УДК 004.82:004:85

Бушменьов.В.Є., здобувач 1 курсу
СО Магістр
Зелінська О.В., доцент кафедри
інформаційних технологій

ПРИКЛАДНЕ ВИКОРИСТАННЯ ОПОРНИХ ВЕКТОРІВ НА ПРИКЛАДІ АНАЛІЗУ ДАТАСЕТУ ЗА ДОПОМОГОЮ МОВИ ПРОГРАМУВАННЯ PYTHON

Донецький національний університет імені Василя Стуса, м. Вінниця

Аналіз даних - дисципліна, що займається збором, вивченням та прийняттям рішень на основі аналізу великих обсягів даних про державу, суспільство, оточуючий світ, тощо. Опорні вектори - це метод аналізу даних для класифікації та регресійного аналізу за допомогою моделей з керованим навчанням з пов'язаними алгоритмами навчання, які називаються опорно-векторними машинами. В основі опорних векторів лежить деяка математична сутність – алгоритм максимізації деякої математичної функції відносно наявного набору даних. [1, 2]

Актуальність полягає в тому, що опорні вектори є корисними для категоризації текстів та гіпертекстів, класифікації зображень, тощо. Алгоритм ОВМ широко застосовується в біологічних та інших науках. Цей статистичний метод підтримує різні набори параметрів, які можуть набувати різну ефективність залежно від вхідного набору даних. [1, 5] Дізнавшись як використовувати цей метод доцільно та ефективно, ми зможемо перенести наші знання на конкретно поставлену проблему. В нашому випадку для дослідження ми будемо використовувати мову програмування “Python”, яка є програмним середовищем для математичних обчислень. [4]

Перейдемо до розгляду такого аналізу, який побудований на основі відкритих даних, які представлені на віртуальній платформі для змагань з машинного навчання та статистичного аналізу даних, в рамках якого статистики, програмісти та добувачі даних конкурують у створенні найкращих моделей для передбачення та візуального опису даних, запропонованих компаніями або іншими учасниками “Kaggle”. Ми будемо використовувати датасет “Kaggle” для прикладного дослідження методу опорних векторів. [3]

Опис датасету: в тренувальній виборці міститься 891 рядок з даними про пасажирів, а в тестовій меньше - 418. Цільовим відгуком в даному наборі даних є змінна “Survived”, яка вказує на те, чи вижила людина на пароплаві внаслідок кораблекрушіння чи померла. Поля “Pclass”, “Age”, “Sibsp”, “Parch”, “Fare” є числовими і вимагають перетворення. Візуальне вивчення даних показало, що признаки “Age” та “Fare” можуть не набувати значень взагалі, тобто комірки в деяких рядках є не заповненими. Для людей на пароплаві,

значення яких немає, я заміняю значення на медіану серед непустих значень цього признака. [3]

Під час аналізу ми будемо використовувати такі ядра опорних векторів : лінійне, поліноміальне та радіальне ядро. Варто зазначити, що даний датасет - є не збалансованим. Тобто, врятованих пасажирів набагато менше за тих, хто не вижив. Тобто в нас йде розподіл 38% до 62% процентів - свідчить про, те що датасет не є збалансованим.

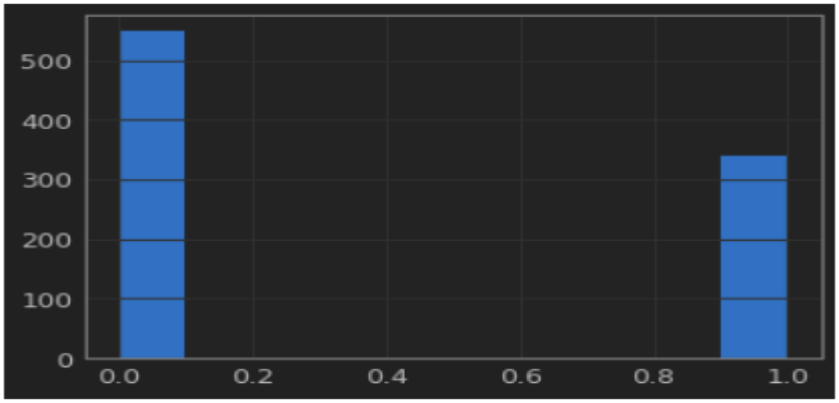

Рисунок 1 - Розподіл відгуків пасажирів

Для відбору найефективніших параметрів для метода опорних векторів ми будемо використовувати пошук по сітці з 5 кросс-валідаційними фолдами та з параметрами, які відображені нижче [Рисунок 2].

```
param_grid = {
    'preprocessor__num__imputer__strategy': ['mean', 'median'],
    'classifier__C': [0.01,0.1,0.5, 1.0, 1.5,2,5, 10],
    'classifier__class_weight': [None,"balanced"],
    'classifier__kernel':['linear', 'poly', 'rbf'],

}
```

Рисунок 2 - Пошук по сітці

Після цього, скориставшись отриманими підібраними найкращими параметрами підсумуємо результати алгоритму на тестовій та тренувальній виборці [Рисунок 3].

