Санкт-Петербургский политехнический университет Петра Великого

Институт Компьютерных наук и кибербезопасности

**Разработка графических приложений**

Отчет по Лабораторной работе 1 на тему:

«Классификация данных с помощью классификатора

K ближайших соседей»

Направление 09.04.01\_15 «Технологии проектирования системного и прикладного программного обеспечения»

Студент группы 5140901/31502 \_\_\_\_\_\_\_\_\_\_\_\_\_\_Алешковский А. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Никитин К.В.

Санкт-Петербург

2023

**Оглавление**

[**1. Аппроксимация плотности распределения.** 3](#_Toc150003408)

[**2. Классификация данных** 13](#_Toc150003409)

[**3. Выводы и анализы** 19](#_Toc150003410)

# **1. Аппроксимация плотности распределения.**

Было проведено 25 экспериментов с различными обучающими выборками и разными k, для метода k ближайших соседей. В таблице 1 приведены результаты эксперимента.

Таблица 1 – Результат 25 экспериментов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ntest** | **K** | | **Graph** | **Classification Error** |
| 100 | 3 | Очень плохой результат |  | 0.04808 |
| 15 | Плохой результат |  | 0.04651 |
| 30 | Средний результат |  | 0.04326 |
| 40 | Хороший результат |  | 0.03970 |
| 60 | Хороший результат |  | 0.03958 |
| 2000 | 3 | Очень плохой результат |  | 0.01846 |
| 15 | Плохой результат |  | 0.01812 |
| 30 | Средний результат |  | 0.01681 |
| 40 | Хороший результат |  | 0.01569 |
| 60 | Хороший результат |  | 0.01542 |
| 4000 | 3 | Очень плохой результат |  | 0.01464 |
| 15 | Плохой результат |  | 0.01397 |
| 30 | Средний результат |  | 0.01276 |
| 40 | Хороший результат |  | 0.01255 |
| 60 | Хороший результат |  | 0.01237 |
| 6000 | 3 | Очень плохой результат |  | 0.01420 |
| 15 | Плохой результат |  | 0.01342 |
| 30 | Средний результат |  | 0.01323 |
| 40 | Хороший результат |  | 0.01225 |
| 60 | Хороший результат |  | 0.01224 |
| 8000 | 3 | Очень плохой результат |  | 0.01369 |
| 15 | Плохой результат |  | 0.01351 |
| 30 | Средний результат |  | 0.01314 |
| 40 | Хороший результат |  | 0.01276 |
| 60 | Хороший результат |  | 0.01272 |
|  |  |  |  |  |
|  |  |  |  |  |

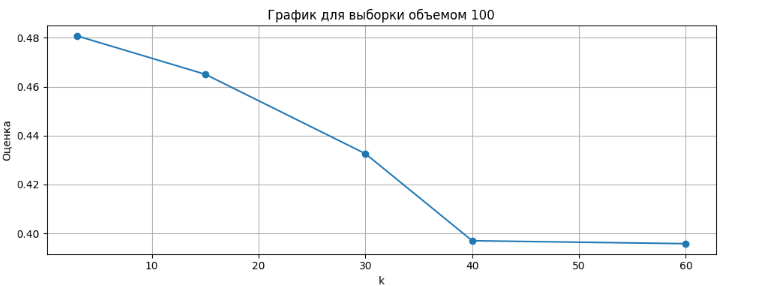
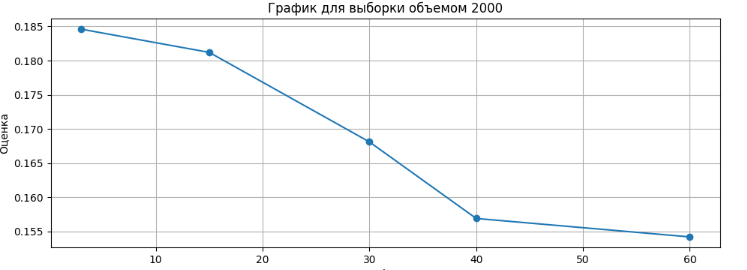


