МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №8
по дисциплине «Машинное обучение»
Тема: Классификация (линейный дискриминантный анализ, метод опорных векторов)

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

Загрузка данных.

Датасет загружен в датафрейм. Выделены данные и их метки, тексты меток преобразованы к числам. Выборка разбита на обучающую и тестовую train_test_split.

Линейный дискриминантный анализ.

Проведена классификация наблюдений с помощью LinearDiscriminantAnalysis. Выявлено 3 неправильно классифицированных наблюдения. Параметры классификатора представлены в табл. 1. Атрибуты классификатора представлены в табл.2.

Таблица 1 – Параметры *LinearDiscriminantAnalysis*

Параметр	Описание
solver	• «svd»: Разложение по сингулярным числам. Не вычисляет ковариационную матрицу, поэтому рекомендуется для данных с большим количеством признаков. • «lsqr»: Решение наименьших квадратов, можно комбинировать с параметром *shrinkage*. • «eigen»: Разложение на собственные значения, можно комбинировать с параметром *shrinkage*.
shrinkage	• «auto»: Автоматическое сжатие по лемме Ледуа-Вольфа. • float from [0, 1]
priors	Класс априорных вероятностей. По умолчанию пропорции классов выводятся из данных обучения.

n_components	Количество компонентов (<= min (n_classes - 1, n_features)) для уменьшения размерности. Если None, будет установлено значение min (n_classes - 1, n_features). Этот параметр влияет только на метод преобразования *transform*.
store_covariance	Если True, явно вычислить взвешенную ковариационную матрицу внутри класса, когда решатель – «svd». Матрица всегда вычисляется и сохраняется для других решателей.

Таблица 2 – Атрибуты *LinearDiscriminantAnalysis*

Атрибут	Описание
coef_	Весовые вектора.
intercept_	Массив прерывания.
covariance_	Взвешенная внутриклассовая ковариационная матрица.
explained_variance_ratio_	Процент дисперсии, объясняемой каждым из выбранных компонентов. Если n_components не задано, то все компоненты сохраняются, а сумма объясненных дисперсий равна 1,0. Доступно только при использовании собственного решателя или «svd».
means_	Средние в классах.
priors_	Вероятности классов.
scalings_	Масштабирование объектов в пространстве, охватываемом центроидами классов. Доступно только для решателей «svd» и «eigen».
xbar_	Общее среднее. Присутствует, только если решатель - «svd».
classes_	Уникальные метки классов.

Точность классификации с помощью LDA составляет 96%.

Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. График представлен на рис. 1.