Training : 0.82
Test Accuracy : 0.84
Test F1 : 0.79

Рисунок 3 - результати пошуку по сітці

Спробуємо покращити результати. Так, як як результати моделі залежать від даних, на яких вона тренується - ми можемо поліпшити їх вигляд. Проаналізуємо залежність головних признаків в датасеті [Рисунок 4].

	Pclass	Age	SibSp	Parch	Fare
Pclass	1.000000	-0.365902	-0.078671	-0.029593	-0.548193
Age	-0.365902	1.000000	-0.160784	-0.279417	0.093143
SibSp	-0.078671	-0.160784	1.000000	0.419741	0.193970
Parch	-0.029593	-0.279417	0.419741	1.000000	0.239264
Fare	-0.548193	0.093143	0.193970	0.239264	1.000000

Рисунок 4 - Кореляція признаків

Ми бачимо, що "SibSp","Parch" дуже корилюють. Створимо з них спільну бінарну колонку. Тобто об’єднаємо в одну. Перетринуємо модель сіткою з попередніми параметрами та звіримо результати [Рисунок 5]. Результати покращились. Отже, які параметри моделі виявились найкращими [Рисунок 6].

```
Training Prefomence : 0.8255977496483825
Test Perfomence : 0.8651685393258427
F1 metric on SVC: 0.8000000000000002
```

Рисунок 5 - результати пошуку по сітці після поєднання признаків "SibSp","Parch" в один бінарний признак.

```
{'classifier__C': 0.5, 'classifier__class_weight': None, 'classifier__kernel': 'poly',
'preprocessor__num__imputer__strategy': 'median'}
```

Рисунок 6 - Найефективніші параметри моделі

Висновок. Отже, під час проведення дослідження ми виявили, як різні параметри можуть впливати на результати роботи статистичного методу опорних векторів. Побачили, що цей метод є ефективним для вирішення проблем бінарної класифікації з дисбалансованим набором даних. Зробили дослідження ефективності різного формату предобробки даних, які подаються на вхід одній й тій самій моделі та побачили їх ефективний вплив. Метод опорних векторів з поліноміальним ядром виявився найкращим на датасеті “Титанік” та набрав 86.5% точності на тестовій виборці.

Список використаних джерел

1. Аналіз даних - [Електронний ресурс]. Режим доступу: https://www.prostir.ua/?news=analiz-danyh-tsykl-bezkoshtovnyh-onlajn-kursiv-vid-prometheus

2. Методи опорних векторів - [Електронний ресурс]. Режим доступу: https://www.dstu.dp.ua/Portal/Data/74/72/3st13-17.pdf

3. Kaggle - [Електронний ресурс]. Режим доступу: https://uk.wikipedia.org/wiki/Kaggle
4. Practical statistics for Data Scientists: П. Брюс, Э. Брюс., O'Really Media Inc., 1005 Gravenstein Highway North, 2018. — 304 с.
5. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. An Introduction to Statistical Learning: with Applications in R. Springer New York Heidelberg Dordrecht London, 2017

УДК 004.82:004:85

Бушменьов.В.Є., здобувач 1 курсу СО Магістр
Нескородєва Т. В., д-р. техн. наук, доцент, завідувач кафедри інформаційних технологій

ДОСЛІДЖЕННЯ ПЕРЕВАГ ХМАРНОГО СЕРВІСУ AMAZON WEB SERVICES - SAGEMAKER ДЛЯ ПРИКЛАДНОГО ЗАСТОСУВАННЯ В МАШИННОМУ НАВЧАННІ

Донецький національний університет імені Василя Стуса, м. Вінниця

Машинне навчання - це підгалузь штучного інтелекту в галузі інформатики, яка часто застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» (тобто, поступово покращувати продуктивність у певній задачі) з даних, без того, щоби бути програмованими явно [5-6]. Amazon Web Services - це сама найпоширеніша у світі хмарна платформа з широкими можливостями, яка надає понад 200 повнофункціональних сервісів для центру обробки даних по всьому світі. [2]

Актуальність полягає в тому, що хмарні технології — відносно нове явище в Інтернеті в цілому і зокрема в машинному навчанні. Одним із основних переваг хмарних обчислень є наявна маштабуємість, цінова ефективність, інноваційність та безпечність. Саме ці фактори надають змогу маштабувати рішення негайно та мати вплив на певний віртуальний продукт. Головною особливістю хмарних сервісів є те, що вони надають можливість бізнесу сконцентруватись на продукті, його гнучкості та ефективності, а не витрачати час на розуміння та керування важкою програмною інфраструктурою. [2]

Перейдемо до розгляду одного із сервісів, які надає хмарний провайдер Amazon Web Services - SageMaker. Amazon SageMaker — це хмарна платформа машинного навчання, запущена в листопаді 2017 року. SageMaker дозволяє розробникам створювати, навчати та розгортати моделі машинного навчання у хмарі. SageMaker також дозволяє розробникам розгортати моделі машинного навчання на вбудованих системах і периферійних пристроях. [1]

Дослідимо, які переваги надає SageMaker: [3, 4]

1. Забезпечує легкий доступ до машинного навчання - дає змогу більшій кількості людей впроваджувати інновації за допомогою машинного