**Отчет по лабораторной работе №1**

*по дисциплине «Машинное обучение»*

Выполнила  
студент гр. 3530904/70105 Л.А. Каргалов

Преподаватель И.А. Селин

Оглавление

[Задание №1 3](#_Toc38468546)

[Задание №2 4](#_Toc38468547)

[Решение 4](#_Toc38468548)

[Задани№3 6](#_Toc38468549)

[Решение А 6](#_Toc38468550)

[Решение В 7](#_Toc38468551)

[Решение С 8](#_Toc38468552)

[Задание №4 8](#_Toc38468553)

[Решение А 9](#_Toc38468554)

[Решение В 9](#_Toc38468555)

[Решение С 10](#_Toc38468556)

[Задание №5 14](#_Toc38468557)

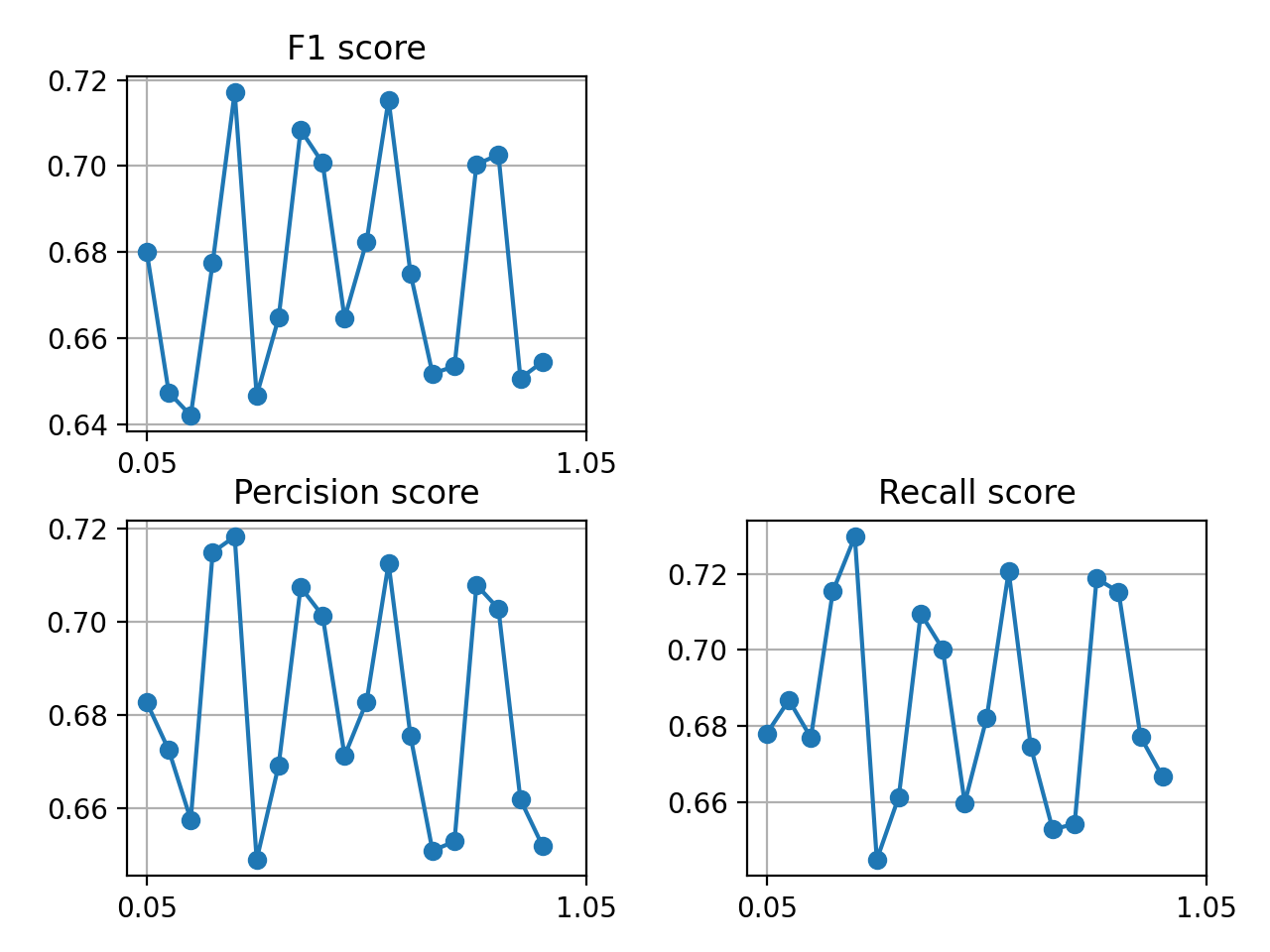
[Решение А 15](#_Toc38468558)

[Решение В 16](#_Toc38468559)

[Задание №6 16](#_Toc38468560)

[Решение 17](#_Toc38468561)

Задание №1   
Исследуйте, как объем обучающей выборки и количество тестовых данных, влияет на точность классификации в датасетах про крестики-нолики (tic\_tac\_toe.txt) и о спаме e-mail сообщений (spam.csv) с помощью наивного Байесовского классификатора. Постройте графики зависимостей точности на обучающей и тестовой выборках в зависимости от их соотношения.   
  
