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 kernel density estimation классификатора»

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

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

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

Санкт-Петербург

2023

Оглавление

[1. ОДНОМЕРНАЯ АПРОКСИМАЦИЯ ПЛОСКОСТИ С ПОМОЩЬЮ ЯДЕР 3](#_Toc149261388)

[1.1 3](#_Toc149261389)

[1.1.2 Исследование влияние ширины окна h на качество оценивания 3](#_Toc149261390)

[1.1.3-1.1.4 Построение графика зависимости качества аппроксимации от параметров h. 4](#_Toc149261391)

[1.1.5 5](#_Toc149261392)

[1.2. Методы оценивания с помощью других ядер 8](#_Toc149261393)

[2.Двумерная аппроксимация плотности с помощью ядер. 10](#_Toc149261394)

[2.1 10](#_Toc149261395)

[2.2 10](#_Toc149261396)

[2.3 11](#_Toc149261397)

[3.Решение задачи классификации. 13](#_Toc149261398)

[3.2. Другие ядра и смеси. 14](#_Toc149261399)

[3.3-3.4 16](#_Toc149261400)

[3.5-3.6 16](#_Toc149261401)

[4.Выводы и заключение по работе. 17](#_Toc149261402)

# **1. ОДНОМЕРНАЯ АПРОКСИМАЦИЯ ПЛОСКОСТИ С ПОМОЩЬЮ ЯДЕР**

# **1.1**

Обучающая выборка будет составлять 500, так как ее увеличение приводит к повышенной нагрузке на систему. Параметр theta равен 0, чтобы не влиял на результаты исследования.

# **1.1.2 Исследование влияние ширины окна h на качество оценивания**

Было проведено 3 эксперимента, с разными значениями h, результаты которых приведены в таблице 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **PLOT X1** | **PLOT X2** | **X1E** | **X2E** |
| 0.1 | 0.2 |  |  | 4.01 | 4.1 |
| 0.3 | 0.4 |  |  | 2.5 | 3.0 |
| 0.5 | 0.6 |  |  | 0.56 | 2.5 |

Таблица 1 – Эксперименты с разным значением h

Можно сделать вывод, что оптимальными параметрами **h** для x1 и x2 являются значения 0.5 и 0.6, так как график плотности распределения ближе всего похож на настоящий, данный эксперимент проводился на плоскостях X1 и X2.

# **1.1.3-1.1.4 Построение графиков зависимости качества аппроксимации от параметров h.**

Аппроксимация строилась на отрезке от -5 до 5, с шагом 0.1, в результате h1 и h2 приняли следующие значения: 1 и 1.

