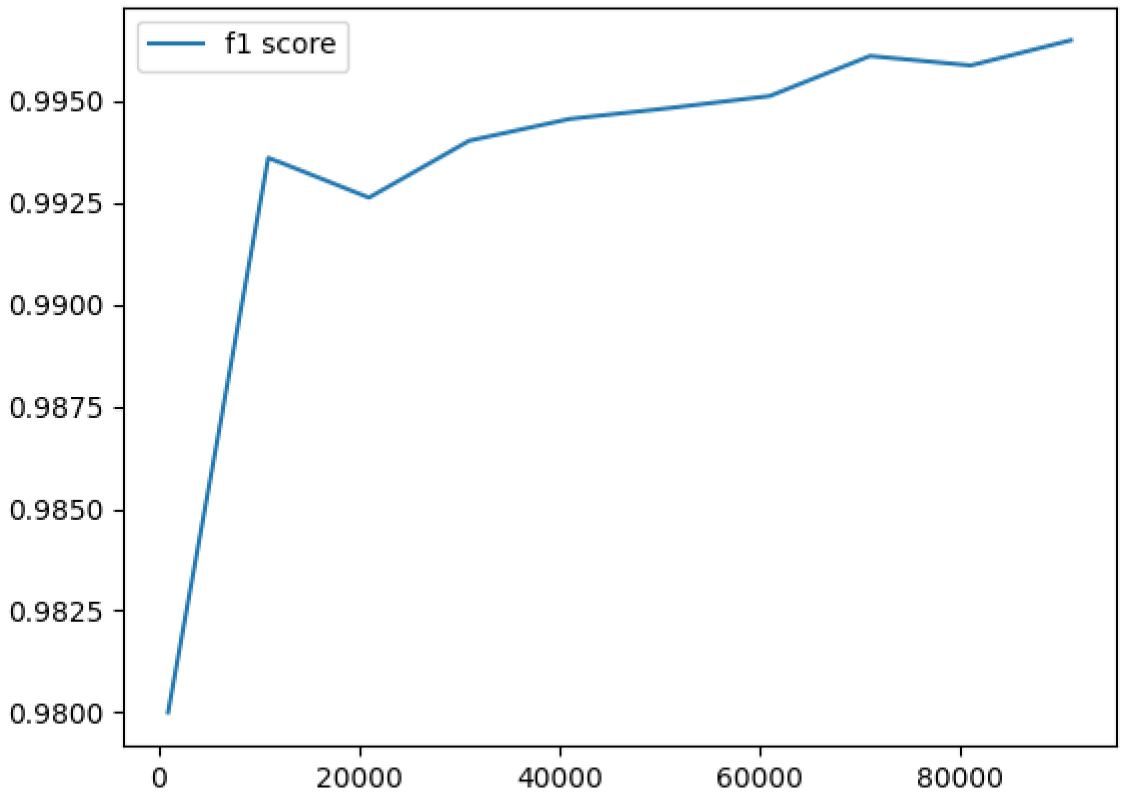

sigmoid
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00