Решение  
Сделаем перебор соотношений обучающей выборки к тестовой от 0.05 до 1 с шагом 0.05



1 Tic tac toe

Изображение выглядит как текст, карта

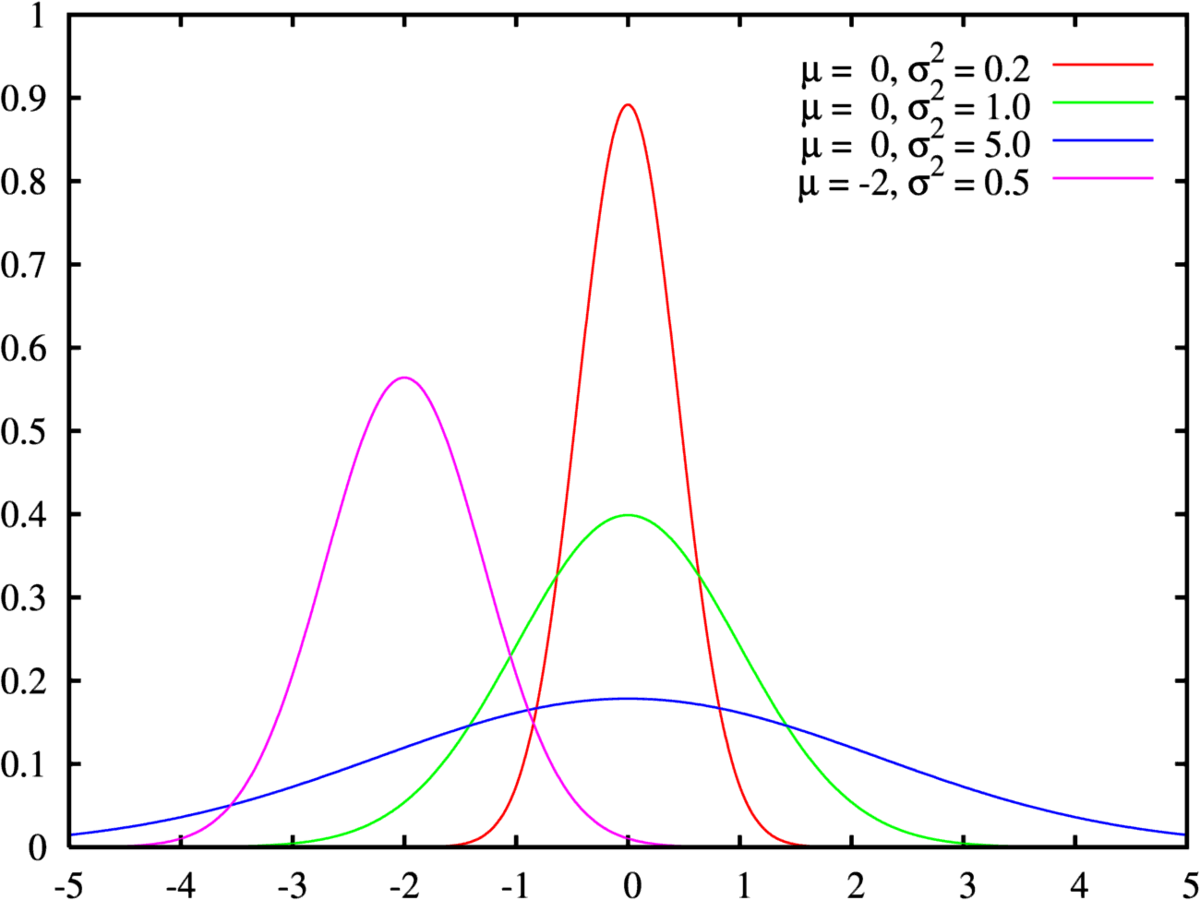
Автоматически созданное описание

2 Spam mails

На графиках видно отношение различных метрик качества к обучающей выборке.

Задание №2  
Сгенерируйте 100 точек с двумя признаками X1 и X2 в соответствии с нормальным распределением так, что одна и вторая часть точек (класс -1 и класс 1) имеют параметры: мат. ожидание X1, мат. ожидание X2, среднеквадратические отклонения для обеих переменных, соответствующие вашему варианту (указан в таблице). Построить диаграммы, иллюстрирующие данные. Построить Байесовский классификатор и оценить качество классификации с помощью различных методов (точность, матрица ошибок, ROС и PR-кривые). Является ли построенный классификатор «хорошим»?

Решение  
Для класса X1 и Х2 возьмем следующие параметры мат ожидания -0.2, 0 и среднеквадратичного отклонения 0.5, 0.2 соответственно.



Получили следующие результаты:

Изображение выглядит как текст, карта

Автоматически созданное описание

Изображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как карта, текст, снимок экрана

Автоматически созданное описание

Поскольку два эти распределения достаточно схожи имеем ошибки в классификации, если распределения будут сильно различаться, то можно добиться нулевой ошибки классификации.

