

Київський національний університет імені Тараса Шевченка
Факультет інформаційних технологій

Кафедра програмних систем і технологій

ЗВІТ

з практичної роботи № 2

Тема: «Алгоритми класифікації»

Дисципліна «Спеціалізоване програмування автоматизованих систем»

Підготував:

студент гр. ПІЗ-33(1)

Мішак Максим

Перевірила:

Ніколаєнко Анастасія Юріївна

Умова : класифікатор найближчого сусіда редукований (косинусідна відстань)

Завдання :

Розробіть класифікатор за варіантом. Підготуйте тренувальний (навчальний) та тестовий набори даних для класифікатора. Дані можна взяти з файлу iris.csv та розділити на дві частини, де 90% віднести до тренувального (навчального) набору, а 10% до тестового. Навчіть класифікатор на тренувальному наборі даних. Застосуйте класифікатор до тестового набору даних і проаналізуйте точність результату – визначте відсоток правильних передбачень серед тестових даних.

Хід роботи :

Метод k -найближчих сусідів — це непараметричний метод навчання з учителем, вперше розроблений Евеліном Фіксом та Джозефом Ходжесом у 1951 році, а пізніше розвинутий Томасом Ковером. Метод використовується як для класифікації, так і для регресії. В обох випадках вхідні дані складаються з k найближчих навчальних прикладів у наборі даних. Результат залежить від того, для чого використовується k -NN для класифікації чи регресії:

- При класифікації k -NN результатом є належність класу. Об'єкт класифікується за допомогою множини голосів його сусідів, при цьому об'єкт відноситься до класу, найбільш поширеного серед його k найближчих сусідів (k — ціле додатне число, як правило, невелике). Якщо $k = 1$, то об'єкт просто приписується до класу цього єдиного найближчого сусіда.
- При k -NN регресії результатом є числове значення властивості об'єкта. Це значення є середнім із значень k найближчих сусідів.

Для виконання лабораторної роботи буде використовуватись дві бібліотеки : pandas , numpy . В файлі даному в методичних указаннях данні мали наступний вигляд :

```
"sepal.length", "sepal.width", "petal.length", "petal.width", "variety"  
5.1, 3.5, 1.4, .2, "Setosa"
```

В датасеті маємо 150 записів , які в подальшому розділяються на два сета : тренувальний та тестовий в співвідношенні 90/10

Код застосунку :

```
import pandas as pd
import numpy as np
DatSetIris = pd.read_csv('iris.csv')
TrainDataSet =
DatSetIris.sample(frac=0.9,random_state =1)
TestDataSet = DatSetIris.drop(TrainDataSet.index)
TrainCol =
TrainDataSet.drop('variety',axis=1).values
TrainRow = TrainDataSet['variety'].values
TestCol = TestDataSet.drop('variety',axis=1).values
TestRow = TestDataSet['variety'].values
def conusdistance(u,v):
    return 1-
np.dot(u,v) / (np.sqrt(np.sum(u**2)) * np.sqrt(np.sum(v*
*2)))
neighbors = []
def findnearesneighbor(TrainCol,TrainRow,TestCol):
    for x in TestCol:
        distance = [conusdistance(x,y) for y in
TrainCol]
        nearest_index = np.argmin(distance)
        neighbors.append(TrainRow[nearest_index])
    return np.array(neighbors)
def results(TestRow,neighbors):
    accuracy = np.sum(TestRow == predictions) /
len(TestRow)
    for i in range(len(neighbors)):
        print(f'Predicted: {neighbors[i]}, Actual:
{TestRow[i]}')
    print(f'\nAccuracy: {accuracy}')
predictions = findnearesneighbor(TrainCol, TrainRow,
TestCol)
results(TestRow, neighbors)
```

Скріншоти виконання :

```
/Users/macbook/Desktop/SPAS2/venv/bin/python  
Predicted: Setosa, Actual: Setosa  
Predicted: Setosa, Actual: Setosa  
Predicted: Setosa, Actual: Setosa  
Predicted: Versicolor, Actual: Versicolor  
Predicted: Versicolor, Actual: Versicolor  
Predicted: Versicolor, Actual: Versicolor  
Predicted: Virginica, Actual: Versicolor  
Predicted: Versicolor, Actual: Versicolor  
Predicted: Versicolor, Actual: Versicolor  
Predicted: Virginica, Actual: Virginica  
Predicted: Virginica, Actual: Virginica  
Predicted: Virginica, Actual: Virginica  
Predicted: Virginica, Actual: Virginica  
Predicted: Virginica, Actual: Virginica  
Predicted: Virginica, Actual: Virginica  
  
Accuracy: 0.9333333333333333  
  
Process finished with exit code 0
```

З результатів обчислення можемо зробити висновок - програма працює добре , з виведених записів бачимо , що всі наведені записи програма передбачила правильно , поле Ассурасу = 0.9(3) , що дорівнює 93% точності .

Висновок :

було виконано роботу по вивченню та написанню програми для класифікації даних по методу Найблищого сусіда , метод є редукований . Відстань між елементами було обраховано за допомогою косинусоїдної відстані