МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ   
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ   
«ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных систем

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ И ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Допущен к защите

Заведующий кафедрой ПМИ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.В. Разова

**Исследование метрических методов классификации**

Курсовой проект по дисциплине  
«Проектная и научно-исследовательская деятельность»

Выполнил студент группы ФИб-2301-51-00     / Д. О. Ощепков/

Руководитель к.п.н., профессор кафедры ПМИ     / Е. В. Котельников/

Работа защищена с оценкой     \_\_\_\_.\_\_\_\_.2023 г.

Члены комиссии:     / /

    / /

Киров 2023

Содержание

[Введение 3](#_Toc38539497)

[1 Название первого раздела 4](#_Toc38539498)

[1.1 Название подраздела 4](#_Toc38539499)

[1.1.1 Название пункта 4](#_Toc38539500)

[1.1.2 Название пункта 5](#_Toc38539501)

[1.2 Выводы по разделу 6](#_Toc38539502)

[2 Название второго раздела 7](#_Toc38539503)

[2.1 Название подраздела 7](#_Toc38539504)

[2.1.1 Название пункта 7](#_Toc38539505)

[2.2 Выводы по разделу 7](#_Toc38539506)

[Заключение 8](#_Toc38539507)

[Библиографический список 9](#_Toc38539508)

[Приложения 10](#_Toc38539509)

[Приложение А. Листинг программы 10](#_Toc38539510)

[Приложение Б. Перечень сокращений и обозначений 11](#_Toc38539511)

# Введение

Метрические методы классификации относятся к области обучения с учителем и основываются на измерении расстояния или сходства между объектами. Задача состоит в разделении данных на заранее определенные классы на основе известных обучающих примеров.

Метрические методы классификации применяются в различных областях жизнедеятельности человека:

* Биомедицина и медицинская диагностика. Метрические методы классификации находят применение в задачах биомедицинской классификации и диагностики, например, в определении заболеваний по медицинским данным или анализе медицинских изображений. [1]
* Обработка естественного языка (Natural Language Processing, NLP). В задачах анализа текста и обработки естественного языка метрические методы классификации используются для распознавания и классификации текстовых данных, например, в определении тональности текста или автоматической классификации текстовых документов. [2]
* Компьютерное зрение и обработка изображений. В области компьютерного зрения метрические методы классификации используются для распознавания и классификации изображений, например, в задачах распознавания образов или детектирования объектов на изображениях. [3]

Основные подходы в метрических методах классификации включают:

1. Метод одного ближайшего соседа (1-Nearest Neighbor). В этом методе объект классифицируется на основе меток его ближайшего соседей в обучающей выборке. Расстояние между объектами определяется с использованием выбранной метрики. Метод 1-NN прост в реализации, но может быть вычислительно требователен при больших объемах данных.
2. Метод k-ближайших соседей (k-Nearest Neighbors, k-NN). Этот метод расширяет метод 1-NN. Отличие заключается в количестве ближайших соседей, которые участвуют в классификации объекта. [5]
3. Взвешенный k-ближайших соседей (Weighted k-Nearest Neighbors): Этот метод расширяет метод k-NN, присваивая различные веса ближайшим соседям. Веса могут быть определены на основе расстояния или других факторов, чтобы учитывать важность каждого соседа при классификации. [6]
4. Метод парзеновского окна (Parzen Window): В этом методе объект классифицируется на основе ядерной функции, которая оценивает плотность объектов каждого класса в окрестности классифицируемого объекта. Ширина окна влияет на важность удаленности объектов при классификации. [7]
5. Метод потенциалов (Potential-based Classification): Этот метод основан на определении потенциала для каждого объекта и применении функции потенциала для классификации. Функция потенциала определяет близость объектов и их принадлежность к классам. [8]

Проблема исследования заключается в отсутствии четких рекомендаций или наставлений для выбора конкретного метрического метода классификации в конкретной задаче. В области машинного обучения существует широкий спектр метрических методов классификации, таких как метод k-ближайших соседей, методы на основе взвешенного голосования, локальные методы взвешенных соседей и другие. Каждый метод имеет свои сильные и слабые стороны, а также определенные условия применимости.

