ZPOiF - Konspekt Projektu

Paulina Przybyłek
Jakub Wiśniewski
Dawid Przybyliński

9 grudnia 2019

1 Cel projektu i jego krótka charakterystyka

Projekt ma na celu stworzenie implementacji ***lasu losowego (Random forest)*** w Javie, działającego na zasadzie biblioteki, który będzie możliwy do wykorzystania na ogólnych danych.

Realizowany jest przez grupę trzyosobową a wyniki postępu prac umieszczane są w repozytorium na Githubie. Wybranym środowiskiem jest ***IntelliJ IDEA***. Czas na jego wykonanie kończy się w dniu 27.01.2020. Sposób realizacji projektu jest zgodny z wytycznymi podanymi przez prowadzącego przedmiot a podział obowiązków wewnątrz grupy będzie ustalany dynamicznie, zależnie od funkcji, które akurat będą do zaimplementowania.
Zakładamy, że każdy członek grupy będzie wystarczająco dobrze zaznajomiony z całym projektem, aby mógł on poprawiać błędy i implementować nowe funkcjonalności na własną rękę.

2 Zadania realizowane w ramach projektu

Poniżej wymieniono funkje, kóre zamierzone są w projekcie. Podział jest rozłożony na 10 punktów, które w przybliżeniu powinny stanowić po 10% pracy. Jednak te założenia są ustalane przed tworzeniem tych funkcji, więc możliwym jest, że niektóre z nich mogą być bardziej lub mniej obszerne niż zakładano.

Lista zadań:

1. Klasy do wprowadzania i odczytywania danych do zastosowania przez algorytm.

   Zakładamy, że dane są podawane przez użytkownika w pliku .csv. Przerabiamy je na ramki danych - na kształt listy list.

2. Indeks Giniego/Entropii - algorytmy klasyfikacji. [wybrano indeks Gini'ego]

3. Struktura drzewa. Klasa DecisionTree (klasa wydmuszka do uzupełnienia innymi klasami - m.in. stworzoną klasą Node). [okazało się to punktem bardziej decyzyjnym, pełnym spotkań i wyboru sposobu na budowę drzew]

4. Dokonywanie podziału zmiennych.

5. Metody do budowania drzew.

6. Metoda tworzenia predycji. [początkowo metoda search a potem zdecydowano na wykorzystywanie dominant do obliczania predykcji i tak powstała metoda predict, która zastąpiła wspomniane search]

   Okazało się, że dobre napisanie algorytmu tworzenia drzew zajmuje dużo czasu potrzebnego na zapoznananie się z literaturą o danej tematyce oraz późniejsze opisanie tego w java.

7. Testowanie drzew. Poprawianie i ulepszanie algorytmu.

   W tym momencie powinniśmy zakończyć tworzenie drzew decyzyjnych, dlatego poświecimy więcej czasu na testowanie algorytmu i udoskonalanie go zanim przejdziemy do budowania lasów losowych. [niektóre metody zostały zmienione lub napisane od nowa, inaczej]

8. Podział zbioru na trenujący i testowy. [podział następuje w klasie RandomForest]

9. Struktura lasu losowego. Bagging, "trenowanie" i "testowanie" lasu.

10. Testowanie algorytmu i analiza wyników. Ostatnie szlify nad biblioteką.

Dodatkowo algorytm będzie na bieżąco testowany, aby wyłapać błędy i móc je poprawić przed przejściem do kolejnego punktu.

UWAGA: Wydaje nam się, że projekt może pochłonąć zbyt dużo czasu i pracy nad dopracowaniem wszystkiego - dodatkowo uczymy się pojęcia drzew decyzyjnych, dlatego projekt nie zawiera w zakładanych zadaniach stworzenia intersejsu graficznego.