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 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2020

Цель работы:

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

Загрузка данных.

Был загружен датасет iris.data (фрагмент исходного датасета показан на рис. 1).

	0	1	2	3	4
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Рис. 1. Фрагмент исходного датасета.

Датасет был разделён на данные и метки, причём метки преобразованы к числам с помощью LabelEncoder. Далее датасет был разбит на тестовую и тренировочную выборку. Листинг 1 показывает предобработку датасета.

Листинг 1 — Подготовка датасета

```
data = pd.read_csv('data/iris.data',header=None)

X = data.iloc[:,:4].to_numpy()

labels = data.iloc[:,4].to_numpy()
le = preprocessing.LabelEncoder()
Y = le.fit_transform(labels)

X_train, X_test, y_train, y_test = train_test_split(X, Y,
test_size=0.5)
```

Линейный дискриминантный анализ

1. Была проведена классификация датасета методом линейного дискриминатного анализа. Было неверно классифицировано 2 объекта, точность составила 97.33%.

Параметры классификатора LinearDiscriminantAnalysis:

- solver — метод решения (svd, eigen, lsqr);
- shrinkage — параметр усадки;
- priors — априорные вероятности классов;
- n_components — число компонент для уменьшения размерности;
- store_covariance — флаг, задающий явное вычисление взвешенной ковариационной матрицы внутри класса при solver=svd;
- tol — абсолютный порог для того, чтобы сингулярное число X считалось значимым;
- covariance_estimator — используется для оценки ковариационных матриц вместо эмпирической оценки.

Атрибуты классификатора LinearDiscriminantAnalysis:

- coef_ — вектор(ы) веса;
- intercept_ — массив прерывания;
- covariance_ — взвешенная матрица ковариаций внутри класса;
- explained_variance_ratio_ — процент дисперсии, объясняемый каждым из выбранных компонентов;
- means_ — средние значения по классам;
- priors_ — априорные вероятности по классам;
- scalings_ — масштабирование признаков в пространстве, охватываемом центроидами классов;
- xbar_ — общее среднее;
- classes_ — уникальные метки классов.

2. На рис. 1 показан график зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки с заданным параметром random_state=630442. Качество классификации падает при

уменьшении обучающей выборки, но сохраняется даже при довольно большой тестовой выборке, что можно объяснить хорошей классифицируемостью данных.

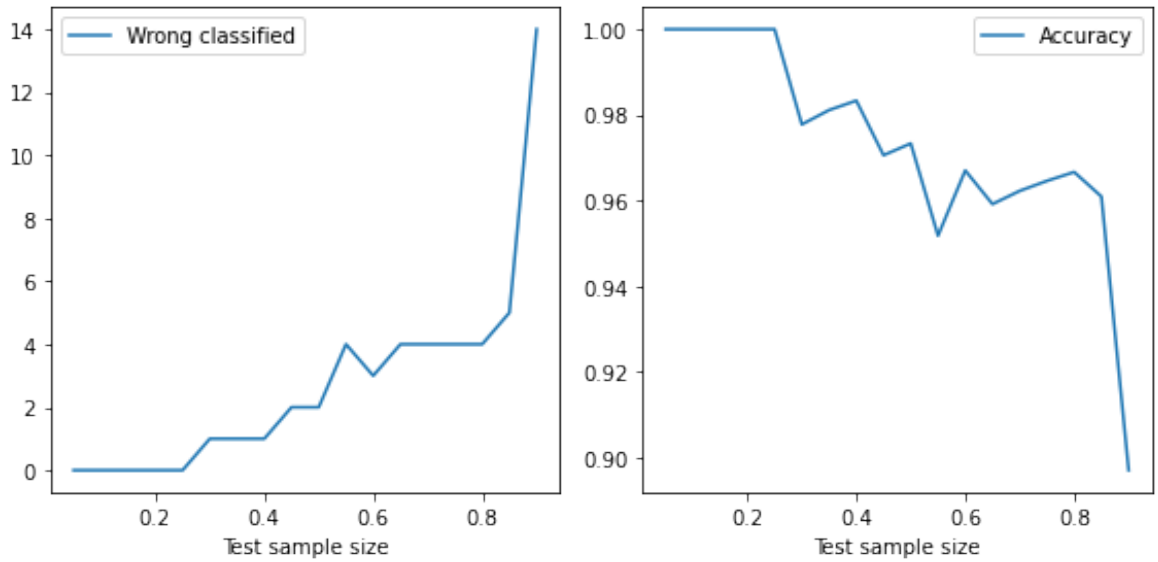

Рис. 1 — График зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки.