Для решения данной проблемы необходимо провести сравнительный анализ различных метрических методов классификации в разных сценариях и задачах. Это позволит определить, какие методы лучше справляются с определенными типами данных, объемом выборки, числом классов и особенностями признаков. Кроме того, рекомендации могут учитывать требования к точности классификации, скорости работы и интерпретируемости модели.

Целью курсового проекта является разработка рекомендации по выбору модели

Для достижения цели необходимо решить следующие задачи.

1. Обзор литературы. Изучить существующие научные статьи, книги и исследования, посвященные метрическим методам классификации. Выявить основные методы и их характеристики, а также рассмотреть примеры исследований, где эти методы применялись в различных задачах.
2. Описание и анализ основных метрических методов классификации
3. Подготовка набора данных. Собрать или использовать доступные наборы данных, соответствующие различным предметным областям. Обеспечить разнообразие данных в терминах признаков, классов, объема выборки и сложности задачи.
4. Реализация и применение каждого из методов для решения задач машинного обучения на различных наборах данных.
5. Анализ результатов и сравнение эффективности каждого метода на различных наборах данных. Определить подходящие метрики оценки качества классификации и провести серию экспериментов, сравнивая различные метрические методы классификации на подготовленных наборах данных. Сравнивать производительность методов в терминах точности, полноты, F-меры.
6. Сделать выводы о применимости каждого метода в различных сценариях и задачах классификации.

Курсовой проект включает два раздела. В первом разделе … Во втором разделе …

# Разбор метрических методов классификации

## Постановка задачи

Пусть на множестве объектов задана функция расстояния Существует целевая зависимость , значения которой известны только на объектах обучающей выборки Множество классов конечно. Требуется построить алгоритм классификации , аппроксимирующий целевую зависимость на всём множестве [4, c. 2]

Другими словами задача классификации заключается в том, чтобы разделить разные ситуации или объекты на группы или классы. Мы имеем некоторые данные об объектах, для которых мы знаем, к каким классам они относятся. Эти данные называются обучающей выборкой. Однако у нас есть и другие объекты, для которых мы не знаем, к каким классам они относятся. Наша задача состоит в разработке алгоритма, который сможет классифицировать любой из этих неизвестных объектов и определить, к какому классу он относится, исходя из имеющихся данных.

## Метрики оценки качества

Для оценки качества понадобятся следующие выкладки:

|  |  |  |  |
| --- | --- | --- | --- |
| Category | | Expert judgments | |
| YES | NO |
| Classifier  Judgement | YES |  |  |
| NO |  |  |

Таблица 1.2.1. The Contingency Table for Category . [9, p. 33]

|  |  |  |  |
| --- | --- | --- | --- |
| Category set | | Expert judgments | |
| YES | NO |
| Classifier  Judgement | YES |  |  |
| NO |  |  |

Таблица 1.2.2 The Global Contingency Table [9, p. 33].

Здесь – Истинно положительный (True Positive), – Ложно Положительный (False Positive), – Ложно Отрицательный (False Negative), – Истинно Отрицательный (True Negative).

### Precision и recall

Precision (точность) измеряет, насколько точно модель классифицирует положительные примеры. Она вычисляется как отношение числа правильно классифицированных положительных примеров к общему числу примеров, которые модель отнесла к положительному классу. Для класса :

Recall (полнота) измеряет, насколько хорошо модель обнаруживает все положительные примеры. Она вычисляется как отношение числа правильно классифицированных положительных примеров к общему числу положительных примеров в данных. Для класса :

Таким образом, precision полезна в задачах, где важно минимизировать ложно положительные результаты, а recall полезна в задачах, где важно минимизировать ложно отрицательные результаты.

Выбор между этими двумя метриками зависит от исходной задачи.

Одним из примеров, когда стоит предпочесть одну метрику другой, может быть задача классификации медицинских изображений для обнаружения редких заболеваний. В таком случае, полнота (recall) может иметь большее значение, поскольку нежелательно пропустить даже один положительный случай.

Напротив, если классификация используется для фильтрации спама в электронной почте, точность (precision) может иметь большее значение. В этом случае, важно минимизировать количество ложно положительных предсказаний, чтобы не пропустить важные сообщения, помечая их как спам.

