

ХМАРНЕ РІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ МЕДИЧНИХ ДАНИХ НА ОСНОВІ КОНСТРУКТОРА МАШИННОГО НАВЧАННЯ AZURE

Машинне навчання (МН) є основою для більшості рішень штучного інтелекту, а створення інтелектуального рішення часто починається з використання МН для навчання прогнозної моделі за допомогою зібраних даних за минулі періоди.

МН широко використовується в медицині. На основі зібраних даних про пацієнтів, навчена модель може застосовуватися для передбачення ризику або виявлення прихованих симптомів захворювання і підбору відповідних методів лікування, що особливо важливо для лікування захворювань, які вкрай бажано діагностувати на ранніх стадіях.

Навчання і розгортання ефективної моделі МН вимагає багато часу, зусиль і ресурсів. Машинне навчання Azure – це хмарна служба, яка допомагає спростити деякі завдання і скоротити час, що витрачається на підготовку даних, навчання моделі і розгортання прогнозної служби.

Існує чотири основні кроки для створення додатка (або моделі) машинного навчання:

Крок 1. Вибір та підготовка набору даних для навчання.

Крок 2. Вибір алгоритму для роботи з навчальним набором даних.

Крок 3. Навчання алгоритму для створення моделі.

Крок 4. Використання і поліпшення моделі.

Пропонована модель буде розроблятися за допомогою Конструктора Машинного навчання Azure, який дозволяє створювати моделі МН без єдиного рядка коду, шляхом візуального з'єднання наборів даних і модулів обробки на інтерактивному полотні для створення моделей МН.

Один з найважливіших кроків – вибір алгоритму для моделі. Тип алгоритму залежить від типу даних і їх кількості в навчальному наборі, а також від типу задачі, яку потрібно вирішити.

Для рішення задачі були застосовані алгоритми контрольованого навчання, а саме класифікації.

Класифікація – це приклад контрольованого машинного навчання, в ході якого модель навчається за допомогою використання даних, що включають в себе як компоненти, так і відомі значення (мітки). Таким чином, модель навчається виконувати підгонку комбінацій компонентів до мітки. Після завершення навчання можна використовувати навчену модель для прогнозування міток для нових компонентів, для яких мітка невідома.

Серед усіх методів контрольованої класифікації було обрано – дерево прийняття рішень. Це метод застосовується в машинному навчанні, при аналізі даних і в статистиці, та в даному випадку допомагає розв'язати задачу класифікації медичних даних, коли цільова змінна є дискретною.

Існують різні модифікації методу дерев прийняття рішень. Наша модель використовує модуль двокласового збільшеного дерева прийняття рішень, що використовує ансамблі простих методів. Збільшене дерево прийняття рішень – це метод навчання ансамблів в якому друге дерево виправляє помилки першого дерева, а третє дерево виправляє помилки першого і другого дерев і т. д. Прогнози засновані на повному ансамблі дерев, які роблять прогноз.

Після додавання модуля в конвеєр налаштовуються його параметри. Коли основні параметри модуля задані, можна переходити до навчання моделі.

Оцінка моделі проводиться за такими ознаками:

Акуратність. Пропорція точних прогнозів по відношенню до загальної кількості прогнозів, тобто це ймовірність того, що клас буде передбачений правильно.

Точність. Частка позитивних варіантів визначених вірно.

Повнота. Частка варіантів, класифікованих як позитивні, які насправді виявилися позитивними.

Оцінка F1. Загальна метрика, яка об'єднує в собі інформацію про повноту і точність алгоритму.

Матриця неточностей. Це матриця розміром N на N, де N – кількість класів яка представляє собою табличне представлення прогнозованих і фактичних значень для кожного можливого класу.

Для наочної оцінки якості алгоритму застосовується ROC-крива. Крива будується на площині, визначеній осями повноти по вертикалі і частоти помилкових спрацьовувань по горизонталі.

Кількісну інтерпретацію ROC дає показник AUC – площа, обмежена ROC-кривою і віссю частки помилкових позитивних класифікацій. Чим вище показник AUC, тим якісніше класифікатор, при цьому значення 0,5 демонструє непридатність обраного методу класифікації.