Рисунок 1 – График для выборки объемом 100

Рисунок 2 – График для выборки объемом 2000

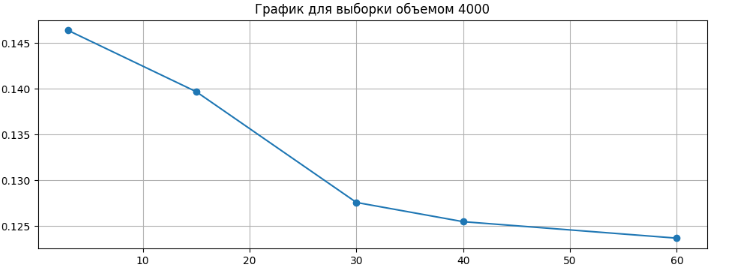


Рисунок 3 – График для выборки объемом 4000

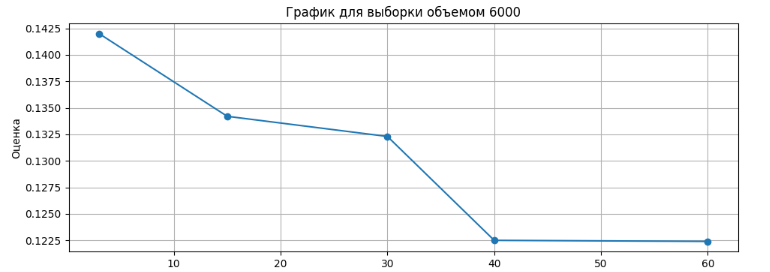


Рисунок 4 – График для выборки объемом 6000

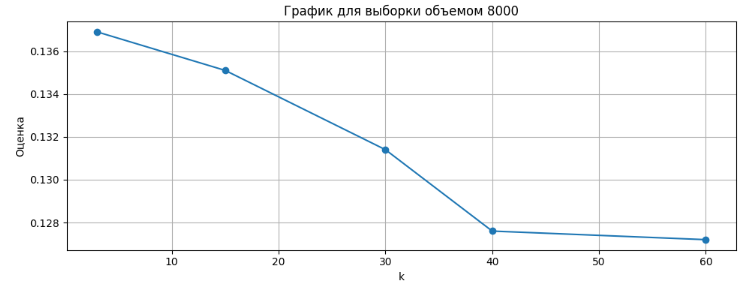


Рисунок 5 – График для выборки объемом 8000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ntrain | Classification Error | K | X1 plot | X2 plot |
| Plot error X1 |
| Plot error X2 |
| 100 | 0.04808 | 60 |  |  |
| 99.9885 |
| 99.9885 |
| 2000 | 0.01846 | 40 |  |  |
| 99.5389 |
| 99.5389 |
| 4000 | 0.01452 | 40 |  |  |
| 99.5119 |
| 99.5119 |
| 6000 | 0.01423 | 40 |  |  |
| 99.4875 |
| 99.4875 |
| 8000 | 0.01420 | 60 |  |  |
| 99.4923 |
| 99.4923 |

Таблица 2 - Результаты

Как можно видеть, наилучший результат достигается при объёме обучающей выборке равной 8000, и при размере **k** равным 60.