### метрика и ее частный случай метрика

Является объединением Precision и Recall, показывает приоритет одной метрики над другой. При метрики учитываются в равной степени. [9, p. 33, 36]

-мера полезна если данные несбалансированны, так как она учитывает и точность, и полноту, что позволяет получить общую оценку производительности модели, не учитывая смещение в сторону преобладающего класса.

### метрика и ее частный случай метрика

Accuracy вычисляется как отношение числа правильно классифицированных примеров к общему числу примеров:

[10, стр. 71]

Accuracy может быть полезна, когда классы в данных сбалансированы, то есть количество примеров каждого класса примерно одинаково. В таких случаях она дает общую оценку производительности модели.

### ROC-AUC

## Вычислительная сложность рассматриваемых алгоритмов

Предположим, у нас есть обучающая выборка с N объектами, каждый из которых имеет d признаков, и тестовая выборка с M объектами, которые мы хотим классифицировать или предсказать их значения. Также предположим, что для поиска ближайших соседей используется простой перебор всех объектов.

Для каждого объекта из тестовой выборки мы должны найти k ближайших соседей в обучающей выборке. Сложность этого этапа зависит от выбранного алгоритма поиска ближайших соседей и может быть разной.

Простой перебор. В наивном подходе, где перебираются все объекты обучающей выборки, сложность составляет так как нам нужно сравнить каждый объект из тестовой выборки с каждым объектом из обучающей выборки.

Использование структур данных, таких как k-d деревья или шаровые деревья: В этом случае сложность может быть снижена до или даже если используется приближенный метод поиска ближайших соседей.

Таким образом, общая вычислительная сложность метода k-NN может быть представлена как для простого перебора или для использования структур данных, учитывающих оптимизации поиска ближайших соседей или даже для приближенных алгоритмов.

В методе потенциальных функций для нахождения потенциалов используется алгоритм с вычислительной сложностью , где обозначает количество итераций в алгоритме, а обозначает сложность вычисления потенциала для одного объекта. может быть равно для простого перебора, для использования структур данных, учитывающих оптимизации поиска ближайших соседей или для приближенных алгоритмов.

## Метод ближайшего соседа

### Формулировка

Самый простой из метрических методов. Он относит классифицируемый объект к тому классу, которому принадлежит ближайший обучающий объект:

[4, c. 3-4]

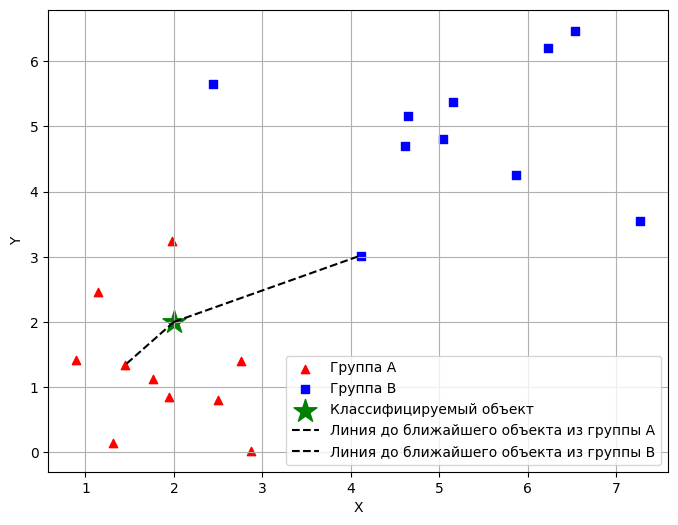


Рисунок 1.1 Визуализация метода ближайшего соседа. Классифицируемый объект следует отнести к классу A, так как ближайший к нему объект относится к этому классу.

### Достоинства и недостатки

Обучение сводится к запоминанию выборки . Единственное достоинство этого алгоритма – простота реализации. Недостатков гораздо больше:

* Неустойчивость к погрешностям. Если среди обучающих объектов есть выброс объект, находящийся в окружении объектов чужого класса, то не только он сам будет классифицирован неверно, но те окружающие его объекты, для которых он окажется ближайшим, также будут классифицированы неверно.
* Отсутствие параметров, которые можно было бы настраивать по выборке. Алгоритм полностью зависит от того, насколько удачно выбрана метрика ρ.
* В результате – низкое качество классификации. [4, c. 3-4]