# Задани№3

Постройте классификатор на основе метода k ближайших соседей для обучающего множества Glass (glass.csv). Посмотрите заголовки признаков и классов. Перед построением классификатора необходимо также удалить первый признак Id number, который не несет никакой информационной нагрузки.

* 1. Постройте графики зависимости ошибки классификации от количества ближайших соседей.
  2. Определите подходящие метрики расстояния и исследуйте, как тип метрики расстояния влияет на точность классификации.
  3. Определите, к какому типу стекла относится экземпляр с характеристиками: RI =1.516 Na =11.7 Mg =1.01 Al =1.19 Si =72.59 K=0.43 Ca =11.44 Ba =0.02 Fe =0.1

Решение А  
Построим график соотношения разных метрик качества к значению параметра k. Параметры k будем варьировать от 1 до 20 с шагом 1.

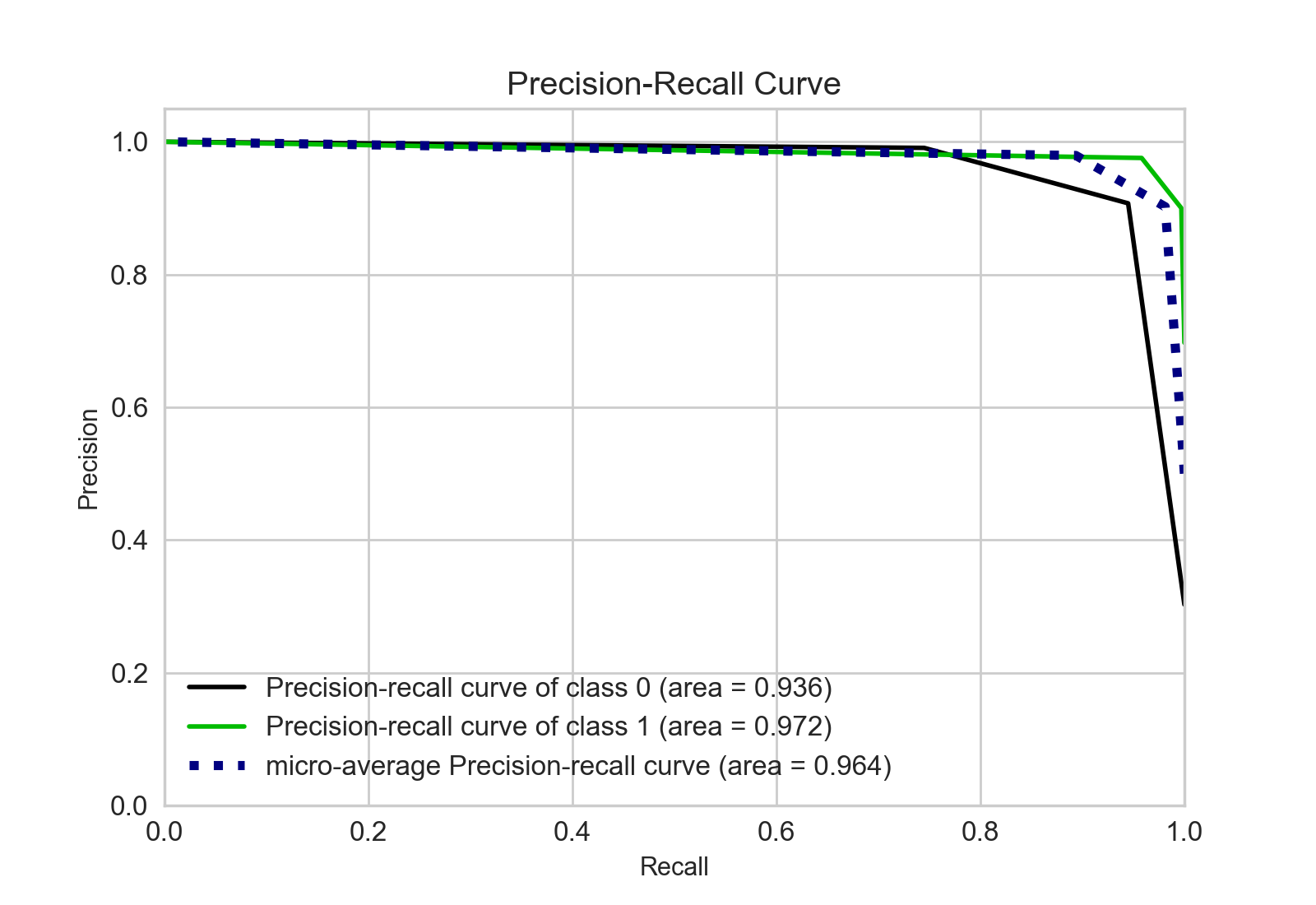
Изображение выглядит как текст, карта

Автоматически созданное описание

## Решение В

Из предыдущих графиков видно что лучше всего для конкретно этой выборки подходит параметр k = 2. Изображение выглядит как карта, текст

Автоматически созданное описание



## Решение С

При k = 2, обученных классификатор относит стекло с данным составом  
RI =1.516 Na =11.7 Mg =1.01 Al =1.19 Si =72.59 K=0.43 Ca =11.44 Ba =0.02 Fe =0.1

к типу со значением 0.