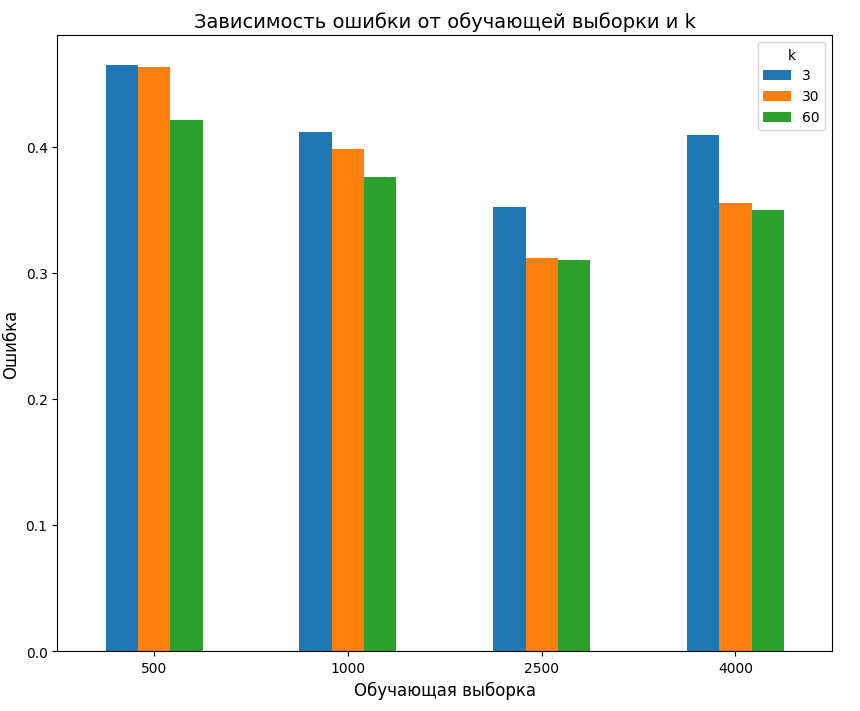
# **2. Классификация данных**

Было проведено 4 эксперимента, с использованием **Евклидовой метрики**, а также разными объёмами обучающих выборок. Целью экспериментов является поиск минимальной ошибки, а также поиск оптимальных параметров **K** и **nTrain**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| nTrain | K | Error | Graph |  |
| 500 | 3 | 0.04653 |  |  |
| 30 | 0.04633 |  |  |
| 60 | 0.04213 |  |  |
| 1000 | 3 | 0.04117 |  |  |
| 30 | 0.03986 |  |  |
| 60 | 0.03765 |  |  |
| 2500 | 3 | 0.03529 |  |  |
| 30 | 0.03123 |  |  |
| 60 | 0.03106 |  |  |
| 4000 | 3 | 0.0410 |  |  |
| 30 | 0.03555 |  |  |
| 60 | 0.03500 |  |  |

Самый лучший результат был достигнут при объёме обучающей выборке равного 2500 и K=30, средняя ошибка составила 3%, что является самым удовлетворительным результатом.

Теперь по результатам эксперимента, можно построить график, который показывает, как зависит ошибка от k и выборки обучения.



Далее будут проведены эксперименты с различными метриками.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **nTrain** | **K** | **Error** |
| euclidian | 1000 | 25 | 0.04091 |
| 50 | 0.03621 |
| 4000 | 25 | 0.03722 |
| 50 | 0.03601 |
| cityblock | 1000 | 25 | 0.0366 |
| 50 | 0.0360 |
| 4000 | 25 | 0.0356 |
| 50 | 0.0355 |
| chebychev | 1000 | 25 | 0.0375 |
| 50 | 0.0364 |
| 4000 | 25 | 0.0360 |
| 50 | 0.0355 |
| minkovski | 1000 | 25 | 0.0364 |
| 50 | 0.0360 |
| 4000 | 25 | 0.0360 |
| 50 | 0.0355 |
| Correlation | 1000 | 25 | 0.0355 |
| 50 | 0.0355 |
| 4000 | 25 | 0.0355 |
| 50 | 0.0355 |
| cosine | 1000 | 25 | 0.1296 |
| 50 | 0.1269 |
| 4000 | 25 | 0.1316 |
| 50 | 0.0355 |
| mahalanobis | 1000 | 25 | 0.0366 |
| 50 | 0.0361 |
| 4000 | 25 | 0.0357 |
| 50 | 0.0355 |