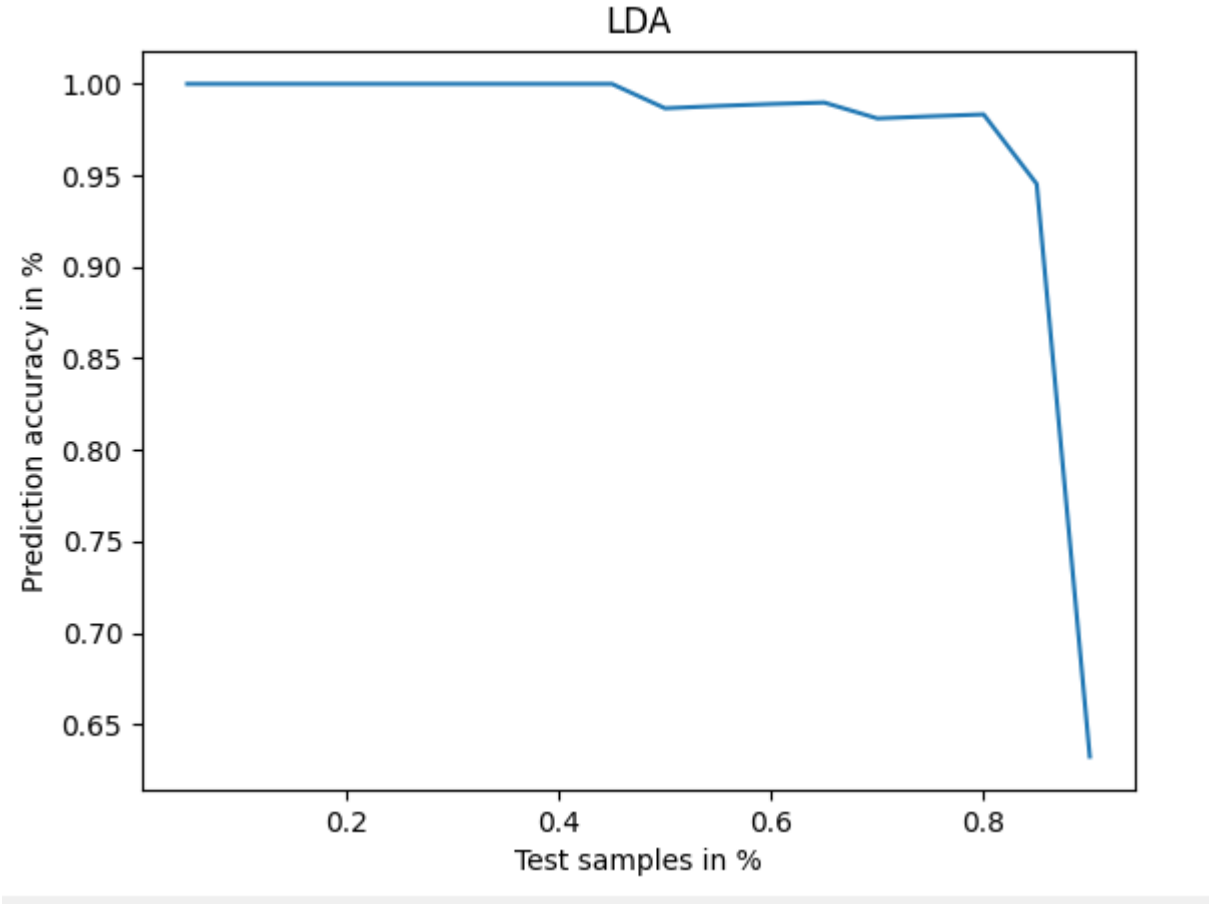

Рисунок 1 – Классификация LDA

Функция transform проецирует данные для максимизации разбиения классов. LDA пытается определить атрибуты, на которые приходится наибольшая разница между классами. В частности, LDA, в отличие от PCA, является контролируемым методом, использующим известные метки классов. Сравнение LDA и PCA представлно на рис. 2 и 3.