### Алгоритм

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21 | Алгоритм метода ближайшего соседа:  Вход:  - trainSet: Обучающая выборка с N объектами  - testSet: Тестовая выборка с M объектами  - K: Количество классов  Выход:  - Предсказанные классы или значения для объектов из testSet    Для каждого объекта testObject из testSet:  Найти ближайший объект trainObject из trainSet к testObject:  minDistance = бесконечность  closestObject = None  Для каждого объекта trainObject из trainSet:  distance = Расстояние между testObject и trainObject  Если distance < minDistance:  minDistance = distance  closestObject = trainObject    Предсказать класс testObject на основе класса closestObject  Вернуть предсказанные классы или значения для объектов |

## Метод k ближайших соседей

### Формулировка

Метод ближайших соседей заключается в том, что каждый из соседей голосует за отнесение объекта к своему классу . Алгоритм относит объект u к тому классу, который наберёт большее число голосов:

[4, стр. 4]

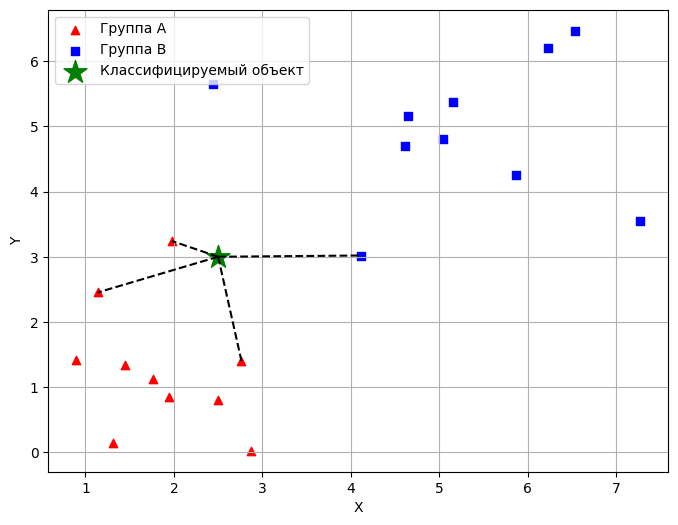


Рисунок 1.2 Визуализация метода k ближайших соседей. В качестве параметра k взято 4 соседа. Объект следует отнести к классу A, так как среди 4 его соседей 3 относится к этому классу, и лишь один объект относится к классу B.

### Алгоритм

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | Алгоритм взвешенных ближайших соседей:  Вход:  - trainSet: Обучающая выборка с N объектами  - testSet: Тестовая выборка с M объектами  - k: Количество соседей  Выход:  - Предсказанные классы или значения для объектов из testSet    Для каждого объекта testObject из testSet:  Найти k ближайший объектов trainObjects из trainSet к testObject:    class\_count = Cловарь, ключ - класс, значение - кол-во объектов  Для каждого объекта trainObject из trainObjects:  distance = Расстояние между testObject и trainObject  class\_count[trainObject.class] += 1    Предсказать класс testObject как наиболее встречающийся класс  Вернуть предсказанные классы для объектов |

### Достоинства и недостатки

Достоинства:

* Простота реализации

Недостатки:

* Приходится хранить обучающую выборку целиком
* Поиск ближайшего соседа предполагает сравнение классифицируемого объекта со всеми объектами выборки за операций. Для задач с большими выборками или высокой частотой запросов это может оказаться накладно. Проблема решается с помощью эффективных алгоритмов поиска ближайших соседей, требующих в среднем операций.
* В простейших случаях метрические алгоритмы имеют крайне бедный набор параметров, что исключает возможность настройки алгоритма по данным.