Задание №4  
Постройте классификаторы на основе метода опорных векторов для наборов данных из файлов svmdataN.txt и svmdataNtest.txt, где N – индекс задания:

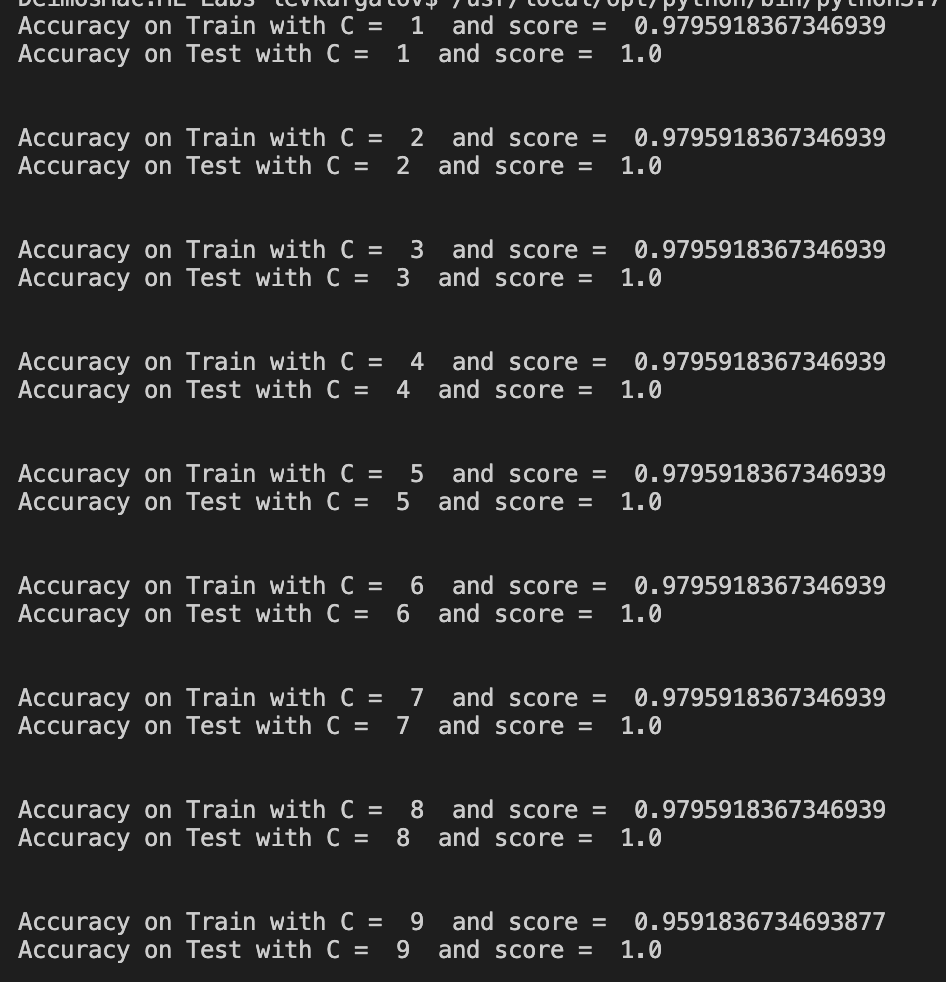
* 1. Постройте алгоритм метода опорных векторов с линейным ядром. Визуализируйте разбиение пространства признаков на области с помощью полученной модели (пример визуализации). Выведите количество полученных опорных векторов, а также матрицу ошибок классификации на обучающей и тестовой выборках.
  2. Постройте алгоритм метода опорных векторов с линейным ядром. Добейтесь нулевой ошибки сначала на обучающей выборке, а затем на тестовой, путем изменения штрафного параметра. Выберите оптимальное значение данного параметра и объясните свой выбор. Всегда ли нужно добиваться минимизации ошибки на обучающей выборке?
  3. Постройте алгоритм метода опорных векторов, используя различные ядра (линейное, полиномиальное степеней 1-5, сигмоидальная функция, гауссово). Визуализируйте разбиение пространства признаков на области с помощью полученных моделей. Сделайте выводы.
  4. Постройте алгоритм метода опорных векторов, используя различные ядра (полиномиальное степеней 1-5, сигмоидальная функция, гауссово). Визуализируйте разбиение пространства признаков на области с помощью полученных моделей. Сделайте выводы.