(b) Объем обучающей выборки

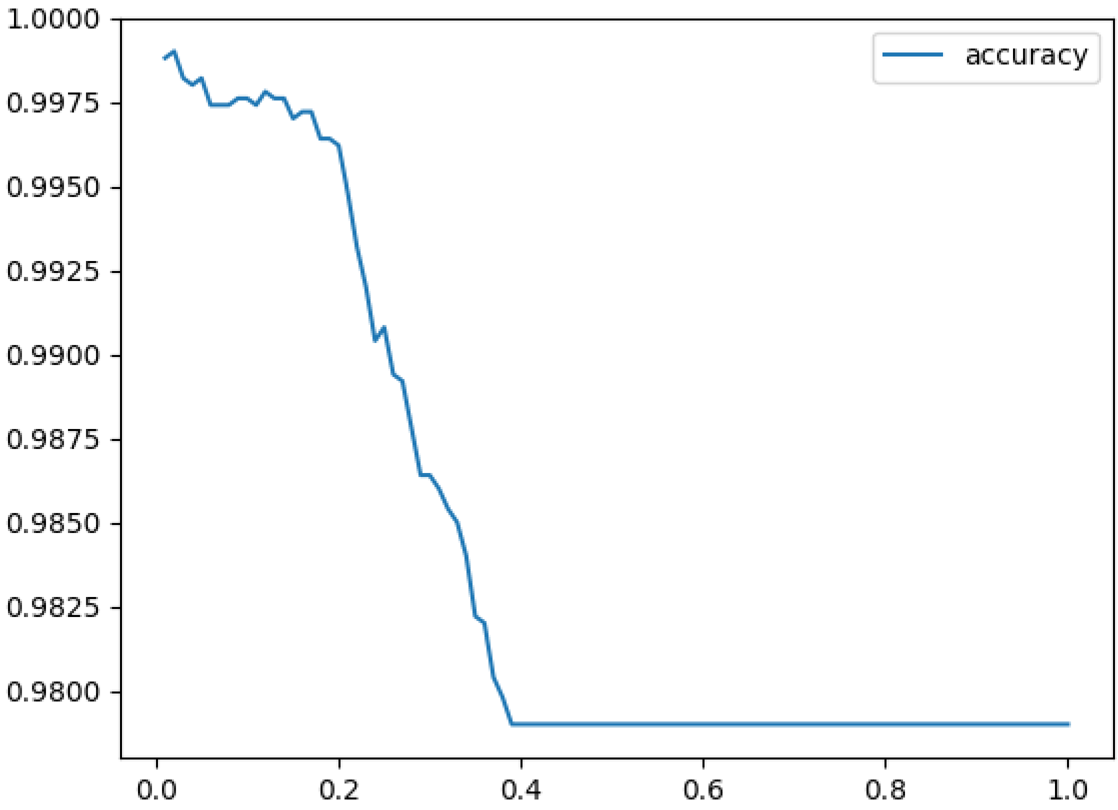


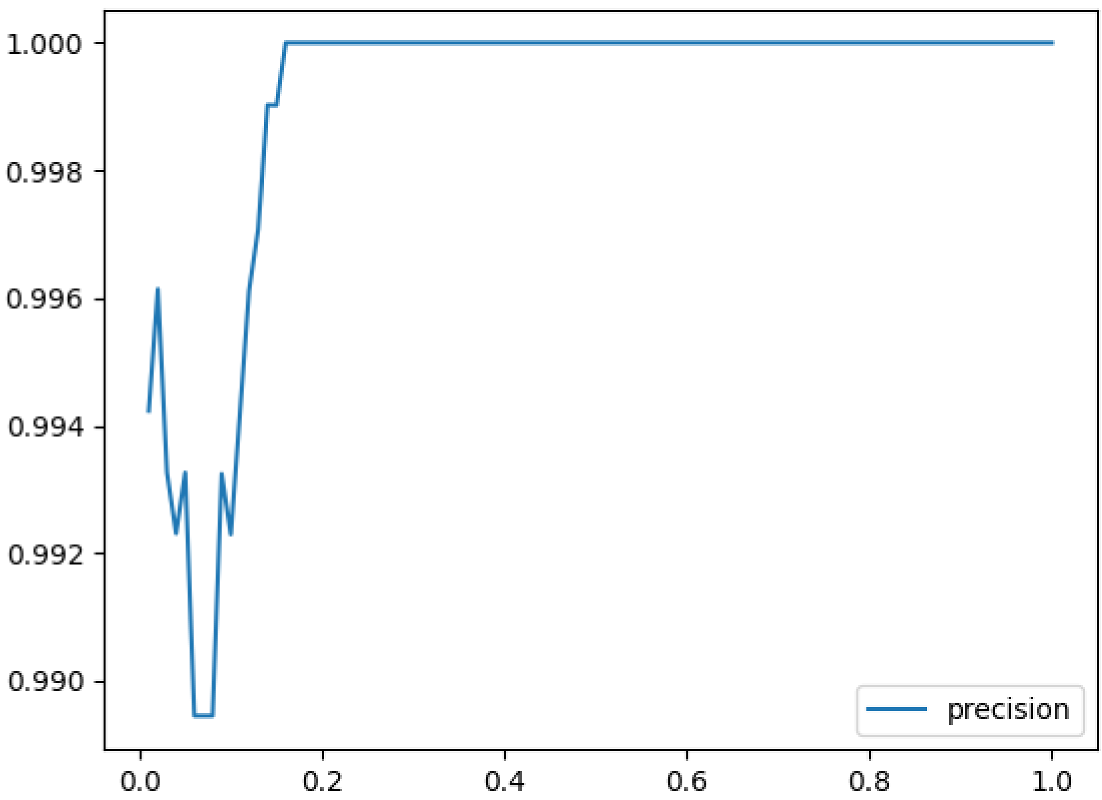


