УДК 004.4:004.05
О.А. Кучеренко[1], О.О. Кучеренко[2]
[1]Тернопільський національний технічний університет імені Івана Пулюя, Україна
[2]Національний університет «Львівська політехніка», Україна

ОСОБЛИВОСТІ ПЕРЕДОБРОБКИ ДАНИХ ДЛЯ МЕТОДІВ ПРОГНОЗУВАННЯ

O.A.Kucherenko, O.O.Kucherenko

FEATURES DATA PREPROCESSING FOR FORECASTING METHODS

Попередня обробка даних (ПОД) у машинному навчанні (МН) - це важливий крок, який допомагає підвищити якість даних та сприяти вилученню з них змістовної інформації. Фактично - це метод підготовки (очищення та систематизації) необроблених даних, щоб зробити їх придатними для побудови та навчання моделей МН. Простіше кажучи, ПОД у МН - це метод інтелектуального аналізу даних, котрий трансформує необроблені дані в зрозумілий формат.

Коли йдеться про побудову моделі МН, така ПОД є першим кроком, що знаменує початок процесу. Як правило, реальні дані є неповними, непослідовними, неточними (містять помилки чи викиди) та часто не мають конкретних значень/тенденцій атрибутів. Саме тут у сценарій входить ПОД - вона допомагає очищати, форматувати і систематизувати необроблені дані, тим самим роблячи їх готовими до використання в моделях МН.

Можна схематично виділити етапи ПОД у МН:

– отримання даних. Щоб побудувати та розробити моделі МН, необхідно спочатку отримати відповідну кількість даних. Вона буде складатися з даних, зібраних з кількох та розрізнених джерел, які потім будуть об'єднані у належному форматі для формування. Формати таких наборів даних (НД) різняться залежно від застосування сценарію. Для прикладу, набір бізнес-даних абсолютно інший, ніж НД для медицини;

– імпорт бібліотеки. Найбільш використовуваними бібліотеками при обробці даних є бібліотеки pandas та numpy;

– імпорт даних. На цьому етапі необхідно імпортувати НД, зібраних для поточного проекту МН. У процесі імпорту НД необхідно витягти залежні та незалежні змінні. Для кожної моделі МН необхідно розділити незалежні змінні (матрицю функцій) та залежні змінні НД. Щоб отримати незалежні змінні, необхідно використовувати функцію iloc[] бібліотеки Pandas;

– виявлення та обробка відсутніх значень. При ПОД дуже важливо ідентифікувати та правильно обробляти відсутні значення, в іншому випадку ви можете зробити неточні та помилкові висновки на основі даних. Зайве говорити, що це завадить вашому проекту МН. У БД немає пропущених значень, тому цей крок опускається;

– кодування категоріальних даних. Вони належать до інформації, що має певні категорії в НД. Моделі МН, в основному, ґрунтуються на математичних рівняннях. У НД категоріальної змінної є стовпець з інформацією про рентабельність підприємства;

– розділення НД. Будь-який НД для моделі МН повинен бути поділеним на два окремих НД - навчальний і тестовий. Перший – це підмножина НД, котра застосовується для навчання моделі МН. Другий – це підмножина НД для тестування моделі МН. Зазвичай НД розбивається на співвідношення 70% та 30% або співвідношення 80% та 20%. Процес поділу залежить від форми та розміру аналізованого НД;

– масштабування функцій. Означає кінець ПОД у МН. Це метод стандартизації незалежних змінних НД у межах певного діапазону.