## Решение А

Изображение выглядит как снимок экрана

Автоматически созданное описание

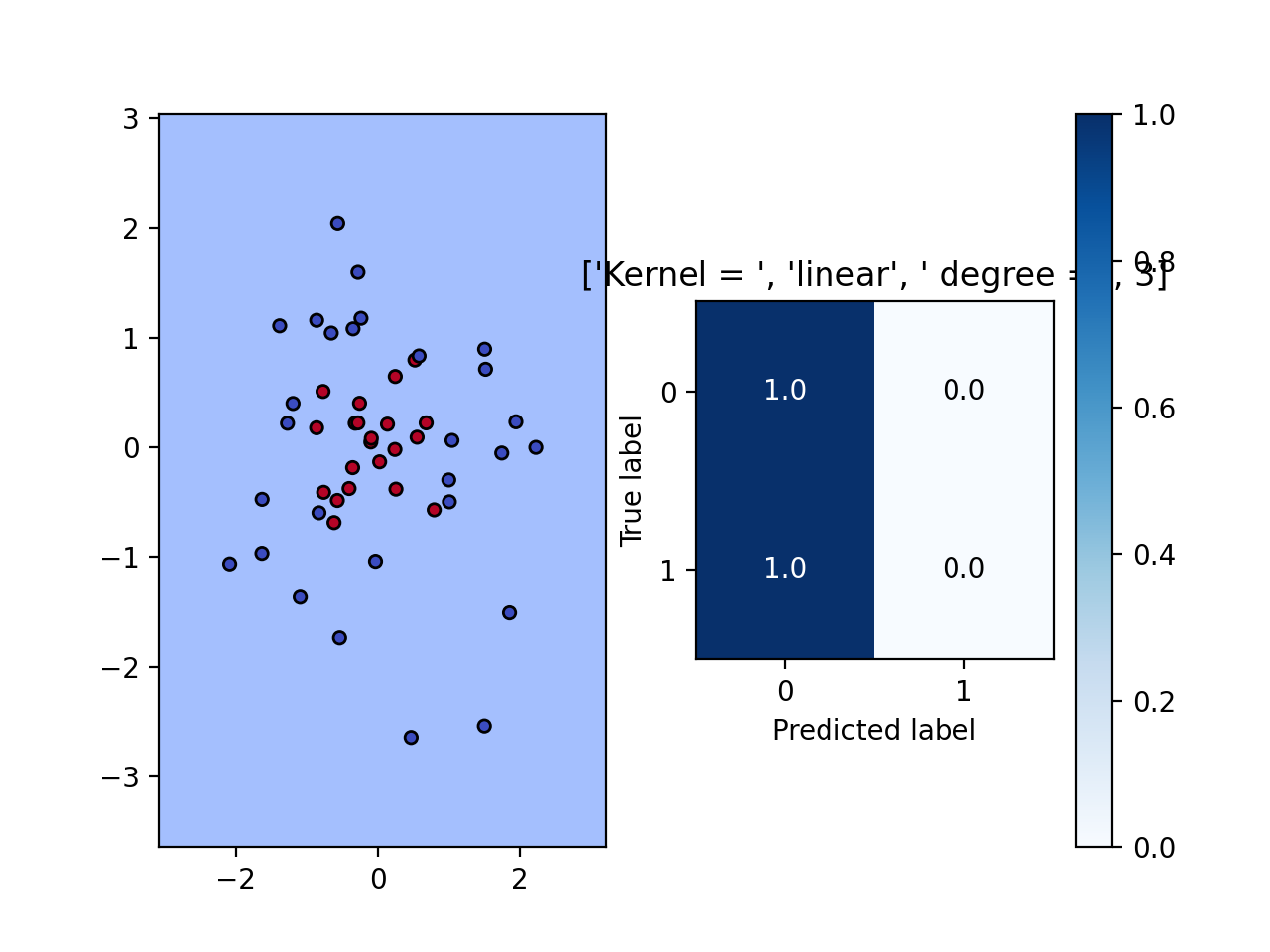
## Решение В

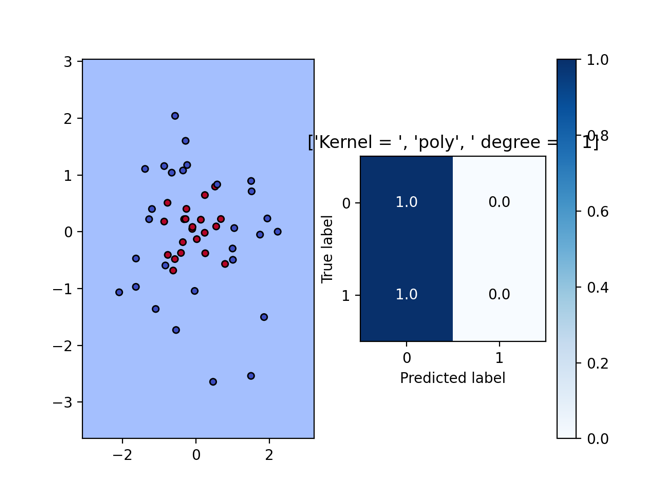