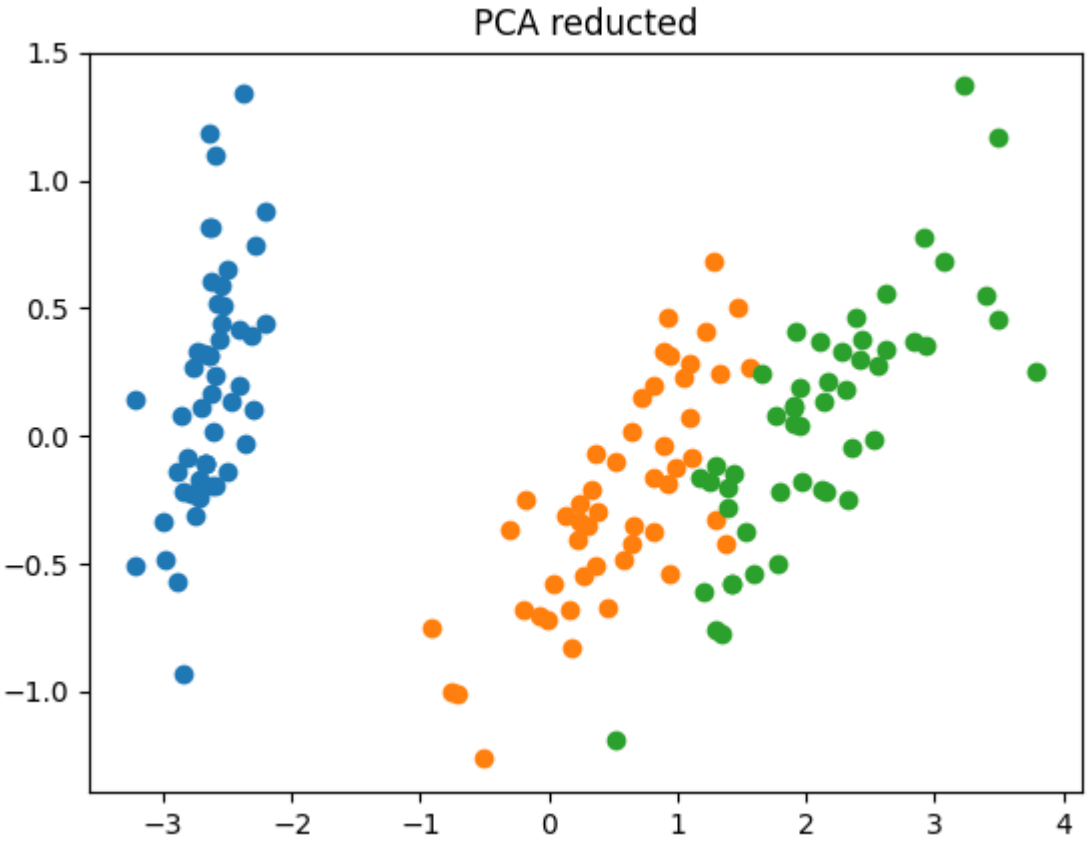

Рисунок 2 – Сокращение PCA

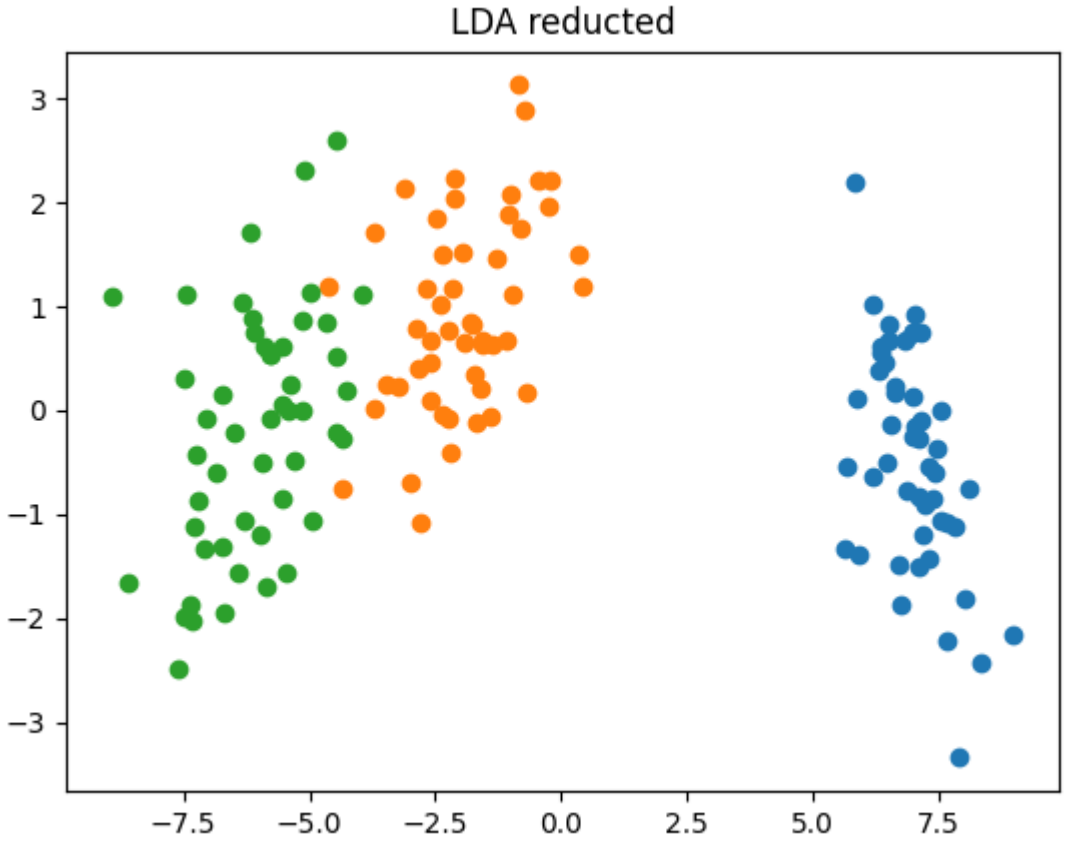

Рисунок 3 – Сокращение LDA

Работа классификатора исследована при различных параметрах solver, shrinkage. Результаты представлены на рис. 4 – 6.

