

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ АЛГОРИТМІВ ЛІНІЙНОЇ РЕГРЕСІЇ У СЛУЖБІ МАШИННОГО НАВЧАННЯ MICROSOFT AZURE

Технології хмарних обчислень у сучасному інформаційному світі стають все популярнішими. Однією з найбільш розвинених хмарних платформ є Microsoft Azure, яка допомагає клієнтам і партнерам створювати та реалізовувати стратегії для досягнення поставлених цілей в хмарному середовищі. Перелік пропонованих послуг Azure включає і технології штучного інтелекту, зокрема, технології машинного навчання (Machine Learning, ML). Спеціально створена Служба машинного навчання Azure (Azure Machine Learning) надає ряд послуг щодо розробки і розгортання рішень, а Конструктор дозволяє спростити процес створення, навчання та оцінки моделей. Конструктор – це спеціалізований засіб візуального проектування, котрий дозволяє створювати моделі машинного навчання без жодного рядка коду і містить ряд розроблених алгоритмів машинного навчання. Кожен алгоритм призначений для вирішення різних типів проблем машинного навчання.

Часто розробники постають перед питанням: «Який алгоритм машинного навчання слід використовувати?». Обраний алгоритм залежить, головним чином, від двох різних аспектів сценарію обробки і аналізу даних:

1. *Що ви хочете зробити з даними?* Зокрема, яким є бізнес-питання, на яке ви шукаєте відповідь за допомогою навчання на попередніх даних?

2. *Які вимоги до сценарію обробки і аналізу даних?* Зокрема, яка точність, час навчання, лінійність, кількість параметрів і число функцій, підтримуваних рішенням?

Якщо бізнес-питання можна сформулювати у вигляді «Скільки?», то варто застосовувати групу алгоритмів, що називається регресією. Група алгоритмів регресії включає: лінійну регресію, регресію лісу прийняття рішень, регресію зростаючого дерева прийняття рішень, регресію нейронної мережі.

Відповіддю на друге питання при виборі сценарію є налаштування відповідного алгоритму регресії. Якщо вимоги до сценарію обробки можна визначити як достатньо висока точність, малий час навчання, лінійність алгоритму та невелика кількість параметрів, то найкращим рішенням є застосування алгоритму лінійної регресії. Лінійна регресія – це загальновідомий статистичний метод, який був прийнятий в машинному навчанні досить давно, доповнений багатьма вдосконаленнями для підгонки результатів і вимірювання помилок. Даний метод, як правило, добре працює на великих і розріджених наборах даних, що не мають складних трендів.

Класична задача регресії включає в себе одну незалежну змінну і залежну змінну. Нажаль, множинна лінійна регресія та багатофакторна лінійна регресія не підтримуються, натомість можна скористатися іншими видами регресії. А для створення регресії з декількома мітками (прогнозування кількох залежних змінних в рамках однієї моделі) рекомендується створювати окреме навчання для кожного виходу, який потрібно спрогнозувати. Для вимірювання помилок і відповідності лінії регресії можна застосувати звичайний метод найменших квадратів, або градієнтний спуск.

Метод найменших квадратів є одним з найчастіше використовуваних методів лінійної регресії, що базується на обчисленні похибки як суми квадрата розбіжності фактичного значення і прогнозованої лінії, та підганяє модель, мінімізуючи квадратичну похибку. При виборі цього методу у моделі додатково можна налаштувати ваговий коефіцієнт спрощення рівня «основний» L2 та початкове число випадкових чисел для заповнення генератора випадкових чисел.

L1- і L2-регуляризація – це два тісно пов'язаних методи, які застосовують для зменшення ступеня перенавчання моделі, що забезпечує більш якісне прогнозування. L1-регуляризація іноді дає корисний побічний ефект видалення непотрібних функцій, присвоюючи їх вагам значення 0.0. Однак L1-регуляризація не працює з усіма методами навчання. L2-регуляризація працює з усіма формами навчання, але не забезпечує неявної селекції функцій. На практиці слід використовувати метод проб і помилок, щоб визначити, яка форма регуляризації краще для конкретного завдання.

Метод градієнтного спуску зменшує кількість помилок на кожному кроці процесу навчання моделі. При виборі цього варіанту для методу рішення можна задати різні параметри для управління розміром кроку, частотою навчання і т.д. Цей параметр також підтримує використання вбудованого очищення параметрів. Таким чином, якість розробленої регресійної моделі машинного навчання залежить від якості навчання відповідного модуля і потребує ітераційного та ретельного процесу налаштування параметрів.