3. Функция transform классификатора LinearDiscriminantAnalysis выполняет понижение размерности данных. Результат применения функции показан на рис. 2 — размерность уменьшена до двух.

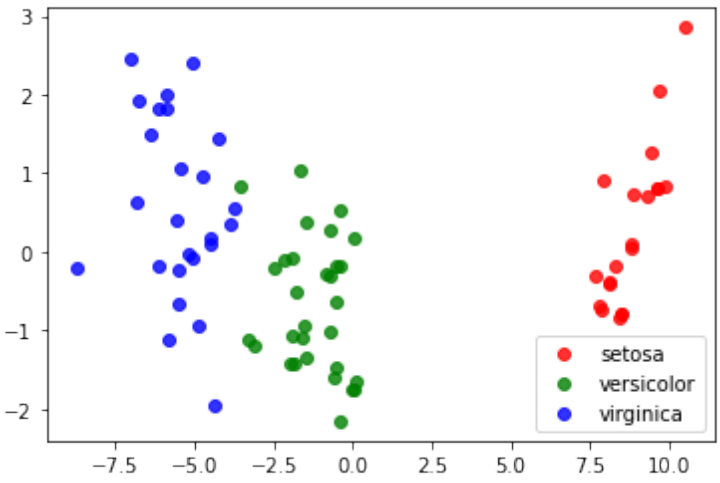

Рис. 2. Визуализация применения функции transform.

4. Для классификатора LinearDiscriminantAnalysis доступно три типа решателя, по умолчанию используется svd (рис. 1). Аналогичное пункту 2 исследование было выполнено для решателей lsqr и eigen. На рис. 3-4 показаны графики, никакой разницы с svd не обнаружено. Вероятно, размер датасета не позволяет оценить разницу в решателях.

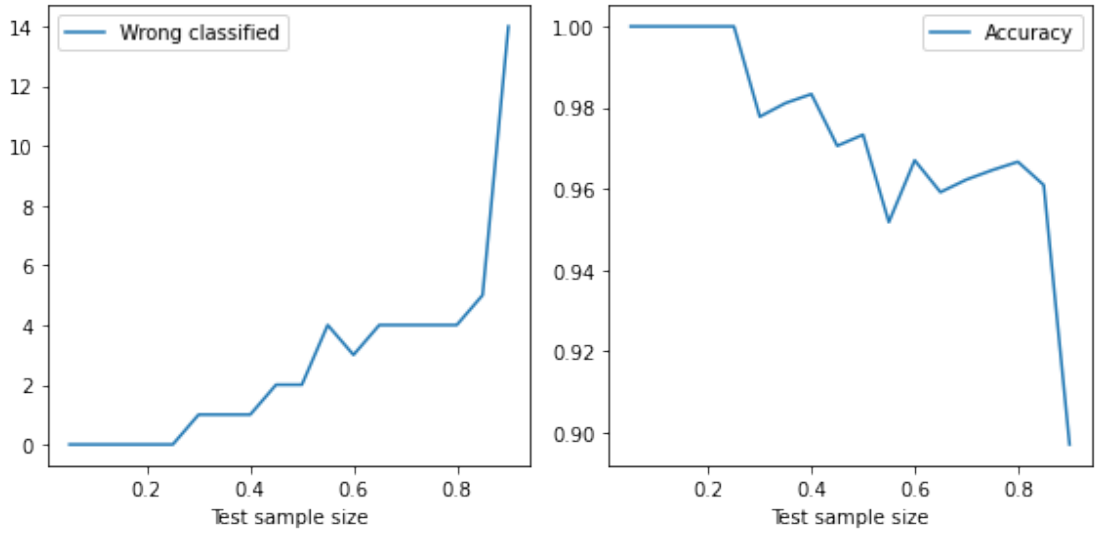

Рис. 3 — График зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки для solver=eigen.