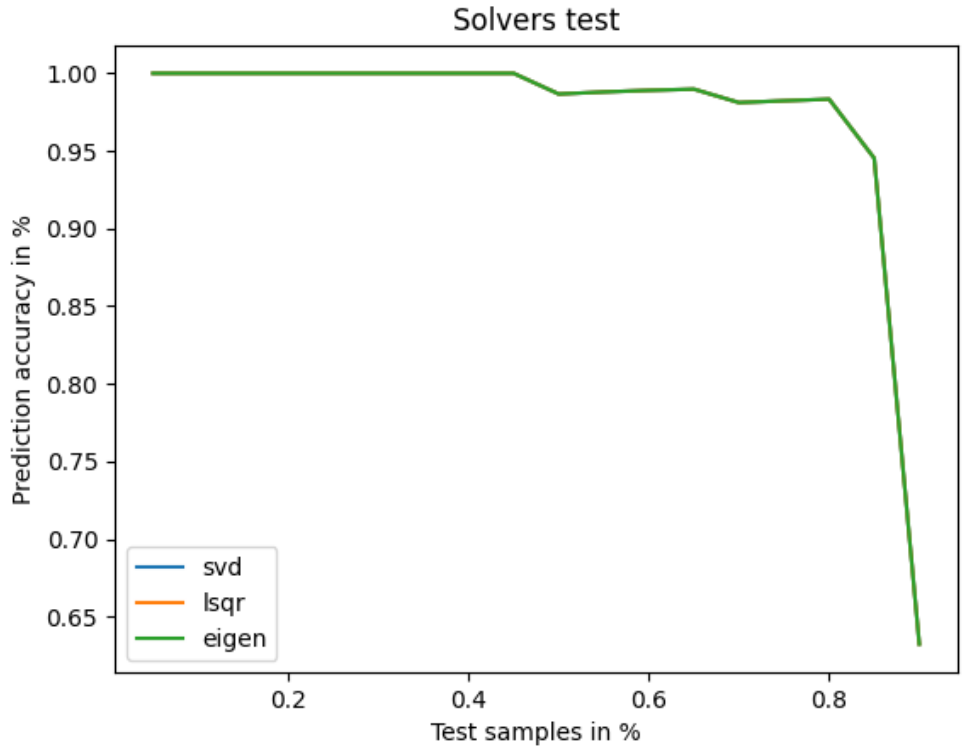

Рисунок 4 – Тест параметра Solver

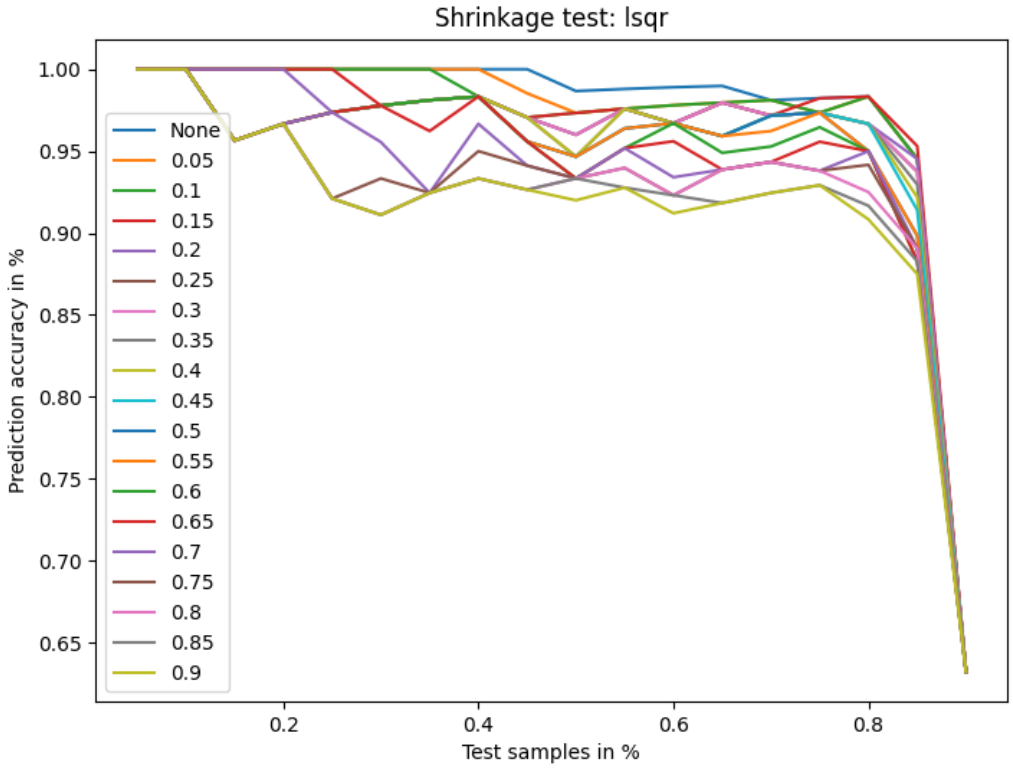

Рисунок 5 – Тест параметра Shrinkage: lsqr solver

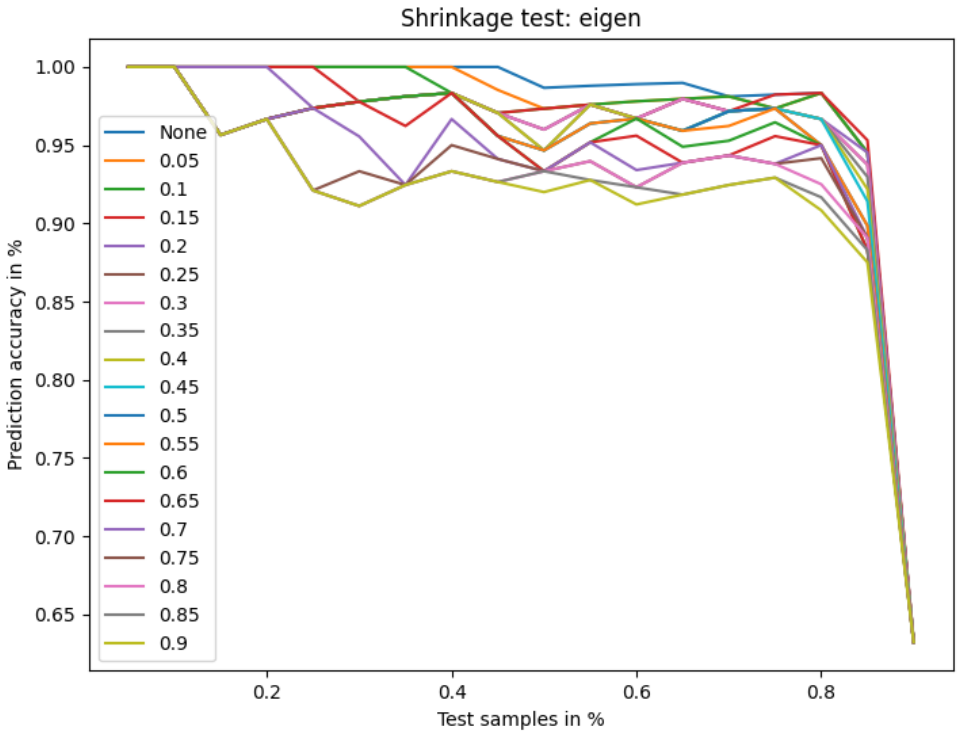

Рисунок 6 – Тест параметра Shrinkage: eigen solver

Заданы собственные значения априорных вероятностей классов, результаты представлены на рис. 7.