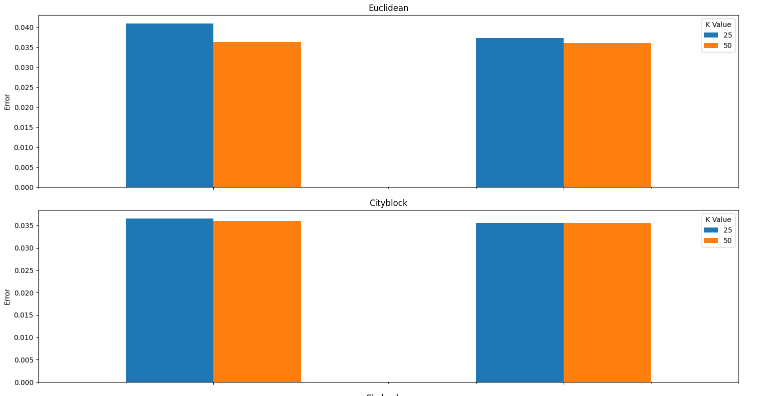


Рисунок 6 – Диаграмма результатов

# 

Рисунок 7– Диаграмма результатов

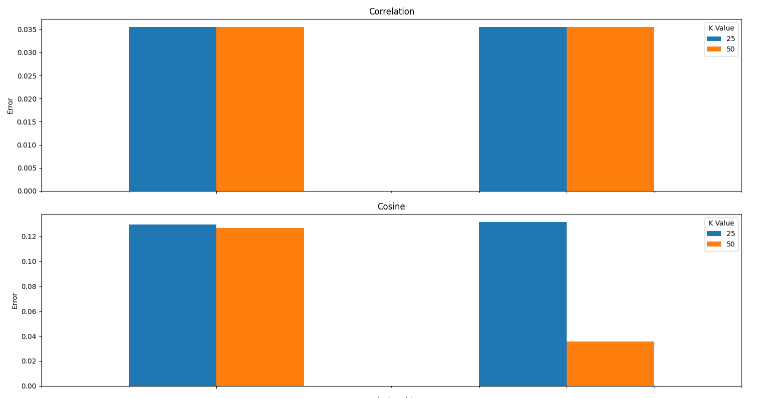


Рисунок 8– Диаграмма результатов

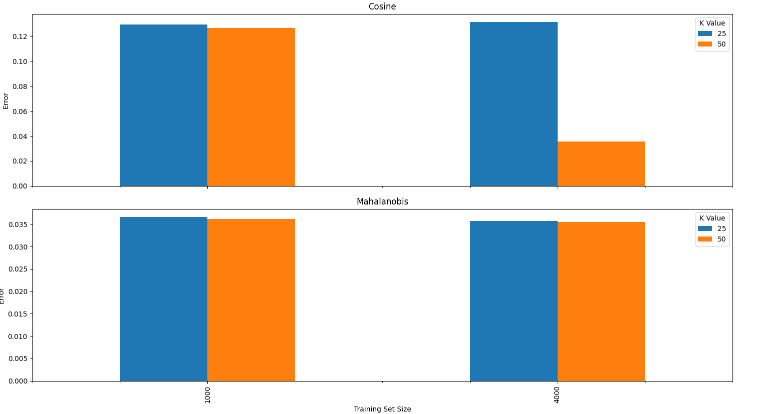


Рисунок 9– Диаграмма результатов

# **3. Выводы и анализы**

С помощью данного классификатора получилось добить наилучшего результата, 0.0355, что является показателем качественного оценивания.

Число K напрямую влияет на способность к классифицированию, так как чем оно больше, тем больше озвучивает объектов, однако нужно подбирать аккуратно, чтобы не вылезло коллизий, это же и касается объема обучающей выборки, если выбрать ее слишком большой, то может случиться переобучение, что скажется негативно на результатах классификации.

Так же их взаимосвязь очень велика, потому что при оптимальной обучающей выборке, получается подобрать оптимальное значение параметра K.

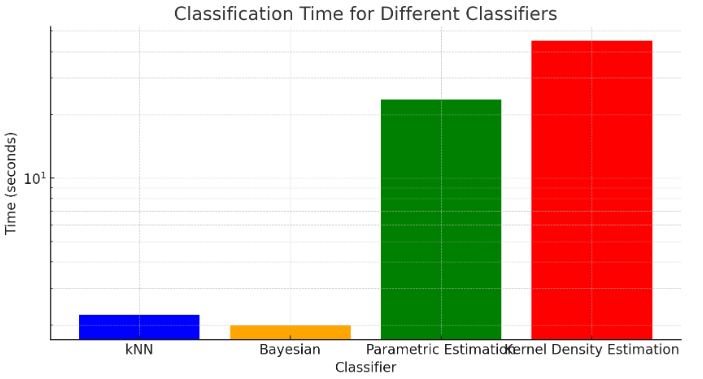


Рисунок 10 – Временные затраты при использовании различных классификаторов

1. Рассматривая классификацию с помощью метода kNN, можно сделать вывод, что данный подход при должной настройке параметров по качеству классификации не уступает рассмотренным в предыдущих пунктах методам. Подбор параметров для данного алгоритма является значительно более простым, чем для других рассмотренных подходов.

2. Качество оценивания плотности исходного распределения при использовании алгоритма kNN в среднем совпадает с качеством оценивания при применении метода с использованием ядер. Также стоит отметить, что совпадает не только численная величина ошибки аппроксимации, но и сама форма аппроксимирующей поверхности для алгоритма kNN и алгоритма с использованием ядер.

3. В общем случае увеличение числа k позволяет получать более гладкие области принятия решений и информацию в форме вероятностей. Однако слишком большое k нарушает локальность оценки, ухудшая качество классификации и увеличивая вычислительные затраты. Однозначно сказать, какая метрика лучше или хуже нельзя, так как выбор метрики сильно зависит от характера исходного распределения. Метрика тем лучше, чем лучше она отделяет признаки разных классов и сближает признаки одинаковых классов.

4. В среднем, чем больше объем обучающей выборки, тем лучше качество классификации, однако начиная с некоторого порога прибавка в эффективности становится несущественной. Также, чем больше объем выборке, тем больше и соответствующее оптимальное значение k.

5. Сравнивая быстродействие алгоритма kNN с рассмотренными ранее подходами, можно сделать вывод о том, что данный алгоритм является достаточно быстрым (если сравнивать с методом оценивания при помощи ядер). Стоит отметить, что алгоритм kNN может быть легко распараллелен, что позволит еще увеличить его быстродействие.