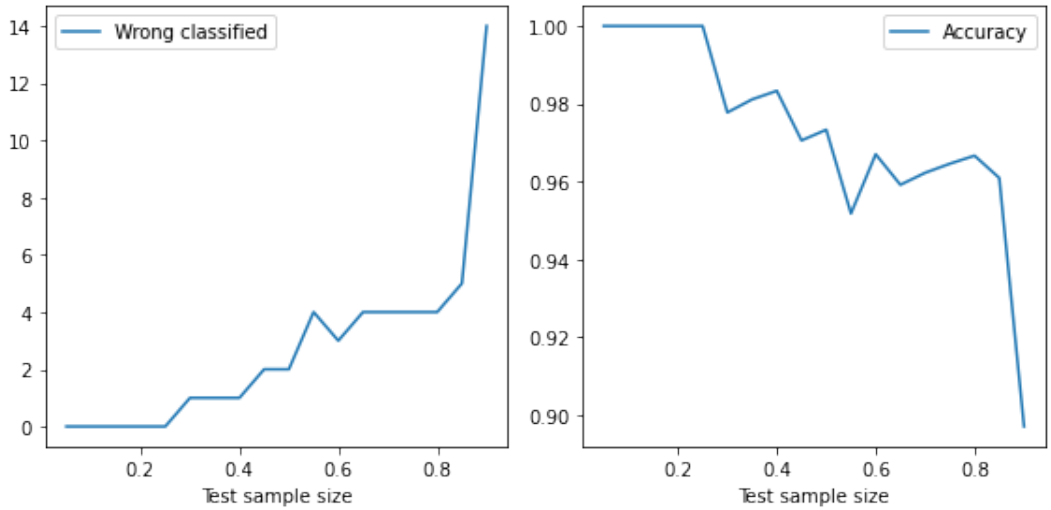

Рис. 4 — График зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки для solver=lsqr.

На рис. 5 показана зависимость точности от параметра shrinkage при solver=eigen. С увеличением значения точность незначительно снижается.

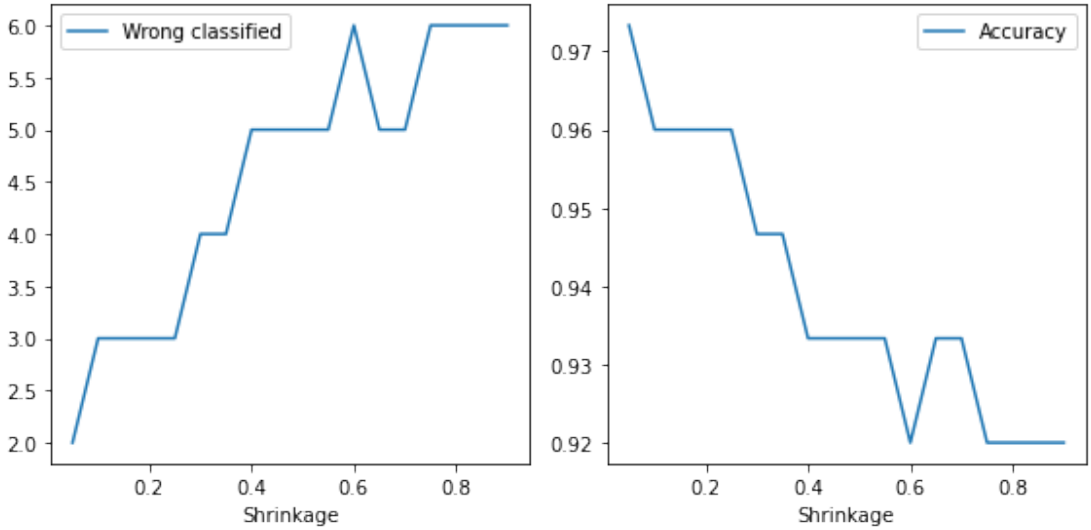

Рис. 5 — График зависимости точности классификации и числа неверно классифицированных точек от параметра shrinkage.

5. На рис. 6 показано исследование точности классификации при заданных априорных вероятностях. Принципиальных отличий от случая, когда вероятности вычисляются по умолчанию, нет.