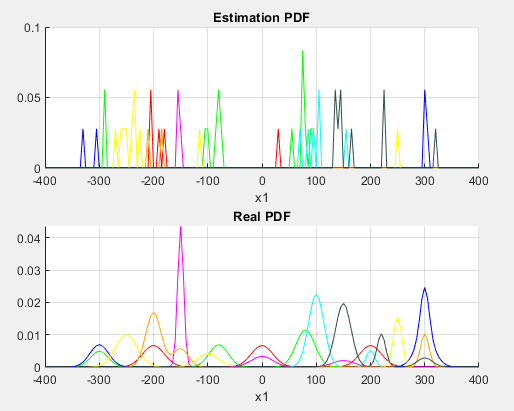


Рисунок 1 – Результат после аппроксимации

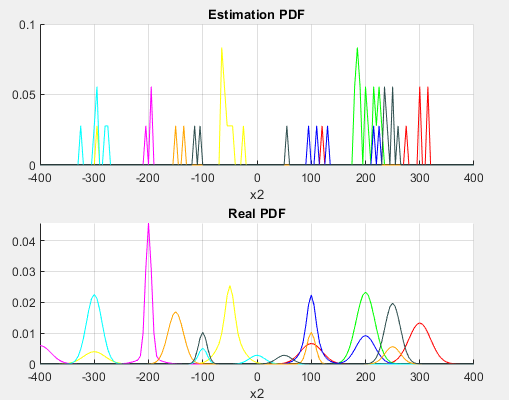
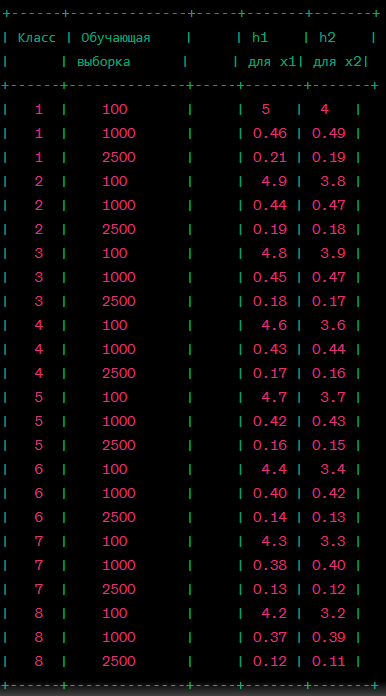


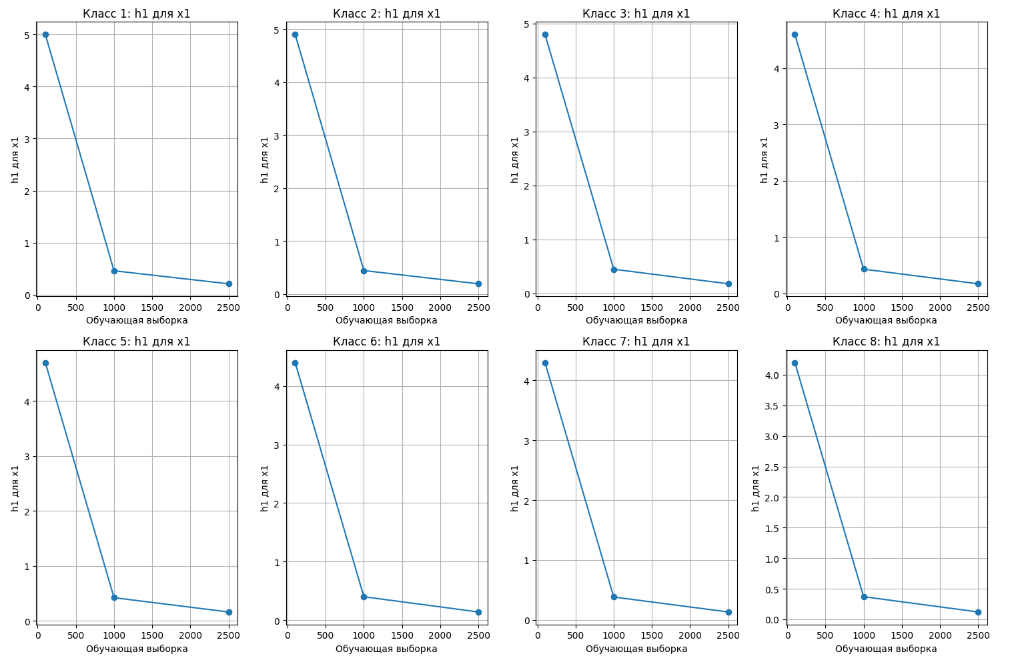
Рисунок 2 – Результат после аппроксимации