[4, с. 5]

## Метод взвешенных k ближайших соседей

### Формулировка

Недостаток kNN в том, что максимальная сумма голосов может достигаться на нескольких классах одновременно. В задачах с двумя классами этого можно избежать, если брать только нечётные значения k. Более общая тактика, которая годится и для случая многих классов – ввести строго убывающую последовательность вещественных весов , задающих вклад i-го соседа в классификацию:

Критерий выбора последовательности является эвристикой. [4, с. 4]

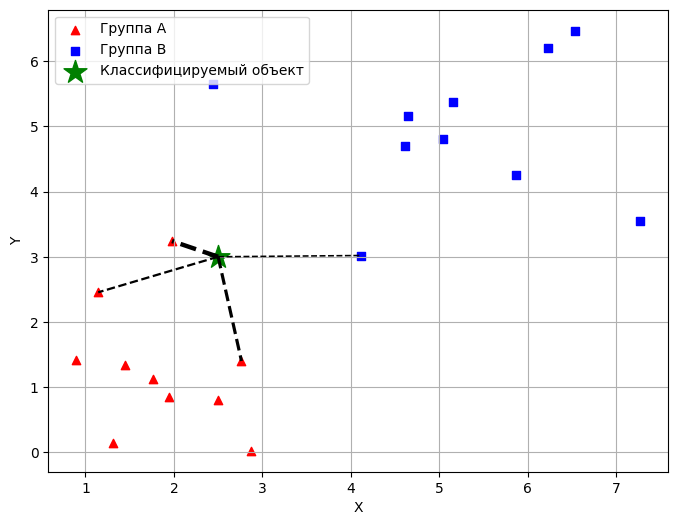


Рисунок 1.3 Иллюстрация метода взвешенных ближайших соседей. Чем толще линия, тем больший вклад вносит сосед в классификацию объекта.

### Алгоритм

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20 | Алгоритм взвешенных ближайших соседей:  Вход:  - trainSet: Обучающая выборка с N объектами  - testSet: Тестовая выборка с M объектами  - k: Количество соседей  Выход:  - Предсказанные классы или значения для объектов из testSet    Для каждого объекта testObject из testSet:  Найти k ближайший объектов trainObjects из trainSet к testObject:    class\_count = Cловарь, ключ - класс, значение - сумма весов объектов  Для каждого объекта trainObject из trainObjects:  distance = Расстояние между testObject и trainObject  class\_count[trainObject.class] += Вес(distance)    Предсказать класс testObject как наиболее весомый класс  Вернуть предсказанные классы для объектов  Вернуть предсказанные классы для объектов |

### Достоинства и недостатки

Метод имеет достоинства и недостатки метода k ближайших соседей, но добавляет возможность введения различных модификаций весовой функции, что является преимуществом.

## Метод парзеновского окна фиксированной ширины

### Формулировка

Ещё один способ задать веса соседям – определить как функцию от расстояния а не от ранга соседа. Введём функцию ядра невозрастающую на , и рассмотрим алгоритм

Параметр h называется шириной окна и играет примерно ту же роль, что и число соседей k. [4, c. 5]

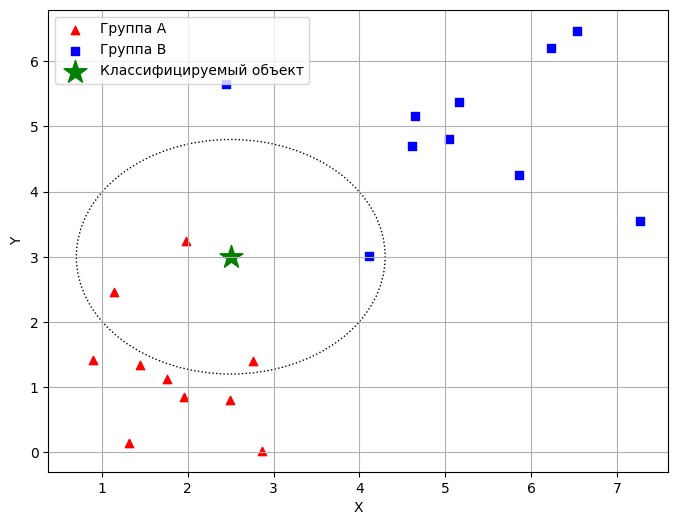


Рисунок 1.4 Иллюстрация метода парзеновского окна фиксированной ширины. При использовании финитного ядра (функция принимает нулевое значение за пределами некоторого компакта), таким образом объекты за пределами параметра h (ширина окна) не вносят свой вклад в классификацию объекта. При использовании нефинитных все объекты принимают участие в классификации, но чем дальше объект, тем меньший вклад он вносит.

