

**Московский авиационный институт  
(национальный исследовательский университет)**

Институт №8 «Информационные технологии и прикладная математика»  
Кафедра 806 «Вычислительная математика и программирование»  
Дисциплина «Искусственный Интеллект»

**Лабораторная работа №1**

Тема: Линейные модели

Студент: Глушатов И.С.  
Группа: М8О-307Б-19  
Преподаватель: Ахмед Самир. Х.  
Дата:  
Оценка:

Москва, 2022

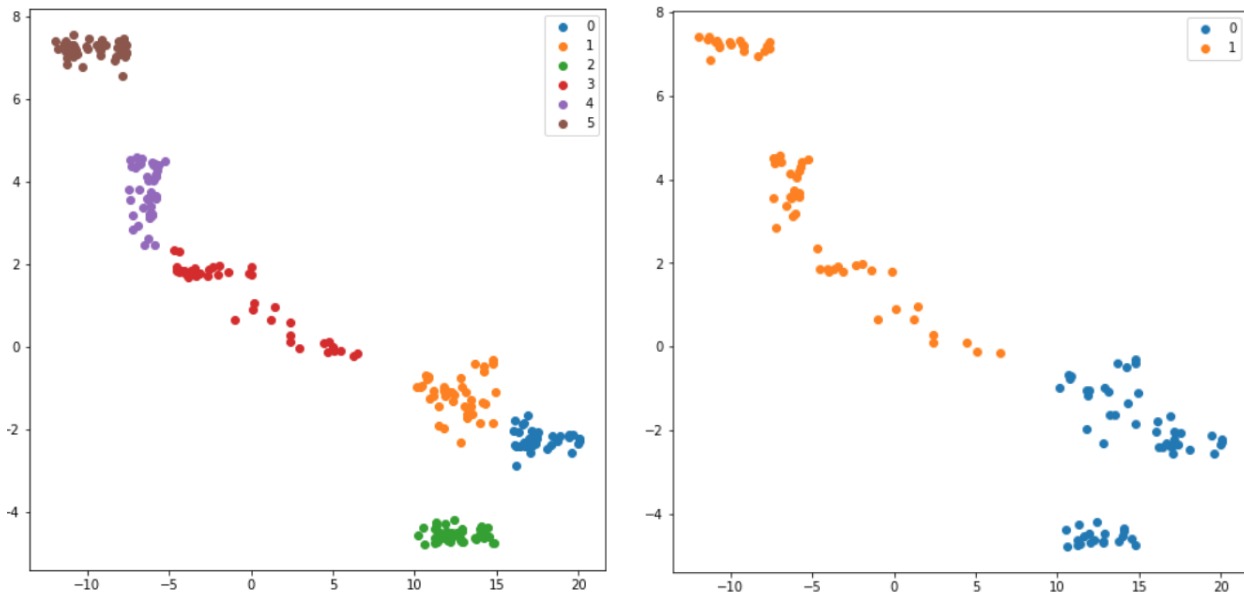
**Цель работы:** научиться реализовывать различные линейные модели классификаторов (KNN, SVN, Logistic Regression и Naive Bayes), работать в pipeline, кросс-валидацией и проанализировать результаты с помощью различных метрик.

**Задание:**

Вы собрали данные и их проанализировали, визуализировали и представили отчет своим партнерам и спонсорам. Они согласились, что ваша задача имеет перспективу и продемонстрировали заинтересованность в вашем проекте. Самое время реализовать прототип! Вы считаете, что нейронные сети переоценены (просто боитесь признаться, что у вас не хватает ресурсов и данных), и считаете, что за машинным обучением классическим будущее и потому собираетесь использовать классические модели. Вашим первым предположением является предположение, что данные и все в этом мире имеет линейную зависимость, ведь не зря же в конце каждой нейронной сети есть линейный слой классификации. В качестве первых моделей вы выбрали, линейную / логистическую регрессию и SVM. Так как вы очень осторожны и боитесь ошибиться, вы хотите реализовать случай, когда все-таки мы не делаем никаких предположений о данных, и взяли за основу идею "близкие объекты дают близкий ответ" и идею, что теорема Байеса имеет ранг королевской теоремы. Так как вы не доверяете другим людям, вы хотите реализовать алгоритмы сами с нуля без использования scikit-learn (почти). Вы хотите узнать, насколько хорошо ваши модели работают на выбранных вам данных и хотите замерить метрики качества. Ведь вам нужно еще отчитаться спонсорам!

