МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра МОЭВМ

# ОТЧЕТ

по лабораторной работе № 8

по дисциплине «Машинное обучение»

Тема: Классификация (линейный дискриминантный анализ, метод опорных векторов)

|  |  |  |
| --- | --- | --- |
| Студент гр. 8304 |  | Сергеев А.Д. |
| Преподаватель |  | Жангиров Т.Р. |

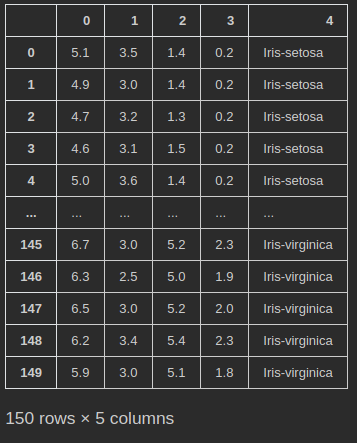
Санкт-Петербург 2021

# Цель работы

Ознакомиться с методами классификации модуля *Sklearn*.

**Загрузка данных**

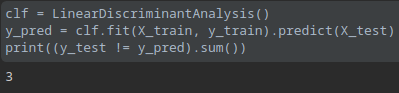
1. Данные загружены в датафрейм.

Рисунок 1 – Исходные данные

1. Выделены данные и их метки, тексты меток преобразованы к числам.
2. Выборка разбита на обучающую и тестовую *train\_test\_split*.

**Линейный дискриминантный анализ**

1. Проведена классификация наблюдений при помощи LDA.

Рисунок 2 — Классификацияпри помощи LDA

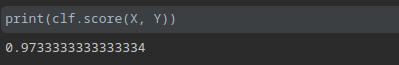
Выявлено 3 неправильно классифицированных наблюдения.

Таблица 1 – Атрибуты *LDA*

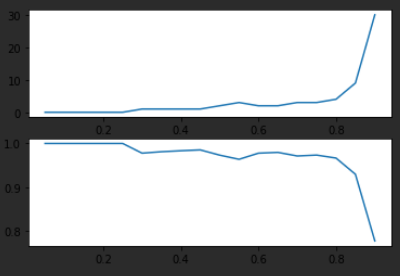
|  |  |
| --- | --- |
| **Атрибут** | **Описание** |
| coef\_ | Векторы веса |
| intercept\_ | Срок перехвата |
| explained\_variance\_ratio\_ | Взвешенная внутриклассовая матрица ковариаций |
| means\_ | Классовые средние |
| priors\_ | Приоры класса |
| scalings\_ | Масштабирование объектов в пространстве, охватываемом центрами классов |
| xbar\_ | Общее среднее |
| classes\_ | Уникальные лейблы класса |
| n\_features\_in\_ | Количество видимых деталей во время посадки |
| feature\_names\_in\_ | Названия видимых особенностей во время посадки |

|  |  |
| --- | --- |
| **Параметр** | **Описание** |
| solver | Используемый метод решения |
| shrinkage | Параметр усадки |
| priors | Класс априорных вероятностей |
| n\_components | Количество компонентов для уменьшения размерности |
| store\_covariance | Флаг для вычисления взвешенной ковариационной матрицы внутри класса. |
| tol | Абсолютный порог, чтобы единичное значение X считалось значимым, используется для оценки ранга X |
| covariance\_estimator | Используется для оценки ковариационных матриц вместо эмпирической оценки ковариации |

1. Точность классификации получена с помощью функции score() и составляет 93.3%.

Рисунок 3 — Точность классификации

1. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки (0.05 до 0.95 с шагом 0.05), random\_state = 830434.

Рисунок 4 – График зависимости точности классификации и неправильно классифицированных наблюдений от размера тестовой выборки

1. Функция transform применяется для уменьшения размерности данных.

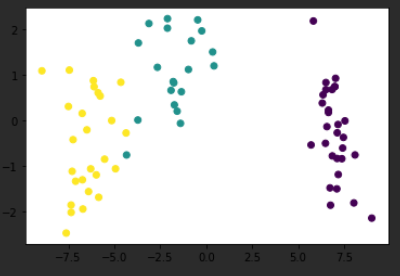
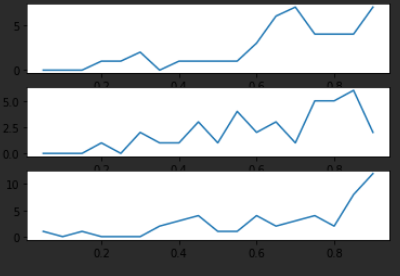


Рисунок 5 – Результат применения функции transform

1. Исследована работа классификатора при различных параметрах solver и shrinkage

Рисунок 6 — Точность классификации при различных значениях параметра solver

svd - разложение по сингулярным значениям (по умолчанию). Не вычисляет ковариационную матрицу, поэтому этот метод решения рекомендуется для данных с большим количеством функций.

lsqr - решение методом наименьших квадратов. Можно комбинировать с оценкой усадки или настраиваемой ковариационной оценкой.

eigen - разложение по собственным значениям. Можно комбинировать с оценкой усадки или настраиваемой ковариационной оценкой.

