УДК 004

Вознюк М. Ю. студент, гр. ІПЗм-20-2,
Левківський В. Л., ст. викладач кафедри
Державний університет «Житомирська політехніка»

АНАЛІЗ МОДЕЛЕЙ КЛАСИФІКАЦІЇ В PYTHON ДЛЯ СТВОРЕННЯ СИСТЕМИ АВТОМАТИЧНОЇ КАТЕГОРИЗАЦІЇ ПУБЛІКАЦІЙ БЛОГУ

Категоризація публікацій блогу передбачає створення веб-додатка, який, ґрунтуючись на попередніх записах, показує категорію до якої можна віднести нову публікацію. Це досягається за допомогою моделі класифікації машинного навчання з учителем.

Для вибору моделі класифікації проведено аналіз трьох популярних моделей класифікації. Цей аналіз дав наступні результати:

Наївний баєсів класифікатор (Naive Bayes) – це один з найпростіших класифікаторів машинного навчання. Як випливає з назви, цей алгоритм припускає, що всі змінні в наборі даних «наївні» тобто, не корелюють один з одним.

Такий класифікатор обчислює можливість приналежності об’єкта до якогось класу. Ця ймовірність обчислюється з шансу, що якась подія відбудеться, з опорою на події, що вже відбулися. Кожен параметр об’єкта, що класифікується, вважається незалежним від інших параметрів.

Переваги Naive Bayes: легко і швидко передбачає клас тестового набору даних; добре справляється із багатокласовим прогнозуванням; добре працює з категоріальними ознаками (порівняно з числовими), продуктивність наївного байєсовського класифікатора краща, ніж в інших простих методів, більш того, йому потрібно менше навчальних даних.

Недоліки Naive Bayes: обмеженням даного алгоритму є припущення незалежності ознак, однак у реальних завданнях цілком незалежні ознаки трапляються вкрай рідко; якщо змінна має категорію (в тестовому наборі даних), яка не спостерігалася в навчальному наборі даних, то модель надасть 0 (нульову) ймовірність і не зможе зробити прогноз.

Класифікатор дерева рішень (Decision Tree Classifier) – це класифікатор, що розбиває дані на все менші та менші підмножини на основі різних критеріїв, тобто у кожної підмножини своя сортуюча категорія. З кожним поділом кількість об’єктів певного критерію зменшується. Класифікація добігає кінця, коли мережа дійде до підмножини лише з одним об’єктом.

Переваги Decision Tree: інтерпретованість моделі; дерева рішень можуть легко візуалізуватися як сама модель (дерево), так і прогноз для окремого взятого тестового об’єкта (шлях у дереві); швидкі процеси навчання та прогнозування; мінімальна кількість параметрів моделі; підтримка і числових та категоріальних ознак.

Недоліки Decision Tree: дерева дуже чутливі до шумів у вхідних даних; розділяюча межа, побудована деревом рішень, має свої обмеження і на практиці дерево рішень за якістю класифікації поступається деяким іншим методам.

Метод k-найближчих сусідів (K-Nearest Neighbors) – теж досить популярний метод класифікації. На рівні інтуїції суть методу така: подивися на сусідів, які переважають, такий і ти. Формально основою методу є гіпотеза компактності: якщо метрика відстані між прикладами введена досить вдало, то подібні приклади набагато частіше лежать у одному класі, ніж у різних.

Переваги K-Nearest Neighbors: проста реалізація; можна адаптувати під необхідне завдання вибором метрики чи ядра; непогана інтерпретація, можна пояснити, чому тестовий приклад класифіковано саме так.

Недоліки K-Nearest Neighbors: метод вважається досить швидким, але в реальних завданнях, як правило, кількість сусідів, що використовуються для класифікації, буде більшою (100-150), і в такому разі алгоритм працюватиме не так швидко, як дерево рішень; немає теоретичних підстав вибору певної кількості сусідів — лише перебір; якщо в наборі даних багато ознак, то важко підібрати відповідні ваги та визначити, які ознаки не важливі для класифікації/регресії.

Отже, як показує даний аналіз, всі розглянуті класифікатори можна використовувати для вирішення завдання створення класифікатора тексту. Однак, наївний баєсів класифікатор має найбільше вагомих плюсів відносно мінусів. Більш того, він успішно застосовується у багатьох додатках, від текстової аналітики та спам-фільтрів до систем рекомендацій, тому добре підійде і для категоризації публікацій.

Список використаних джерел:

1. Огляд методів класифікації у машинному навчанні за допомогою Scikit-Learn [Електронний ресурс] – Режим доступу до ресурсу: https://tproger.ru/translations/scikit-learn-in-python/.