Projekt zaliczeniowy

Zarys i założenia

Projekt powinien realizować scenariusz pełnego przepływu eksperymentu machine Learning z wykorzystaniem technik przetwarzania dużych zbiorów danych zaprezentowanych na zajęciach i składać się z następujących etapów:

Etap 1 - pozyskanie danych.

Jako dane można wykorzystać gotowe datasety lub web scrapping, dane muszą być udostępnione na licencji do powszechnego użytku. Nie ma ograniczeń co do typu danych (tabelaryczne, obrazy) oraz ich przeznaczenia (klasyfikacja, regresja). Warto jednak zaopatrzyć się w zbiór, który będzie o rozmiarach odpowiednich do zadań, które realizowane są w trakcie zajęć, np. dane tabelaryczne licznone co najmniej w dziesiątkach milionów rekordów, a dane innego typu (np. obrazy) liczone w setkach megabajtów lub gigabajtach. W trakcie procesu budowania projektu można oczywiście pracować na małym wycinku zbioru w celu przyspieszenia prototypowania, ale docelowo eksperyment należy uruchomić na całym zbiorze. Tutaj również warto użyć formatów zapisu, które są dedykowane dla rozwiązań big data (parquet, avro, orc i inne).

Etap 2 - analiza eksploracyjna i preprocessing danych.

Wstępne przetworzenie danych powinno polegać na zdobyciu wiedzy na temat struktury danych, występowaniu (lub nie) wartości brakujących, odstających (ogólnie analiza eksploracyjna). Ta część powinna być udokuumentowana np. poprzez odpowiedni Jupyter notebook z wypisanymi statystykami, wykresami. Etap preprocessingu polega na oczyszczeniu danych (np. pozbycie się brakujących wartości) wybranie tylko najbardziej przydatnych cech zbioru (np. po wcześniejszej analizie eksploracyjnej i badaniu korelacji), dalszy feature engineering (opcjonalnie). Tak przygotowane dane zapisujemy w postaci docelowego źródła dla naszego modelu.

Etap 3 - trening modelu ML.

W zależności od typu danych i celu modelu (klasyfikacja, regresja, model generatywny) dobieramy odpowiedni algorytm z dostępnej bazy bibliotek dla języka Python. Na tym etapie prowadzący dostarczy kilka przykładowych rozwiązań również w oparciu o bibliotekę Dask oraz Spark. Należy dokonać odpowiedniego podziału na dane treningowe i testowe, zaleca się wykorzystanie mechanizmów walidacji krzyżowej lub bootstrapingu (próby permutacyjne).

Etap 4 - ewaluacja modelu i jego utrwalenie.

W tym etapie należy przedstawić wyniki modelu na danych testowych w formie zestawień, wykresów itp. Również model, który został wytrenowany powinien zostać poddany serializacji (zapisaniu na dysku w formie trwałej, np. z wykorzystaniem modułu pickle lub dedykowawnej metody dostępnej w wybranym algorytmie ML), tak aby możliwe było jego szybkie wponowne czytanie i użycie w procesie wnioskowania (inferencji) bez konieczności ponownego szkolenia.

Całość prac powinna się koncentrować na wykorzystaniu mechanizmów big data na poszczególnych etapach, a nie na jak najlepszych wynikach samego modelu.

Jako rozwiązanie należy dostarczyć kod (notebooki, pliki .py) oraz dane (link, dane skompresowane lub reprezentatywną próbkę, jeżeli zbiór jest bardzo duży). Najwygodniej jeżeli będzie to repozytorium git, ale dopuszczalna jest też inna forma.

Termin: 29.01.2025, ale chciałbym w miarę możliwości chociaż kilka projektów zobaczyć już na zajęciach tydzień wcześniej.