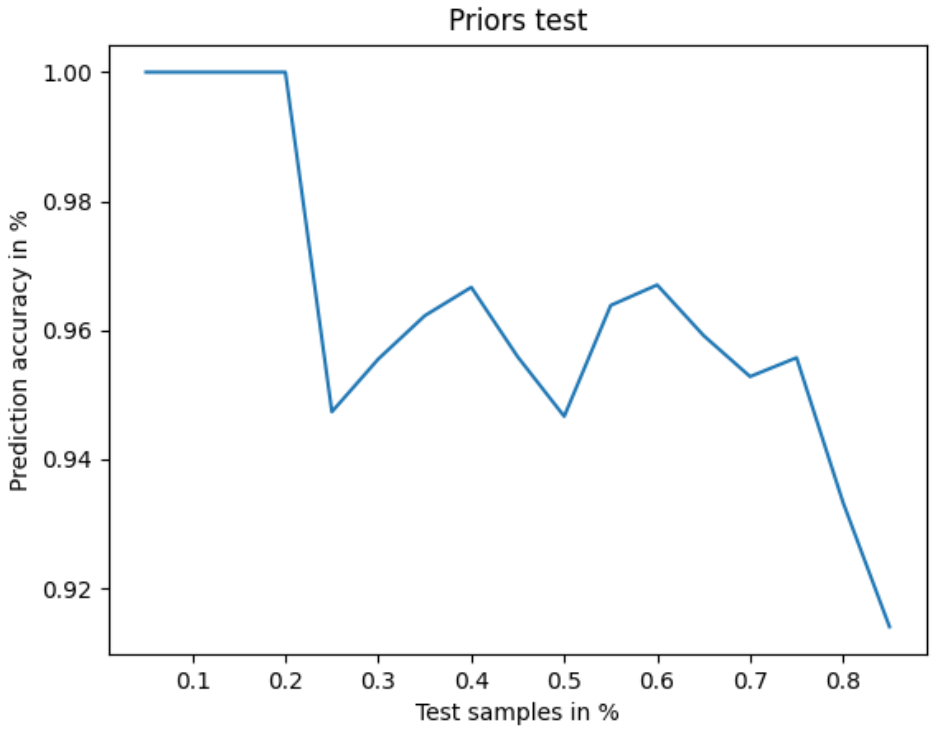

Рисунок 7 – Тест с предустановленным параметром prior = [0.15, 0.7, 0.15]

Метод опорных векторов.

Проведена классификация наблюдений с помощью метода опорных векторов на тех же данных. Выявлено 4 неправильно классифицированных наблюдения.

Точность классификации составляет 95%.

Атрибут support_ хранит индексы опорных векторов, support_vectors_ – сами опорные вектора, n_support_ – количество опорных векторов для каждого класса.

Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. График представлен на рис. 8.

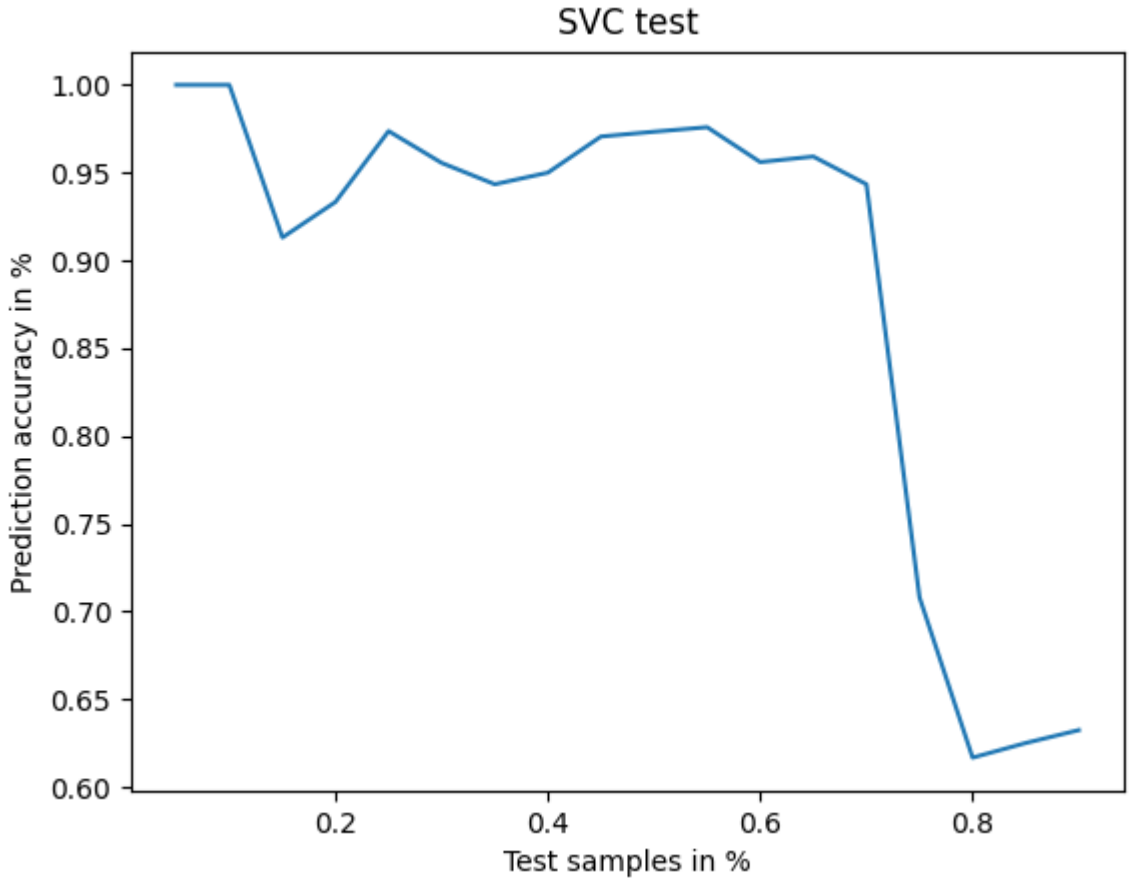

Рисунок 8 – Классификация SVC

Исследована работа метода опорных векторов при различных значениях параметров kernel, degree, max_iter. Результаты представлены на рис. 9 – 11.