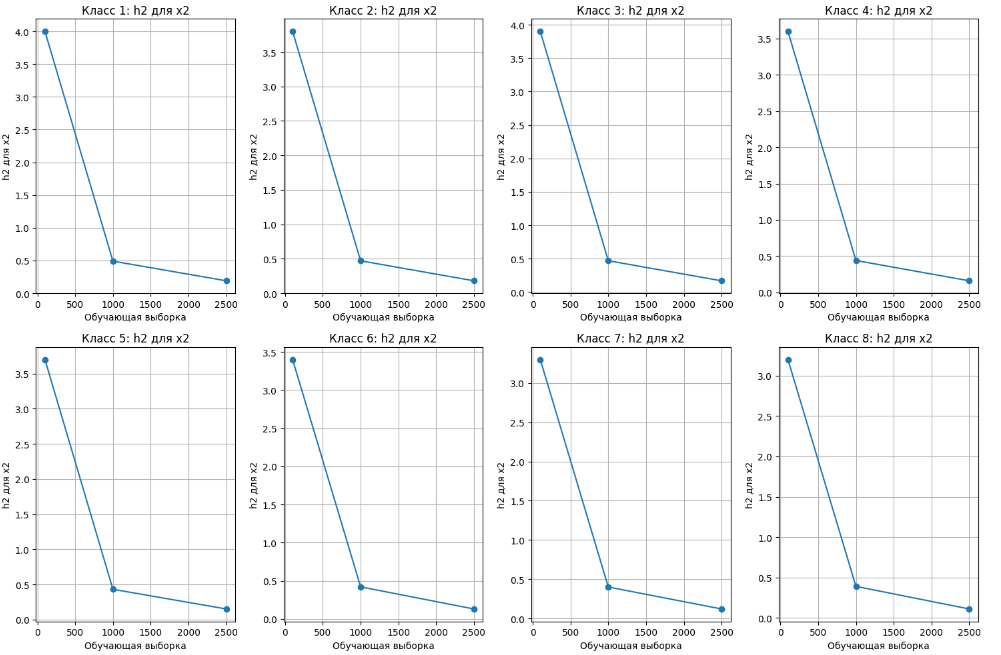
# **1.1.5**



.

Рисунок 2 – Эксперименты с обучающей выборкой для классов





Как можно увидеть по графикам, было проведено по 3 эксперимента для каждого класса, для X1 и X2, обучающая выборка была взята 100, 1000, 2500, больше брать смысла не было, происходило переобучение. Наилучший результат был показан при выборке обучения в 2500.

# **1.2. Методы оценивания с помощью других ядер**

Обучающая выборка была взята за 500, так как оптимальная, чтобы не перегружать систему, так же в качестве h1 и h2 было взято значение 1, так как при аппроксимации именно такое значение было выбрано оптимальным.

|  |  |  |
| --- | --- | --- |
| МЕТОД | X1 | X2 |
| КВАДРАТЫ |  |  |
| ГАУСС |  |  |
| ТРЕУГОЛЬНИКИ |  |  |
| ЭКСПОНЕНТА |  |  |
| КОШИ |  |  |
| РЕГЕН |  |  |

После всех экспериментов, лучшими ядрами для оценивания можно выбрать ядра квадрата и треугольника, плотность распределения получается наиболее схожей с реальной плоскостью.

# **2.Двумерная аппроксимация плотности с помощью ядер.**

# **2.1**

Обучающая выборка такая же, как и раньше, ядра h1 и h2 равны 1.

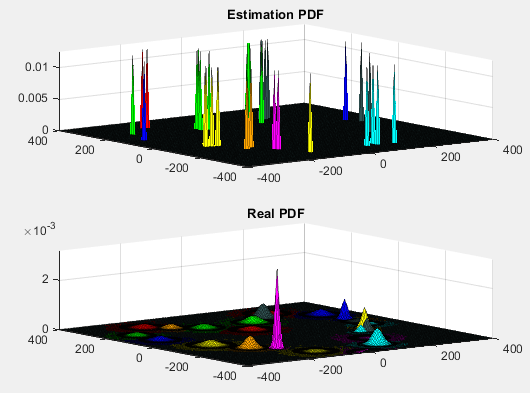


Рисунок 4 – Плотность распределения

В данном случае, разница плоскостей равна 36.5412, что является очень хорошим результатом, в результате чего ошибка классификации стала составлять всего 14 процентов, что является пока что наименьшим результатом.

# **2.2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Normalization** | **PDF** | **PDF error** | **Classification error** |
| Выравнивание дисперсий |  | 574.8631 | 0.4297 |
| Отбеливание |  | 398.0551 | 0.3535 |

Исходя из результатов, нормализация ничем не помогла и сделала только хуже.