Изображение выглядит как снимок экрана

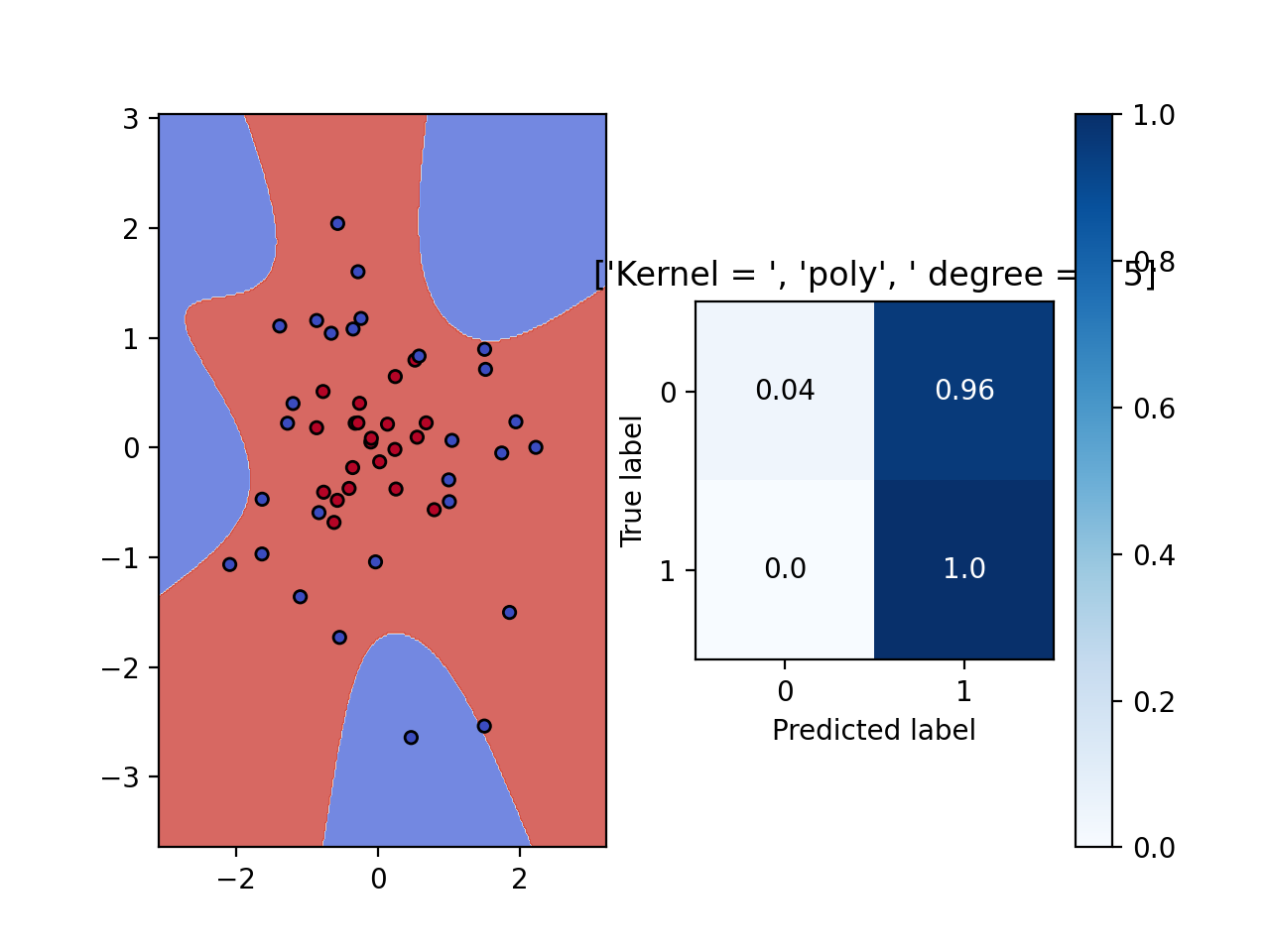
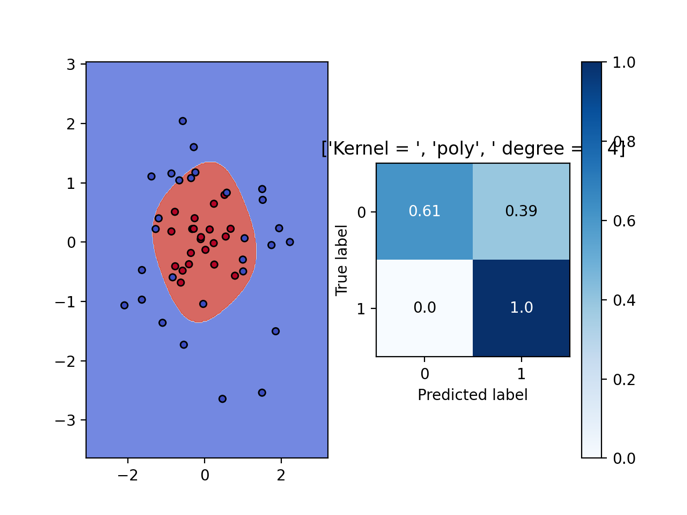
Автоматически созданное описание