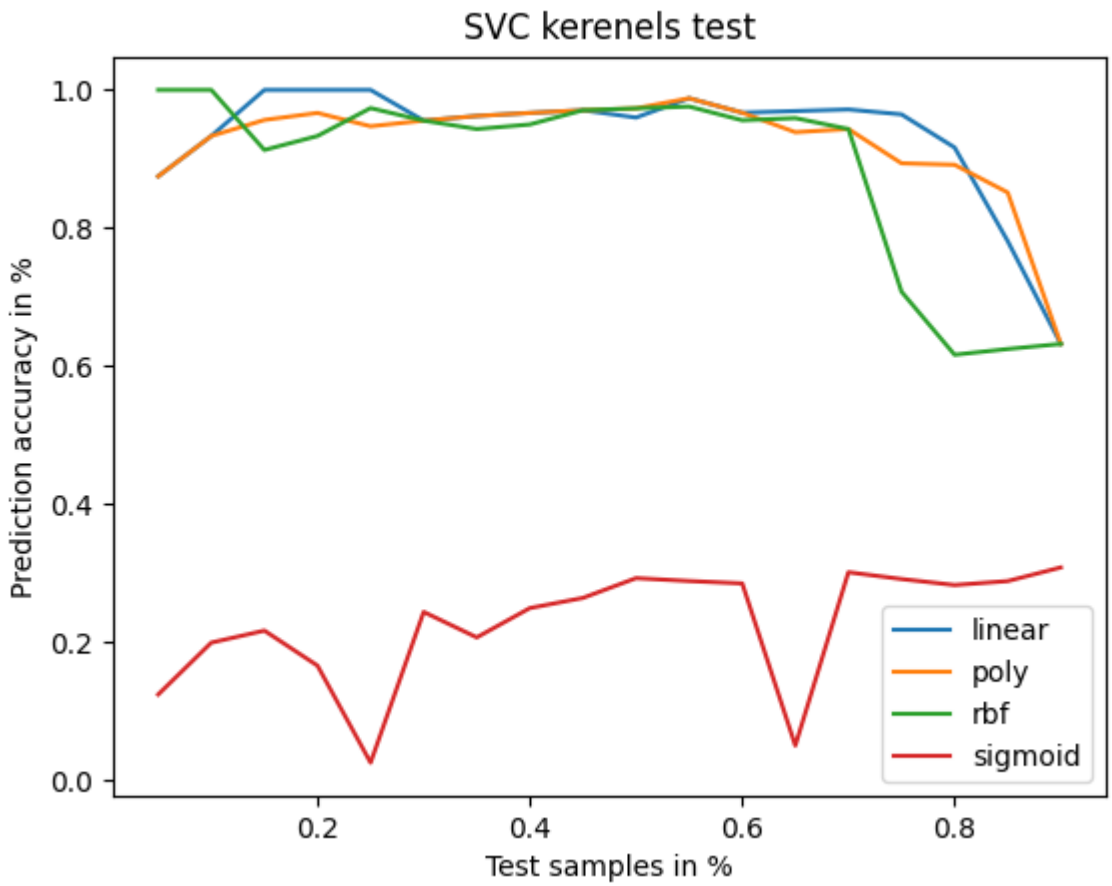

Рисунок 9 – Тест параметра kernel

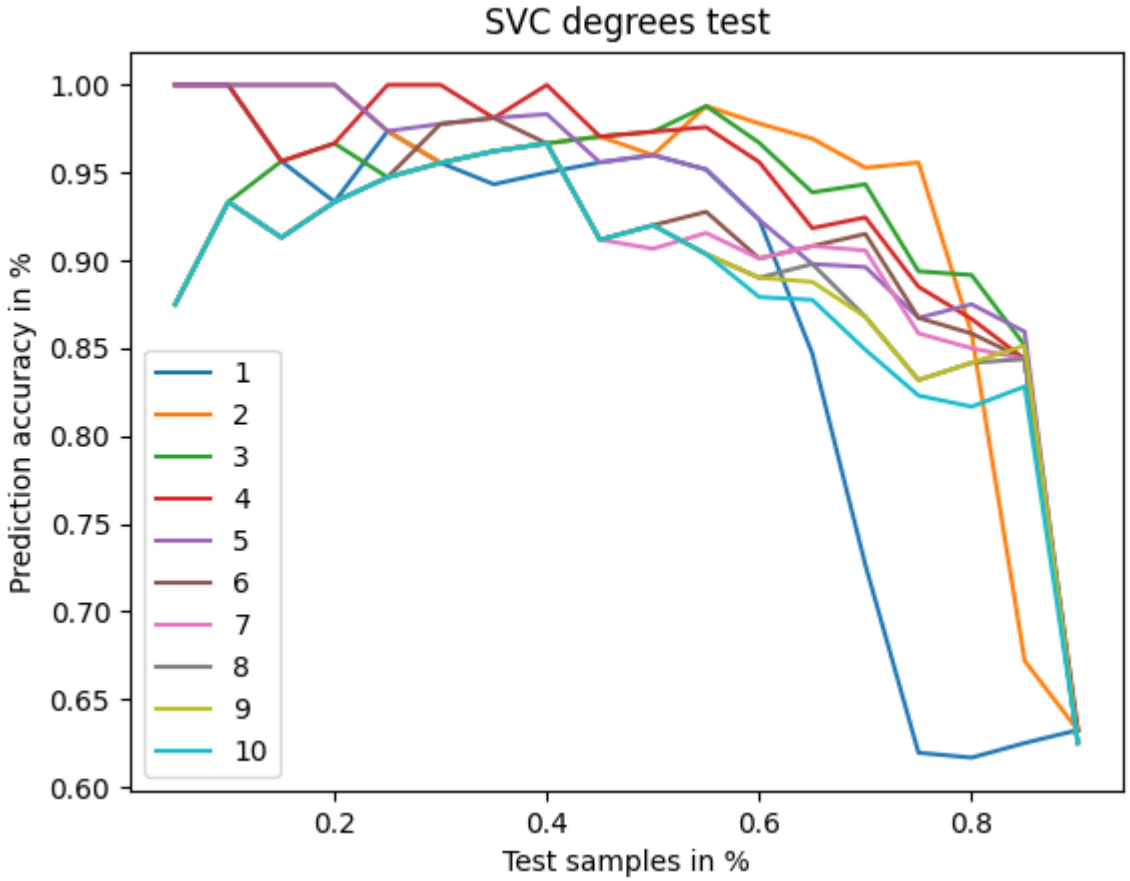

Рисунок 10 – Тест параметра degree

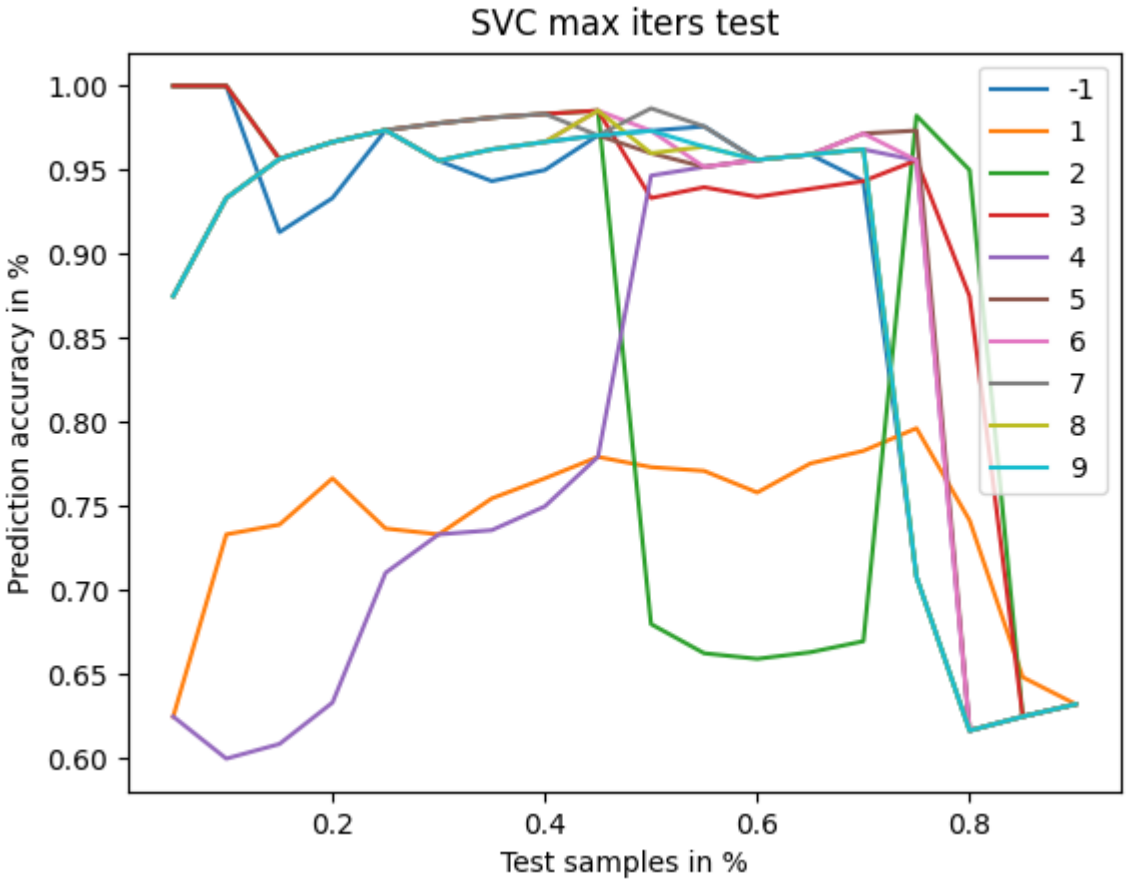

Рисунок 11 – Тест параметра max_iter

NuSVC подобен SVC, но использует параметр для управления количеством опорных векторов.

LinearSVC аналогично SVC с линейным ядром, но лучше масштабируется для большого числа выборок.

Классификация методами NuSVC и LinearSVC представлены на рис. 12 – 13.