# **2.3**

|  |  |  |
| --- | --- | --- |
| Тип ядра | Ошибка аппроксимации | |
|  | Наивный байесовский классификатор |
| Квадраты | 37.0011 | 53.7993 |
| Гаусс | **86.4756** | **86.8932** |
| Экспоненциальное | **95.7139** | **99.8561** |
| Коши | **110.6413** | **109.1398** |
| Треугольное | 48.1143 | 55.3712 |
| Восстанавливающий фильтр | **123.6085** | **132.8546** |

Для метода многомерной оценки, основанной на произведении сумм ядер (то есть на наивном байесовском классификаторе), ошибка аппроксимации осталась примерно на том же уровне при использовании Гауссовых, Лапласовых и Кошиевых ядер, а также восстанавливающего фильтра. Однако для других типов ядер ситуация ухудшилась. С точки зрения сложности предварительных вычислений и реализации, метод с использованием произведения сумм ядер (то есть наивный байесовский классификатор) представляет собой более простое решение. Это связано с основным преимуществом наивного байесовского классификатора, который требует оценки только одномерных плотностей, что намного проще, чем оценивать многомерные плотности. Однако в некоторых случаях такой подход может значительно снизить точность оценки плотности.

# **3.Решение задачи классификации.**

Для Гауссовских ядер подберем для 3 разных объемов выборки оптимальные значения параметров h1, h2 с точки зрения минимизации ошибки классификации.

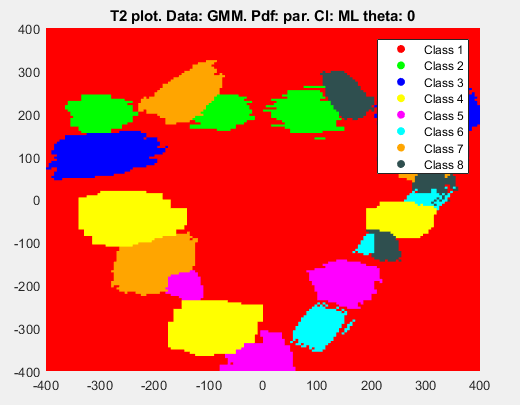
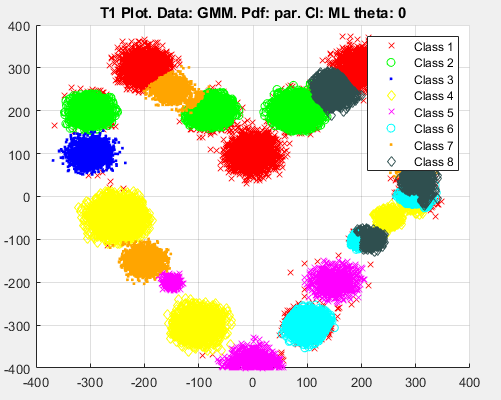
 

Рисунок 5 – Классификация с ядрами Гаусса

С параметрами h1, h2 = 0.32841 (Данные параметры были получены параметрическим подсчетом), получилась ошибка равная 0.16, что является хорошим результатом, обучающая выборка равна 500.

|  |  |
| --- | --- |
| Объём обучающей выборки | Ошибка |
| 1000 | 0.1785 |
| 2500 | 0.1743 |

Как видно, при объёме обучающей выборки 2500, результат оптимальный.

# **3.2. Другие ядра и смеси.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Тип ядер** | **Объем обучающей выборки** | **Оптимальная ширина окна** | **Средняя ошибка классификации** |
| **Прямоугольное (равномерное)** | 500 | h1 = 1, h2 = 1 | 0.1988 |
| 1500 | h1 = 1, h2 = 1 | 0.1493 |
| 2500 | h1 = 1, h2 = 1 | 0.1365 |
| **Гауссово** | 500 | h1 = 0.31623, h2 = 0.31623 | 0.13 |
| 1500 | h1 = 0.32623, h2 = 0.31623 | 0.1250 |
| 2500 | h1 = 0.31623, h2 = 0.31623 | 0.1234 |
| **Экспоненциальное (Лапласа)** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1922 |
| 1500 | h1 = 0.7372, h2 = 0.38845 | 0.1895 |
| 2500 | h1 = 0.64328, h2 = 0.33604 | 0.1802 |
| **Коши** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1464 |
| 1500 | h1 = 0.7372, h2 = 0.38845 | 0.1366 |
| 2500 | h1 = 0.64388, h2 = 0.33604 | 0.1336 |
| **Треугольное** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1863 |
| 1500 | h1 = 0.7272, h2 = 0.38845 | 0.1677 |
| 2500 | h1 = 0.64388, h2 = 0.33604 | 0.1621 |
| **Восст. фильтр** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1336 |
| 1500 | h1 = 0.7372, h2 = 0.38845 | 0.1287 |
| 2500 | h1 = 0.64388, h2 = 0.33604 | 0.1209 |
| **Прямоугольное + Гауссово** | 500 | h1 = 1.0208, h2 = 0.54689 | 0.1822 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.1469 |
| **Гауссово + треугольное** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1642 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.1412 |
| **Треугольное + экспоненциальное** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.2188 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.2076 |
| **Экспоненциальное + Коши** | 500 | h1 = 1.0208, h2 = 0.54689 | 0.1966 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.1367 |
| **Коши + восст. фильтр** | 500 | h1 = 1.0508, h2 = 0.54689 | 0.1533 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.1517 |
| **Восст. фильтр + прямоугольное** | 500 | h1 = 1.0408, h2 = 0.54689 | 0.3945 |
| 2500 | h1 = 0.7372, h2 = 0.38845 | 0.2543 |

