Eksploracja Danych

Rok akad. 2016/17

Zadanie projektowe

Celem zadania jest wykonanie analizy danych w celu rozwiązania problemu postawionego w jednym z 15 niżej podanych tematów. Zadania te dotyczą problemu klasyfikacji oraz analizy statystycznej pod kątem wyciągnięcia pewnych wniosków lub udowodnienia postawionych tez. Każdy zbiór danych ma w opisie postawiony dla niego problem/problemy, jednakże jeśli któryś zespół jest w stanie zaproponować inny problem do rozwiązania/udowodninia przy pomocy otrzymanych danych to droga jest wolna.

Do rozwiązania otrzymanego zadania należy wykorzystać poznane na labolatorium metody i narzędzia analizy, wizualizacji, grupowania oraz klasyfikacji danych. Rozwiązując postawione problemy należy przede wszystkim skupić się na danych wykonując poszczególne kroki:

1. Opisać postawiony problem.
2. Określić liczbę obiektów, liczbę klas, zakresy zmienności poszczególnych atrybutów, ich wartości statystycznych, poziom wypełnienia kolumn, ilość unikalnych danych itp.
3. Przeanalizować korelację między zmiennymi.
4. Przygotować dane do analizy: Imputować brakujące dane lub usunąć rzadko wypełnione kolumny.
5. Przeanalizować podobieństwo między danymi przy pomocy poznanych algorytmów grupowania, wraz z analizą ilości grup.
6. Dla zadań klasyfkacji należy przetestować wybrane klasyfikatory pod kątem doboru ich parametrów.
7. Ocenić czy do poprawnej klasyfikacji należy wykorzystać wszystkie atrybuty, czy wystarczy ich podzbiór, a może należy stworzyć jakieś nowe dane w oparciu o istniejące?
8. W przypadku zadań z ekspoloracji danych należy przetestować różne możliwe przecięcia oraz zwizualizować i opisać otrzymane wyniki.

Projekt wykonujemy w zespołach dwuosobowych. Każde dane można analizować na wiele sposobów, więc proponuję podzielić się pracą, a później zebrać do raportu końcowego wszystkie wyniki, komenatrze oraz zrozumiale opisany sposób analizy. W przypadku klasyfikacji można w oparciu o wcześniej wyuczone klasyfikatory przez poszczególne osoby wykonać ensembling.

Raport ma zostać dostarczony w **ipython notebook**. Preferowanym językiem jest Python, jednakże jeśli ktoś chce i czuje się na siłach wykonać projekt w języku R to nie widzę problemu. Raport posiadający w sobie skrypt należy wgrać na iSOD najpóźniej do **20 stycznia 2017** roku i napisać maila w celu umówienia się na termin obrony prac.

Zbiory danych:

1. Rozpoznawanie płci na podstawie głosu.
2. Określenie czy dana osoba zarabia więcej niż 50tyś dolarów rocznie.
3. Analiza oparta o dane o aresztowaniach w Baltimore.
4. Rozpoznawanie wyłudzeń na kartach kredytowych.
5. Rozpoznawanie płci właściciela profilu na tweeterze.
6. Rozpoznawanie jednego z sześciu stanów aktywności przy pomocy czujników ze smartfonów.

7. Rozpoznawanie typu Pokemona po jego cechach.
8. Analiza płac w San Francisco.
9. Analiza sentymentu wypowiedzi z tweetów
10. Rozpoznawanie nieprawidłowości w kręgosłupie.
11. Rozpoznawanie końcowej oceny studenta (przy pomocy klasyfikacji) zbiór A.
12. Rozpoznawanie końcowej oceny studenta (przy pomocy klasyfikacji) zbiór B.
13. Analiza jakości zębów w zależności od danych demograficznych/statystycznych (np. konsumpcji cukru).
14. Klasyfikacja możliwej niewypłacalności z kard kredytów.
15. Rozpoznawanie kategorii artykułu po tytule i wydawcy.

Zbiory danych należy pobrać z podanych linków. Numer zadania określony jest poprzez resztę z dzielenia numery zespołu przez 15 (reszta 0 oznacza projekt 15). Każdy zespół ma tydzień na poinformowanie mnie o problemie związanym ze zrozumieniem danych. Po tym terminie reklamacje dotyczące zbioru danych do analizy nie będą przyjmowane.

Dr inż. Grzegorz Sarwas
sarwasg@ee.pw.edu.pl