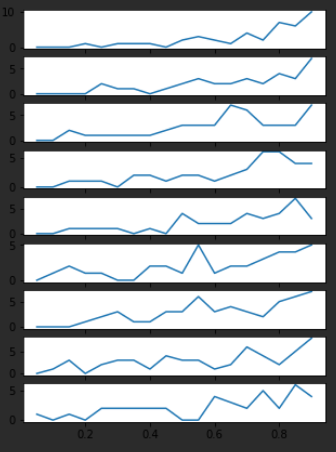


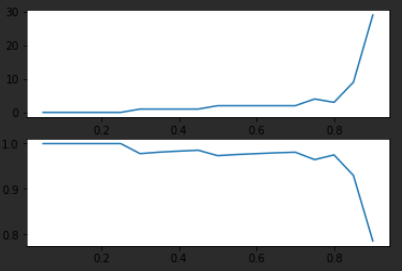
Рисунок 7 — Точность классификации при различных значениях параметра shrinkage

None – без усадки. (по умолчанию)

auto – автоматическая усадка с использованием леммы Ледуа-Вольфа.

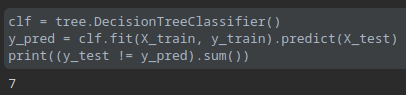
float – значение между 0 и 1. фиксированный параметр усадки.

1. Задана априорную вероятность класса с номером 1 равная 0.7, остальным классам заданы равные априорные вероятности.

Рисунок 8 – Точности классификации и количество неправильно классифицированных наблюдений при заданных априорных вероятностях

**Метод опорных векторов**

1. Проведена классификация при помощи SVM на тех же данных.

Рисунок 9 — Классификация с помощью SVM

Выявлено 7 неправильно классифицированных наблюдения.

1. Точность классификации получена с помощью функции score() и составляет 66.7%.

Рисунок 10 — Точность классификации

1. Выведена следующая информация.

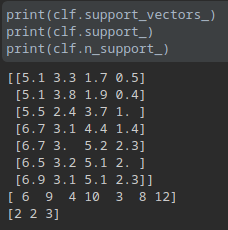


Рисунок 11 – Информация о SVM

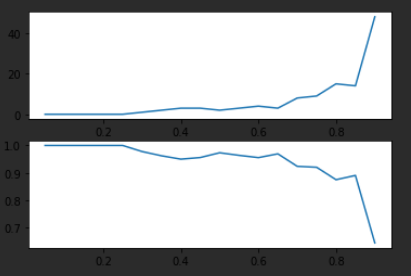
Это информация описывает опорные вектора.

support\_ содержит индексы опорных векторов.

support\_vectors\_ сожержит сами опорные вектора.

n\_support\_ сожержит количество опорных векторов для каждого класса

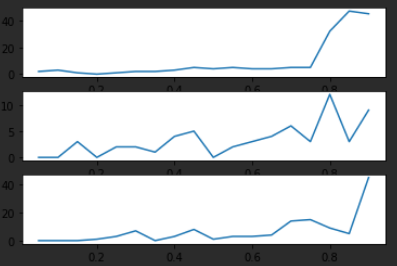
1. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. Размер тестовой выборки изменяется от 0.05 до 0.95 с шагом 0.05. Параметр random\_state выбран равным номеру зачетной книжки (830434).

Рисунок 12 – График зависимости точности классификации и неправильно классифицированных наблюдений от размера тестовой выборки

1. Исследована работа метода опорных векторов при различных значениях kernel, degree, max\_iter.

|  |  |
| --- | --- |
| Параметр | Описание |
| kernel | Тип ядра, который будет использован алгоритмом.  Поддерживается 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'.  Видно, что хуже всего себя показывает ядро sigmoid. |
| degree | Cтепень полиномиальной функции ядра. Только при kernel = 'poly'.  Видно, что лучше всего работает ядро в 2, 3, 4 и 5 степени. |
| max\_iter | Ограничение на количество итераций. При -1 неограниченно.  Видно, что при одной итерации алгоритм показывает наихудшие результаты работы. |

1. Проведено исследование для методов NuSVC и LinearSVC.

Рисунок 13 — График сравнения алгоритмов SVC, NuSVC и LinearSVC.

NuSVC имеет параметр для управления количеством опорных векторов.

LinearSVC аналогичен SVC при kernel = „linear“, но лучше масштабируется.

# Выводы

В ходе выполнения лабораторной работы было произведено знакомство с классификацией методами GaussianNB, MultinominalNB, ComplementNB, BernoulliNB и DecisionTreeClassifier модуля Sklearn.