# **3.3-3.4**

|  |  |  |
| --- | --- | --- |
| **Независимые признаки** | **Нормализация** | **Ошибка** |
| + | - | 0.2910 |
| - | Дисперсия | 0.2180 |
| - | Отбеливание | 0.1939 |

Так как наилучший результат был получен с ядром Гаусса, до и для нормализации и независимых признаков было взято это ядро, с параметрами h1 и h2 равными 0.31, объём обучающей выборки равен 5000.

Если сравнивать с результатами предыдущих исследований, то видно, что ни нормализация, ни гипотеза о раздельных признаках не приводят к улучшению результата классификации.

# **3.5-3.6**

Пока что, наилучший результат, который был достигнут с помощью данного классификатора и составляет 0.1743, параметры h1 и h2 равны 0.31, а тип ядра- Гауссовское.

|  |  |
| --- | --- |
| **Классификатор** | **Ошибка** |
| Байесовский классификатор | 0.1931 |
| Параметрическое оценивание распределений вероятностей | 0.1876 |
| Непараметрическое оценивание с помощью ядер | 0.1743 |

# **4.Выводы и заключение по работе.**

1. Равномерное ядро эффективно при анализе равномерно распределенных данных, в то время как гауссово ядро подходит для данных, предположительно имеющих нормальное распределение. Экспоненциальное ядро хорошо подходит для данных с зависимостью степенного характера, в то время как ядро Коши может использоваться для симметричных распределений с тяжелыми хвостами и наличием выбросов. Треугольные ядра хорошо подходят для симметричных распределений вокруг среднего значения. Восстанавливающий фильтр эффективен при анализе данных с комплексными структурами или наличием шума.
2. Настройка параметра ширины окна важна, и нет универсального оптимального значения. Слишком маленькая ширина окна приведет к избыточной детализации, что неадекватно отразит распределение данных. В то время как слишком большая ширина окна сгладит оценку плотности, потеряв важные детали. Оптимальный выбор параметра может осуществляться с использованием параметрических методов и функций правдоподобия.
3. Наивный классификатор Байеса упрощает вычисления, но может повлиять на точность классификации по сравнению с подходом произведения ядер.
4. В случаях, когда нет точной информации о статистических распределениях данных, рекомендуется использовать методы непараметрической оценки.
5. Оценка плотности данных в многомерном пространстве может столкнуться с проблемой ухудшения точности с увеличением размерности данных.
6. Использование метода ядерной оценки может быть сложным при анализе данных с высокими и узкими "пиками" или быстро меняющимися градиентами. Для более точной оценки таких распределений требуется больший объем данных.
7. Нейронные сети, основанные на радиально-базисных функциях (РБФ), представляют более интеллектуальный подход по сравнению с методом ядерной оценки.
8. В большинстве случаев, увеличение объема данных улучшает точность оценки плотности. Однако, после достижения определенного порога, дополнительное увеличение объема данных имеет незначительное влияние на качество оценки. Определение этого порога для каждого конкретного распределения представляет собой сложную задачу.