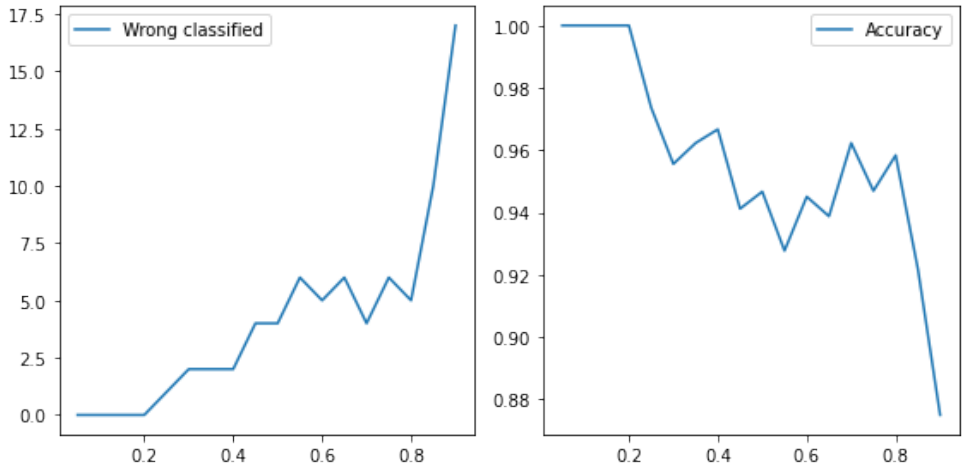

Рис. 6 — График зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки при заданных априорных вероятностях.

Метод опорных векторов

1. Тот же набор данных был классифицирован методом опорных векторов (SVM). Неверно классифицированы 4 точки, точность составила 95.33%.

2. Были получены некоторые атрибуты классификатора SVM:

- support_ — индексы опорных векторов:

```
[ 1 12 15 29  7 11 14 16 19 24 25 27 33 34 35 38 43 51 54 63 66 67
73  5  6  9 13 17 22 30 39 40 50 53 58 60 62 72]
```

- support_vectors_ — опорные векторы:

```
[[4.8 3.4 1.9 0.2]
 [4.5 2.3 1.3 0.3]
 [5.1 3.8 1.9 0.4]
 [5.1 3.3 1.7 0.5]
 [5.7 2.6 3.5 1. ]
 [6.7 3.1 4.7 1.5]
 [5.7 2.8 4.5 1.3]
 [5.  2.3 3.3 1. ]
 [5.5 2.6 4.4 1.2]
 [6.5 2.8 4.6 1.5]
 [6.1 2.9 4.7 1.4]
 [6.2 2.2 4.5 1.5]
 [6.9 3.1 4.9 1.5]
 [6.1 3.  4.6 1.4]
 [4.9 2.4 3.3 1. ]
 [6.4 3.2 4.5 1.5]
 [6.6 3.  4.4 1.4]
 [6.6 2.9 4.6 1.3]
 [6.3 3.3 4.7 1.6]
 [5.9 3.2 4.8 1.8]
 [6.1 2.8 4.  1.3]
 [6.3 2.5 4.9 1.5]
 [6.1 2.8 4.7 1.2]
 [6.2 3.4 5.4 2.3]
 [6.3 2.8 5.1 1.5]
 [6.4 2.8 5.6 2.2]
 [5.8 2.8 5.1 2.4]
 [7.2 3.  5.8 1.6]
 [4.9 2.5 4.5 1.7]
 [6.8 3.  5.5 2.1]
 [6.7 3.  5.2 2.3]
 [6.5 3.  5.5 1.8]
 [6.4 2.8 5.6 2.1]
 [6.1 3.  4.9 1.8]
 [7.7 3.8 6.7 2.2]
 [6.5 3.2 5.1 2. ]
 [5.7 2.5 5.  2. ]
 [5.8 2.7 5.1 1.9]]
```

- n_support_ — количество опорных векторов для каждого класса:

```
[ 4 19 15]
```

3. Были построены графики (рис. 7) зависимости числа неверно классифицированных точек и точности классификации от размера тестовой выборки. При уменьшении тренировочной выборки падает точность классификации, однако темпы снижения не столь высоки в силу хорошей классифицируемости данных.