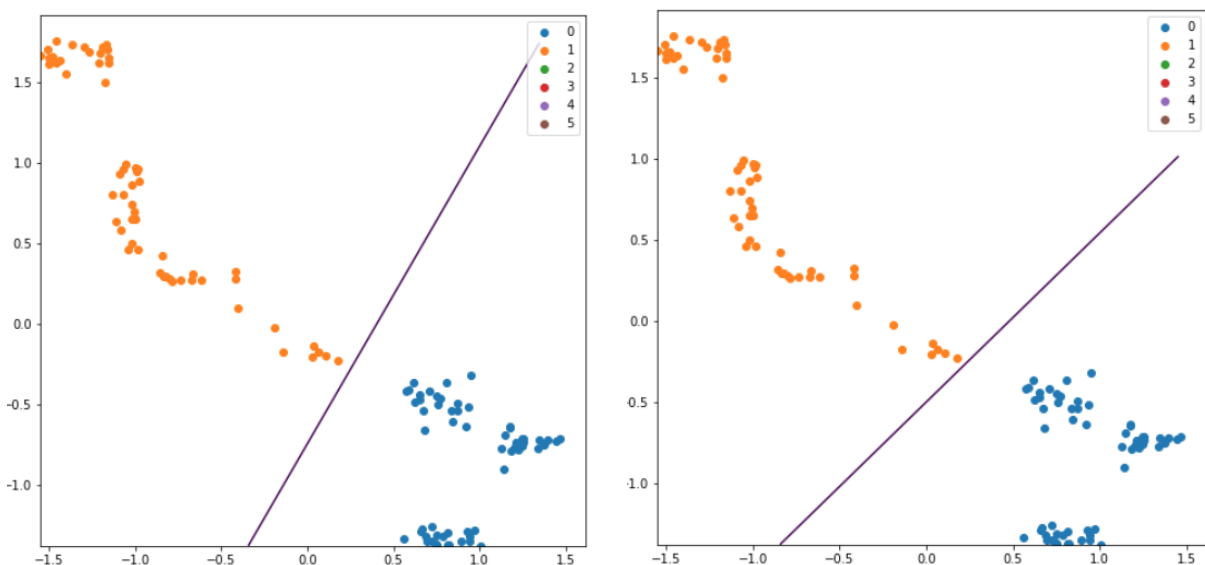
## Ход работы

Проанализировав в прошлой лабораторной работе данные характеристик звезд, я нашел два параметра, позволяющих разделить классы друг от друга. Из четырех характеристик я оставил лишь два: радиус и звездную величину, причем вместо радиуса, я решил рассматривать логарифм радиуса. Получилась такая картина:

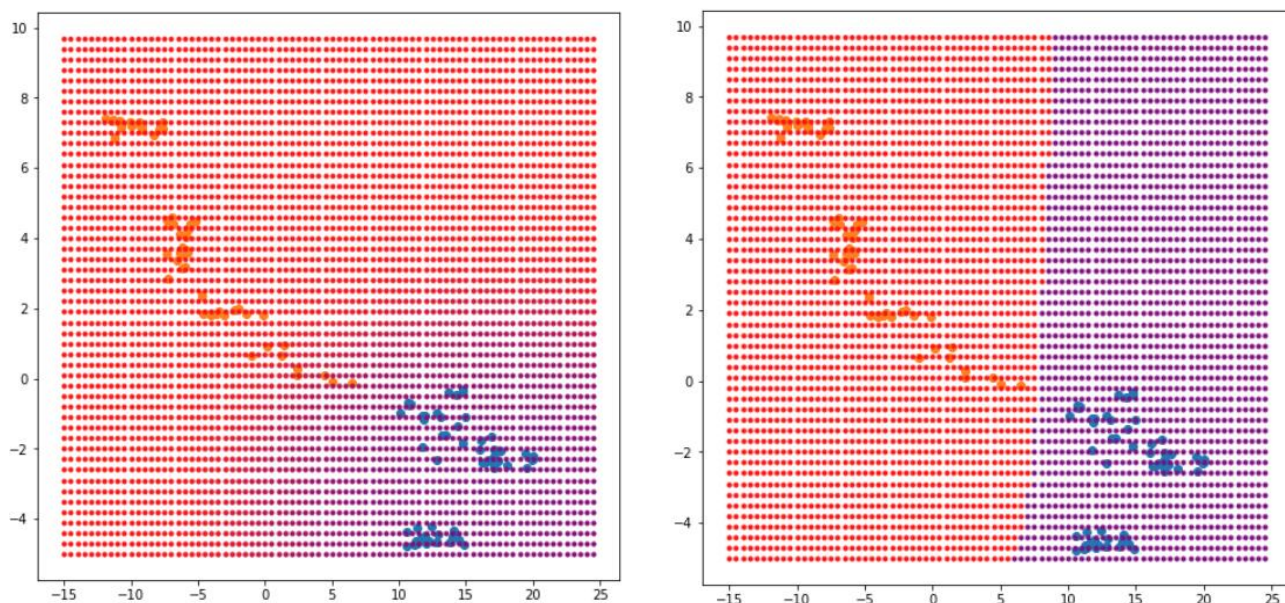


Для решения задачи бинарной классификации я решил в качестве первой группы выбрать классы звезд от 0 до 2, второго – от 3 до 5. Так как в моём случае имеет место быть линейная разделимость, то все модели собственной реализации справились с задачей на отлично. Ассигасу для всех методов был равен единице.

Демонстрация разделяющих плоскостей после обучения:



Слева представлен метод логистической регрессии, а справа – метод опорных векторов. Алгоритмы просто повторяют однослойную модель перцептрона с разными loss-функциями.



Слева демонстрация Гауссовского наивного Байесовского классификатора, справа – метод k-ближайших соседей.

Видно, что граница первого более размыта, у второго – четкая. Гауссовский NBC хорошо подходит для данных, которые в общем напоминают облако, кучку. В случае других форм, типа овалов, кругов или парабол и т. д. алгоритм будет работать явно хуже.

Все бинарные классификаторы работают хорошо и в случае с разделимыми данными сравнить алгоритмы друг с другом не имеет смысла.

## Выводы

В ходе лабораторной работы я реализовал четыре линейных моделей бинарных классификаторов в виде отдельных классов с интерфейсом, приближенным к интерфейсу библиотеки `sklearn`. Они отлично себя показали на моих данных. Больше всего мне понравился гауссовский наивный байесовский классификатор, а именно его концепция, а также его легкость к обобщению на много классовую классификацию. KNN тоже был интересен, и он также легко обобщается на мультиклассы, особенно интересен частный случай ( $k = 1$ ), при котором реализуется тесселяция Воронова. [GitHub работы](#).