## Решение С

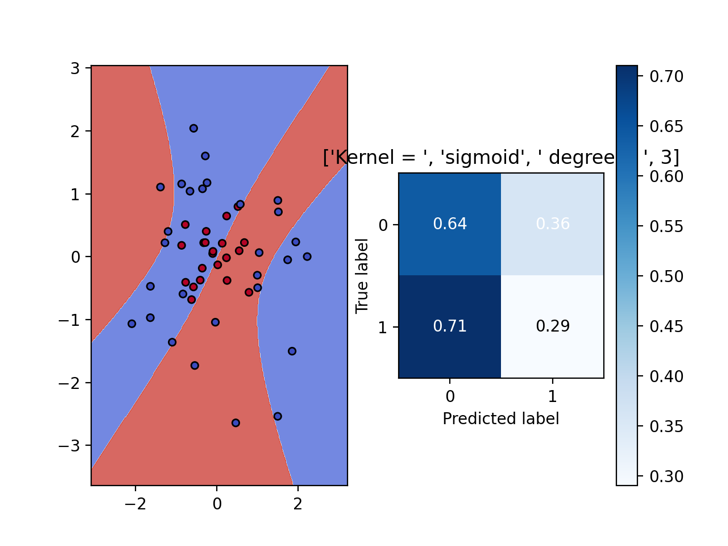
  
Из-за характера расположения точек линейное ядро попросту не может верно производить классификацию на данном датасете.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание  
Полиномиальное ядро справляется с задачей классификации на порядок лучше, но все зависит от степени полинома, как правило четные степени дают меньшую ошибку.

Изображение выглядит как снимок экрана

Автоматически созданное описание

По итогам тестирования разных ядер можно сказать, что лучшим выбором для классификации точек, находящихся внутри другого класса, будет Гауссово ядро.

Решение D

Изображение выглядит как снимок экрана

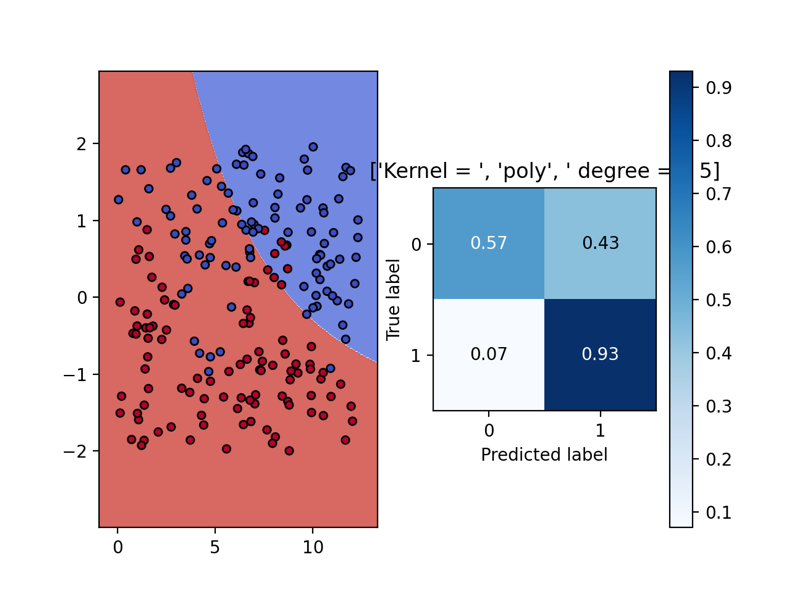
Автоматически созданное описание  
Для такого набора точек линейное ядро имеет явный недостаток, оно не способно выделять классы, находящиеся среди других, хотя и дает относительно не плохой результат.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описание  
Наиболее подходящим снова можно назвать Гауссово ядро.