### Алгоритм

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | Алгоритм взвешенных ближайших соседей:  Вход:  - trainSet: Обучающая выборка с N объектами  - testSet: Тестовая выборка с M объектами  - h: Ширина окна  Выход:  - Предсказанные классы или значения для объектов из testSet    Для каждого объекта testObject из testSet:  Найти k ближайший объектов trainObjects из trainSet к testObject:    class\_count = Cловарь, ключ - класс, значение - сумма вкладов объектов  Для каждого объекта trainObject из trainObjects:  distance = Расстояние между testObject и trainObject  class\_count[trainObject.class] += Ядро(distance/h)    Предсказать класс testObject как класс с наибольшим вкладом  Вернуть предсказанные классы для объектов |

### Преимущества и недостатки

Основное преимущество, которое дает метод парзеновского окна переменной ширины – это способность к учету локальной плотности. Он может хорошо справляться с областями, где данные плотно сгруппированы и имеют четкую границу.

Из недостатков можно выделить:

* Чувствительность к выбору ширины окна. При использовании фиксированной ширины окна метод становится чувствительным к выбору этой ширины. Неправильно выбранная ширина окна может привести к недооценке или переоценке плотности и, как следствие, к неправильным классификационным результатам. Проблема может быть решена, если применить метод парзеновского окна фиксированной ширины.
* Проблема с выбросами. Метод парзеновского окна фиксированной ширины может быть чувствителен к выбросам. Он не всегда хорошо справляется с областями, где есть выбросы или разреженные области данных.

## Метод парзеновского окна переменной ширины ширины

### Формулировка

Возьмём финитное ядро – невозрастающую функцию положительную на отрезке , и равную нулю вне его. Определим h как наибольшее число, при котором ровно ближайших соседей объекта u получают ненулевые веса Тогда алгоритм принимает вид

Выбор ядра K слабо влияет на качество классификации. На практике ядро либо задаётся априори, либо выбирается по скользящему контролю из нескольких стандартных ядер [4, с. 6].

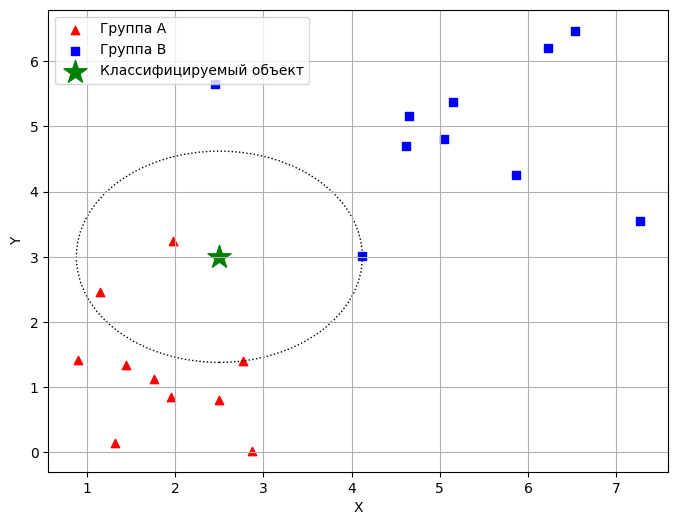


Рисунок 1.4 Иллюстрация метода парзеновского окна переменной ширины. Окружность вокруг классифицируемого объекта влючает k=4 объекта, которые будут учавствовать в его классификации.

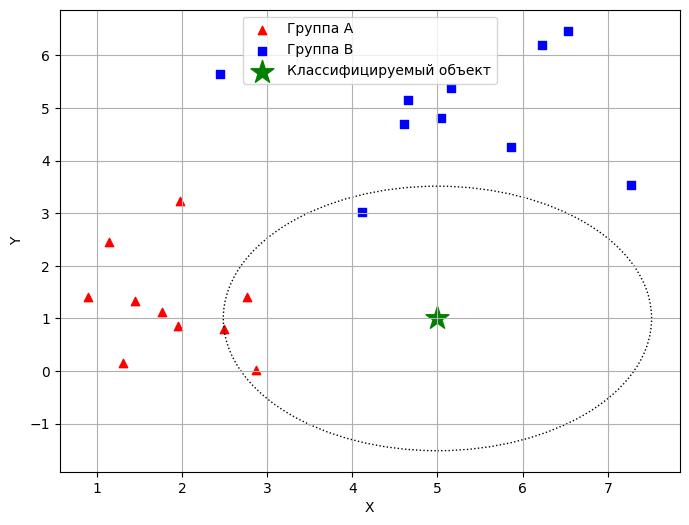


Рисунок 1.5 Иллюстрация метода парзеновского окна переменной ширины. Окружность вокруг другого классифицируемого объекта стала больше, чтобы включить k=4 объекта.