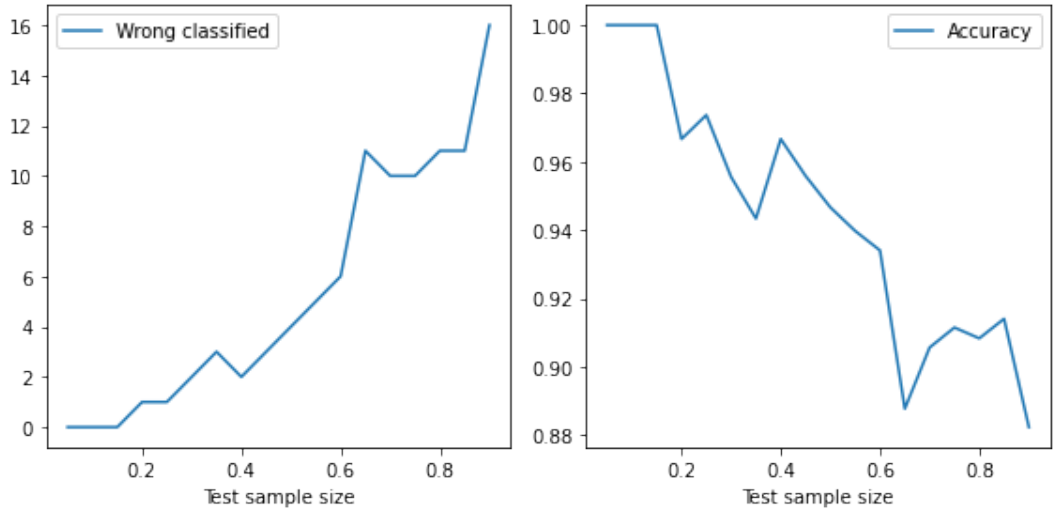

Рис. 7 — График зависимости точности классификации и числа неверно классифицированных точек от размера тестовой выборки для классификатора SVM.

4. Было проведено исследование влияния параметров классификатора SVM на точность классификации. На рис. 8 приведена зависимость точности классификации от степени ядра при использовании полиномиального ядра. Видно, что при высоких степенях точность стремится к нулю.

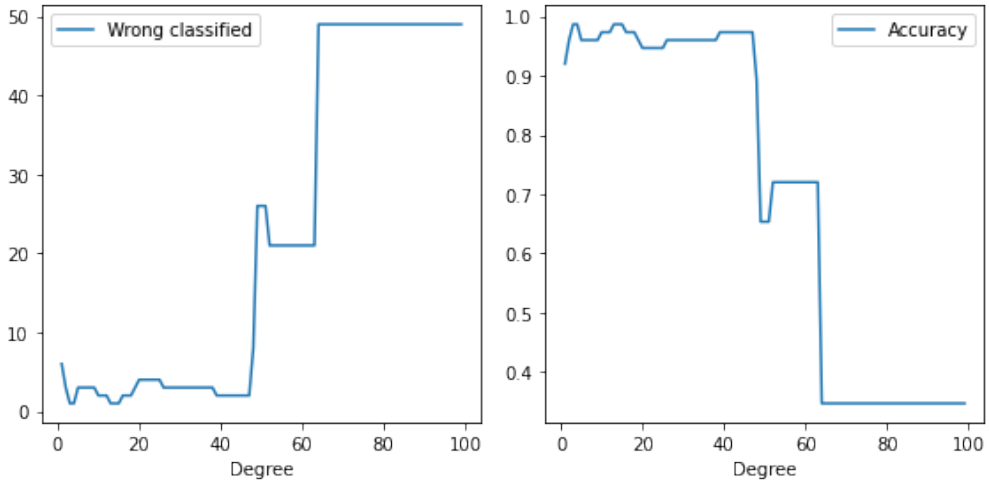

Рис. 8 — График зависимости точности классификации и числа неверно классифицированных точек от параметра degree для классификатора SVM.

На рис. 9 показано аналогичное исследование для параметра max_iter (остальные параметры — по умолчанию). После некоторого порога точность классификации не меняется.