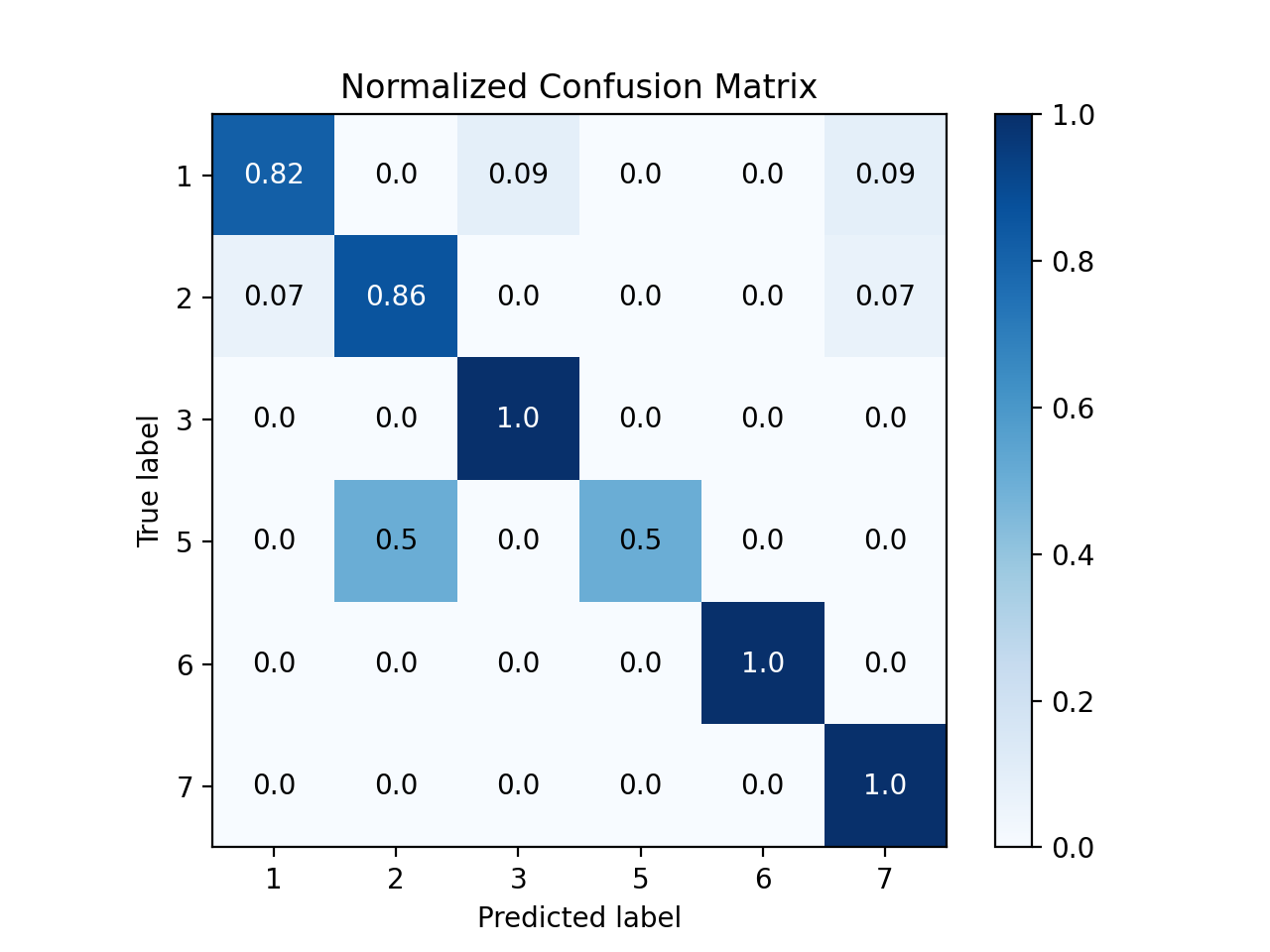
Задание №5  
Постройте классификаторы для различных данных на основе деревьев решений:

1. Загрузите набор данных Glass из файла glass.csv. Постройте дерево классификации для модели, предсказывающей тип (Type) по остальным признакам. Визуализируйте результирующее дерево решения. Дайте интерпретацию полученным результатам. Является ли построенное дерево избыточным? Исследуйте зависимость точности классификации от критерия расщепления, максимальной глубины дерева и других параметров по вашему усмотрению.
2. Загрузите набор данных spam7 из файла spam7.csv. Постройте оптимальное, по вашему мнению, дерево классификации для параметра yesno. Объясните, как был осуществлён подбор параметров. Визуализируйте результирующее дерево решения. Определите наиболее влияющие признаки. Оцените качество классификации.

Решение А  
Производил перебор по трем параметрам, типу расщепления, глубине и количеству признаков, по которым ищется лучшее разбиение в дереве. Так как диапазон параметров был достаточно велик, привожу только те параметры, при которых был получен лучший результат.

Изображение выглядит как текст, компьютер

Автоматически созданное описание

Приведем матрицу ошибок для лучшего результата и визуализацию его дерева решений.  


Изображение выглядит как карта

Автоматически созданное описание  
  
Такое дерево практически не имеет лишних вершин.

Решение В  
Тем же способом нашел оптимальные параметры для данных spam7  
  


Изображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как холодильник, комната

Автоматически созданное описание

Из визуализации дерева решений можно увидеть, что оно достаточно хорошо сбалансировано, но имеет некоторое количество лишних вершин. Так же можно отметить, что на решение по большей части влияют значения 2 и 4 параметров.

Задание №6  
Загрузите набор данных из файла bank\_scoring\_train.csv. На основе как минимум двух из изученных методов постройте классификаторы, предсказывающий значение столбца «SeriousDlqin2yrs». Определите, какой из классификаторов сработал лучше. Оцените качество классификации на bank\_scoring\_test.csv для обоих классификаторов.

Решение  
В данном случае довольно тяжело выделить лучший классификатор, поскольку и один из них не может корректно распознать второй класс. Но можно применить одну хитрость, а именно произвести инверсию классификатора kNN для второго класс, таем самым точность с 0.25 переменится на 0.75 и это уже значительно улучшает ситуацию.

Изображение выглядит как текст, карта

Автоматически созданное описание