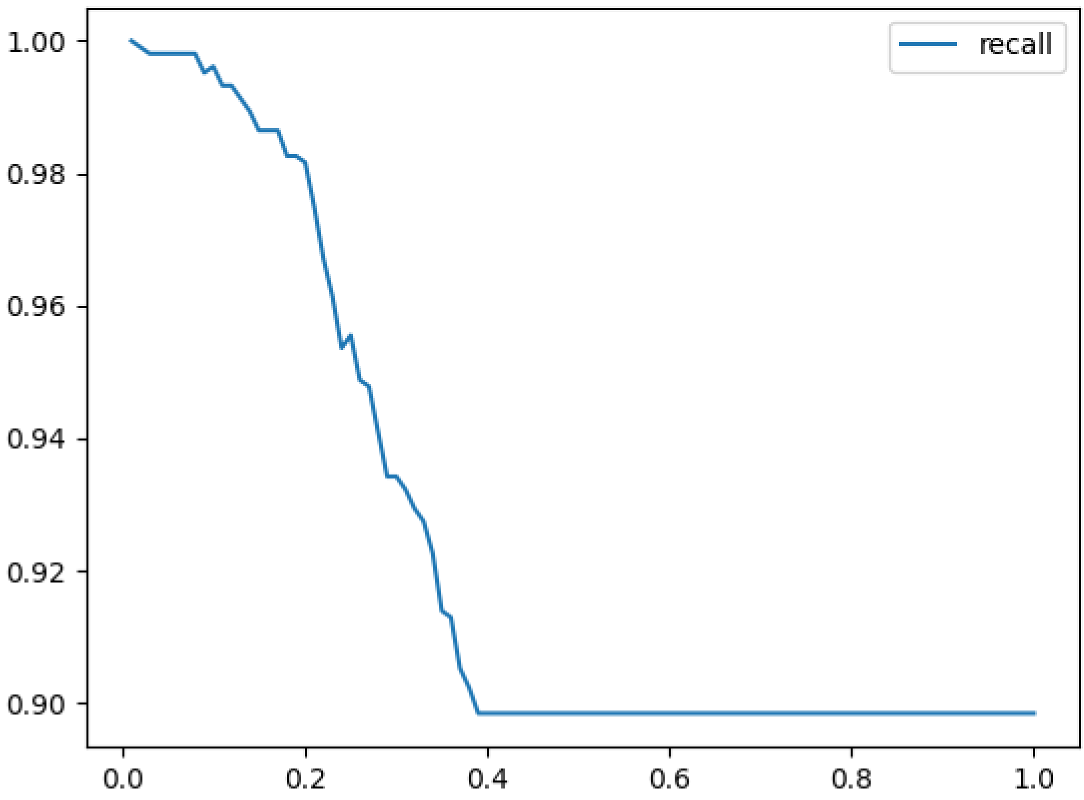

recall
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0
20000
40000
60000
80000