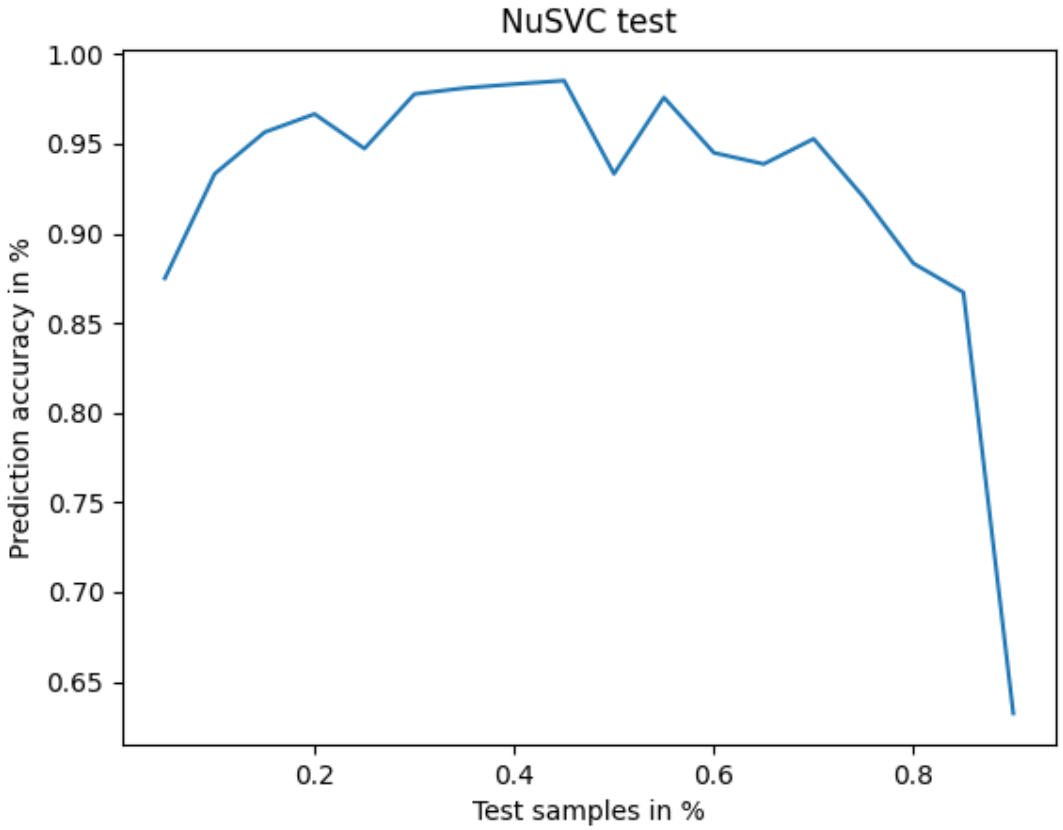

Рисунок 12 – Классификация NuSVC

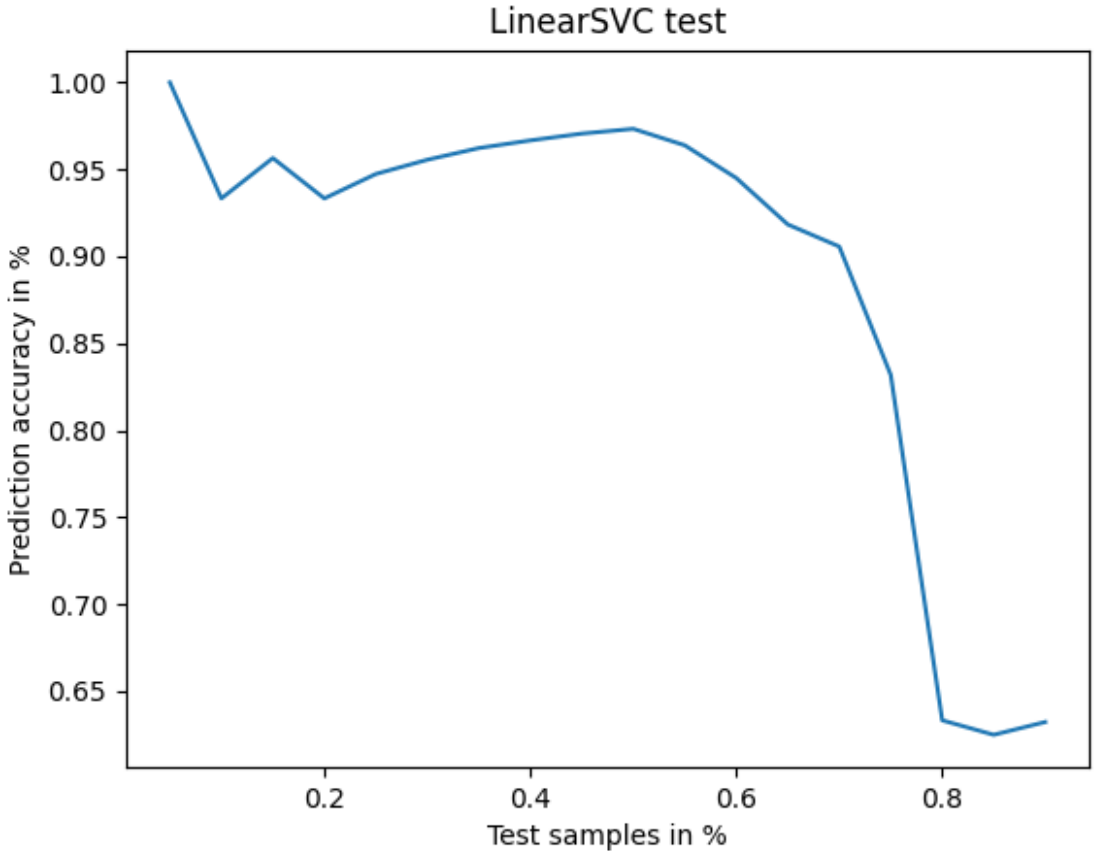

Рисунок 12 – Классификация LinearSVC

Выводы

В ходе лабораторной работы рассмотрены такие методы классификации модуля Sklearn, как LinearDiscriminantAnalysis, SVC, NuSVC и LinearSVC.