КН-Н921б

Сторожук О.Д.

Лабораторна робота №2

Тема роботи: «Метрики якості. Матриця невідповідності. ROC-аналіз»

Мета роботи: Дослідження методів оцінки якості класифікації даних. Розробка програмних засобів для оцінки якості побудованих моделей машинного навчання.

ЗАВДАННЯ ДО ВИКОНАННЯ

Розробити програму генерації псевдовипадкових чисел для 3641 об'єкту, із частотою появи «1», що дорівнює 55. Діапазон значень лежить від 1100 до 1200.

Також необхідно розробити програмний модуль оцінки якості побудованої моделі. При цьому побудувати графік ROC-кривої, розрахувати його площу за методом лівих прямокутників. Знайти порогове значення параметру cut-off. Також додатково знайти наступні параметри:

– True Positive (TP);

– False Positive (FP);

– False Negative (FN);

– True Negative (TN);

– Condition Positive (CP);

– Condition Negative (CN);

– Accuracy (ACC);

– Precision;

– Recall;

– F1 Score

ХІД РОБОТИ

**1. Генерація даних**

Алгоритм роботи генератору складається з наступних послідовних етапів:

– Загальна кількість об'єктів розбивається на 100 окремих об'єктів, якщо є залишок, він також включається окремо;

– Для кожної з групи генеруються 55 номерів об'єктів, відносно яких результат буде “1”;

– Після генерації списку об’єктів із істинним значенням, відбувається перехід до генерації поточних значень об’єктів у заданому діапазоні, в данній роботі від 1100 до 1200, середнє значення дорівнює 1150;

– У процесі генерації значень, якщо поточний об'єкт лежить у списку об'єктів із результатом “1”, то додатково викликається випадкова генерація від 0 до 1, аналог підкидання монети. Якщо в данному випадку генератор видає “1”, то в якості поточного значення об’єкту генерується число із діапазону [1150:1200], якщо “0” – то [1100:1200]. Приклад цього фрагменту наведено нижче:

while len(indexesList) < expectedOneAmountPer100:

    randValue = random.randint(0, 99)

    if indexesList.count(randValue) == 0:

      indexesList.append(randValue)

  for randomRowIndex in range(100):

    for colIndex in range(columnsCount - 1):

      insertValue = 0

      if randomRowIndex in indexesList:

        if random.randint(0, 1) == 1:

          insertValue = random.randint(avgNum, highestNum)

        else:

          insertValue = random.randint(lowestNum, highestNum)

      else:

        insertValue = random.randint(lowestNum, highestNum)

      rowData = rowData + str(insertValue) + ";"

    if randomRowIndex in indexesList:

      rowData = rowData + '1'

    else:

      rowData = rowData + '0'

    rowData = rowData + "\n"

  f.write(rowData)

Це зроблено для того, аби у подальшому площа AUC не дорівнювала значенню, наближеному до 0.5 через специфіку роботи вбудованого генератору чисел у програмі.

**2. Обробка даних**

Для побудови ROC-кривої необхідно визначити TPR і FPR на множині матриці невідповідностей для порогових значень (threshold), які належать множині значень [1100:1200]. Результати розрахунків наведені на рис. 2 і рис. 3.

Table

Description automatically generated

Рисунок 2 – Розраховані значення змінних Condition Positive, Condition Negative, Accuracy, Precision, Recall, F1 Score

Text

Description automatically generated

Рисунок 3 – Порогові значення (threshold), FPR, TPR

Розрахунки наведені вище дозволяють побудувати графік, що наведено на рис. 4. На прикладі також наведений розрахунок мінімальної відстані до точки {FPR = 0 : TPR = 1} за алгоритмом Евклідової відстані, значення порогу, завдяки якому отримується найменша відстань. Також розраховується площа отриманої фігури за методом лівих прямокутників, при цьому отримана площа порівнюється з тією, яка була отримана з бібліотеки sklearn.

Chart, line chart

Description automatically generated

Рисунок 4 – Графік ROC-кривої для сгенерованих даних. Значення площі дорівнює 0.6259, що свідчить про задовільний рівень якості класифікації

Графік був створений із застосуванням бібліотеки matplotlib.pyplot.

ВИСНОВКИ

Під час виконання лабораторної роботи було проведене дослідження методів оцінки якості класифікації даних. Були розроблені програмні засоби для оцінки якості побудованих моделей машинного навчання.