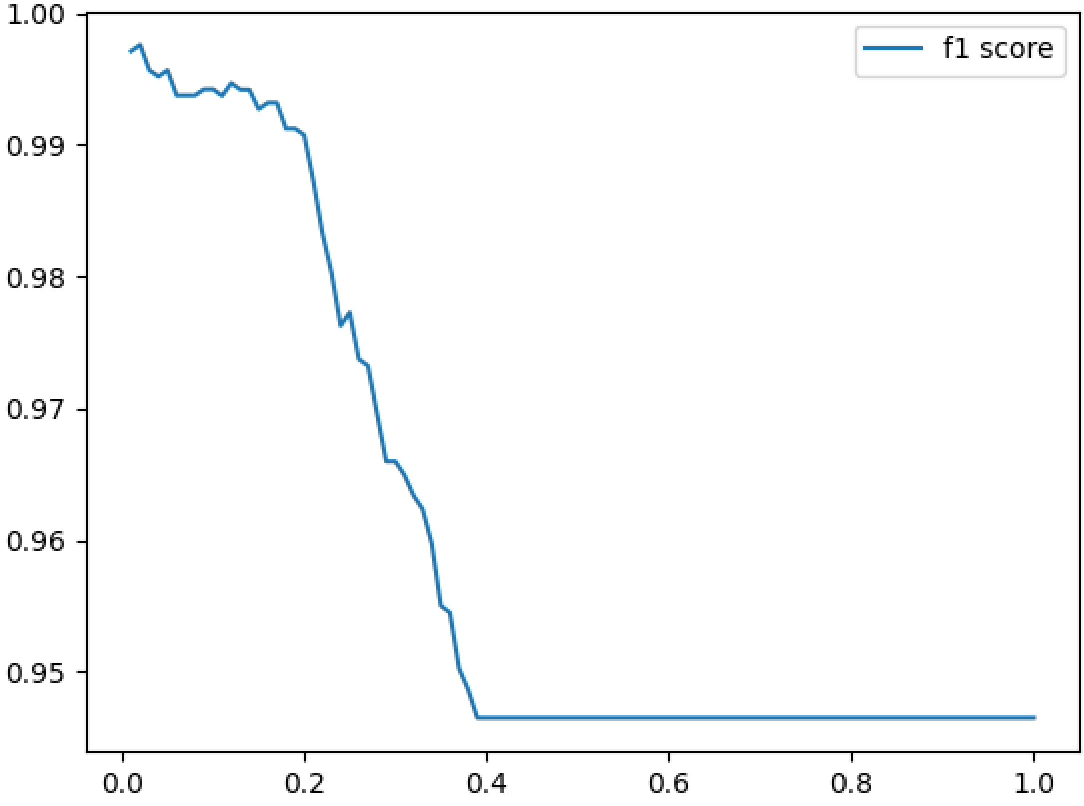

f1 score
0.9950
0.9925
0.9900
0.9875
0.9850
0.9825
0.9800
0
20000
40000
60000
80000

(c) Параметр ν - нижняя граница доли опорных векторов






recall
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
1.0


f1 score
1.00
0.99
0.98
0.97
0.96
0.95
0.0
0.2
0.4
0.6
0.8
1.0

Ссылка на исходный код: https://github.com/appmath-2022/labs/tree/main/Lab8

Вывод:

Изменения nu и размера выборки влияют на конечные результаты модели достаточно минорно. При увеличении выборки, ожидаемо, показатели метрик плавно улучшаются, при увеличении ν наблюдается относительно резкое падение.

В исследовании ядер SVM можно заметить, что хорошо себя показали rbf и полиномиальная со степенью 2. Так как их разделяющие гиперплоскости позволяют максимально точно повторить изначальное деление.

- линейное ядро: $K(x_i, x_j) = x_i x_j$
- полиномиальное ядро со степенью p: $K(x_i, x_j) = (1+x_i x_j)^p$
- RBF: $K(x_i, x_j) = \exp(\gamma||xi-xj||^2)$
- сигмодиное ядро: $K(x_i, x_j) = \tanh(\gamma x_i x+\beta_0)$