### Алгоритм

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21 | Алгоритм взвешенных ближайших соседей:  Вход:  - trainSet: Обучающая выборка с N объектами  - testSet: Тестовая выборка с M объектами  - h: Ширина окна  Выход:  - Предсказанные классы или значения для объектов из testSet    Для каждого объекта testObject из testSet:  Найти k+1 ближайший объектов trainObjects из trainSet к testObject:    h = trainObjects[-1]  trainObjects.pop()  class\_count = Cловарь, ключ - класс, значение - сумма вкладов объектов  Для каждого объекта trainObject из trainObjects:  distance = Расстояние между testObject и trainObject  class\_count[trainObject.class] += Ядро(distance/h)    Предсказать класс testObject как класс с наибольшим вкладом  Вернуть предсказанные классы для объектов |

### Преимущества и недостатки

Метод имеет те же преимущества, что и метод парзеновского окна фиксированной ширины, но при этом решает перечисленные в этой работе его недостатки, а именно:

* Чувствительность к выбору ширины окна.
* Проблема с выбросами.

## Выводы по разделу

Должны быть выводы по каждому разделу.

# Название второго раздела

## Название подраздела

### Название пункта

Пример текста.

Пример формулы во втором разделе:

|  |  |
| --- | --- |
|  |  |

Пример текста.

## Выводы по разделу

Должны быть выводы по каждому разделу.

# Заключение

Пример текста.

# Библиографический список

Xing W., Bei Y. Medical health big data classification based on KNN classification algorithm //IEEE Access. – 2019. – Т. 8. – С. 28808-28819.

Trstenjak B., Mikac S., Donko D. KNN with TF-IDF based framework for text categorization //Procedia Engineering. – 2014. – Т. 69. – С. 1356-1364.

Zhu Y. et al. Private-knn: Practical differential privacy for computer vision //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – С. 11854-11862.

Воронцов К. В. Метрические алгоритмы классификации. – 2013.

Peterson L. E. K-nearest neighbor //Scholarpedia. – 2009. – Т. 4. – №. 2. – С. 1883.

Dudani S. A. The distance-weighted k-nearest-neighbor rule //IEEE Transactions on Systems, Man, and Cybernetics. – 1976. – №. 4. – С. 325-327.

Kwak N., Choi C. H. Input feature selection by mutual information based on Parzen window //IEEE transactions on pattern analysis and machine intelligence. – 2002. – Т. 24. – №. 12. – С. 1667-1671.

Alimardani F. et al. Classification of bipolar disorder and schizophrenia using steady-state visual evoked potential based features //IEEE Access. – 2018. – Т. 6. – С. 40379-40388.

Sebastiani F. Machine learning in automated text categorization //ACM computing surveys (CSUR). – 2002. – Т. 34. – №. 1. – С. 1-47.

Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – Litres, 2022.

# Приложения

## Приложение А. Листинг программы

**Модуль чтения данных (read\_csv.py)**

import csv, sys

filename = 'some.csv'

# Стараемся хорошо комментировать код

with open(filename, newline='') as f:

reader = csv.reader(f)

try:

for row in reader:

print(row)

except csv.Error as e:

sys.exit('file *{}*, line *{}*: *{}*'.format(filename, reader.line\_num, e))

**Модуль записи данных (write\_csv.py)**

**import csv**

# Каждый фрагмент кода должен сопровождаться комментариями

**with** open('some.csv', 'w', newline='') **as** f:

writer = csv.writer(f)

writer.writerows(someiterable)

## Приложение Б. Перечень сокращений и обозначений

|  |  |
| --- | --- |
| ACL | Association for Computational Linguistics (Ассоциация компьютерной лингвистики); |
| AJAX | Asynchronous JavaScript And Xml (технология обращения к серверу без перезагрузки страницы); |
| API | Application Programming Interface (интерфейс прикладного программирования). |