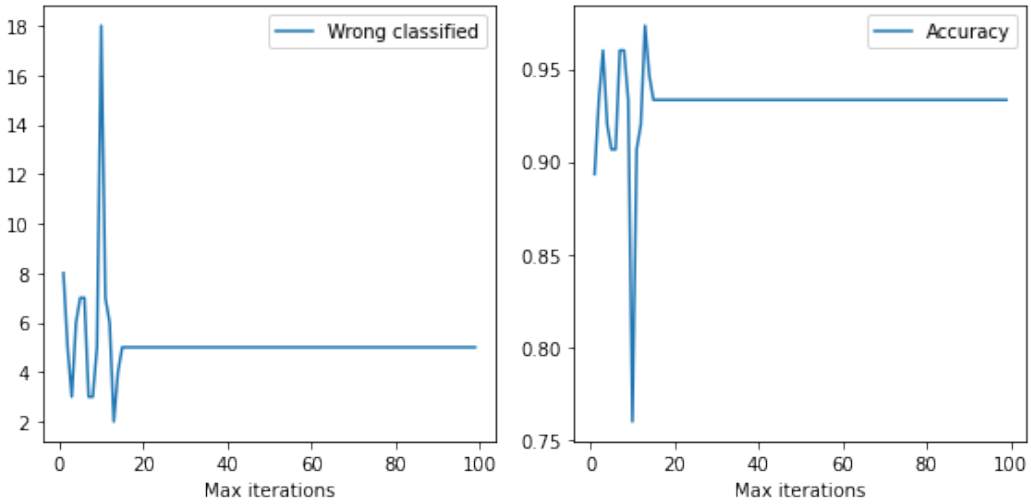

Рис. 9 — График зависимости точности классификации и числа неверно классифицированных точек от параметра max_iter для классификатора SVM.

В таблицу 1 сведены данные исследования влияния различных ядер на классификацию. Ядро sigmoid показывает значительно меньшую точность.

Таблица 1 — Точность классификации для различных ядер SVM

Ядро	Неверно классифицированных	Точность, %
linear	1	98.67
poly	2	98
rbf	2	96.67
sigmoid	52	33.33

5. Было проведено сравнение работы классификаторов SVC, LinearSVC и NuSVC, результаты классификации с параметрами по умолчанию приведены в таблице 2. LinearSVC повторяет логику SVC при линейном ядре, но добавляет ряд параметров для этого ядра. NuSVC позволяет задать коилчество опорных векторов. График точности для разных ядер в зависимости от числа опорных векторов приведён на рис. 10. Чем больше векторов, тем ниже точность, причем для ядра sigmoid точность падает быстрее всего.

Таблица 2 — Сравнение работы классификаторов SVC, LinearSVC, NuSVC

Классификатор	Неверно классифицированных	Точность, %
SVC	2	96.67
LinearSVC	4	96
NuSVC	3	97.33

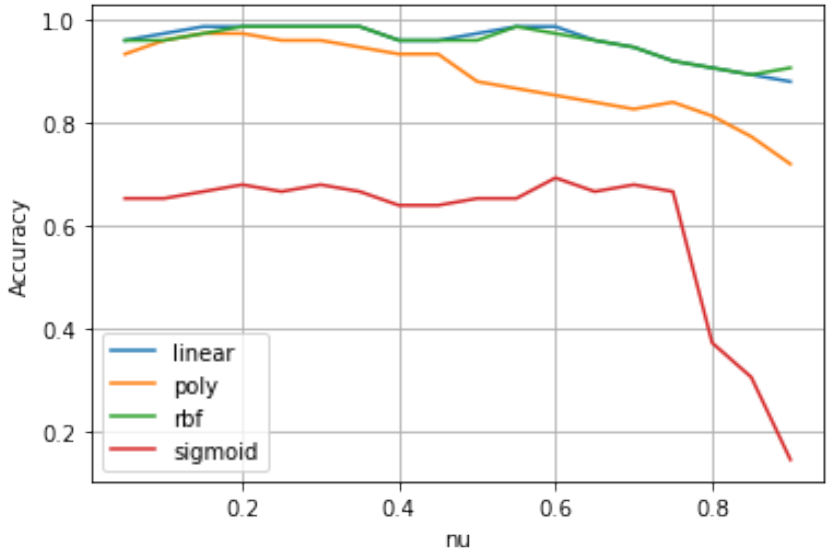

Рис. 10. Зависимость точности от числа опорных векторов для различных ядер NuSVC.

Вывод:

В результате выполнения лабораторной работы были изучены классификаторы, основанные на методах опорных векторов и линейного дискриминантного анализа.

Были изучены параметры классификаторов LinearDiscriminantAnalysis, SVC, LinearSVC, NuSVC и их влияние на точность классификации.