Eksploracja danych

Piotr Lipiński

Lista zadań nr 5 – Klasyfikacja danych i drzewa decyzyjne

Zadanie 1. *(1 punkt)*
a) Napisz własny klasyfikator K-Nearest Neighbours (KNN). Zadanie można wykonać w Matlabie, Pythonie lub innym języku programowania. W zadaniu można użyć wcześniej zaimplementowanej funkcji liczenia odległości (na przykład z zadania 0 z listy 3).
b) Napisany klasyfikator KNN przetestuj na danych IRIS. Podziel dane losowo na dwie części: 100 wektorów danych użyj jako dane uczące do stworzenia klasyfikatora i 50 wektorów danych użyj jako dane testowe do przetestowania stworzonego klasyfikatora. Powtórz ten eksperyment kilkukrotnie i porównaj wyniki.
c) Przeprowadź też podobny test klasyfikatora KNN na danych Optical Recognition of Handwritten Digits (https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits).
d) Zrób cross validation klasyfikatora KNN na obu zestawach danych: podziel dane na 10 części, kolejno bierz jedną z nich, traktuj ją jako dane testowe, a pozostałe 9 części jako dane uczące, stwórz klasyfikator na danych uczących i przetestuj na danych testowych, odnotuj liczbę błędów, powtórz obliczenia dla kolejnych części danych, policz całkowity błąd klasyfikatora sumując odnotowane liczby błędów.

Zadanie 2. *(1 punkt)*
W pakiecie SciKit do Pythona dostępnych jest kilka popularnych algorytmów klasyfikacji danych przy użyciu drzew klasyfikacyjnych. Zapoznaj się z nimi wykonując skrypt umieszczony w materiałach do wykładu. Do rysowania drzew użyj programu Graphviz.
a) Jaką miarę różnorodności stosuje algorytm konstrukcji drzew klasyfikujących użyty w skrypcie? Dla danych Titanic sporządź drzewa klasyfikacyjne stosując indeks Giniego oraz entropię.
b) Podziel dane Titanic na dane uczące i dane testowe (jak w zadaniu 1b), stwórz drzewa klasyfikacyjne na danych uczących i przetestuj na danych testowych.
c) Spróbuj ograniczyć głębokość drzewa. Zobacz jak wpływa to na wyniki (zarówno na danych uczących jak i na danych testowych).
d) Spróbuj przyciąć drzewo techniką omawianą na wykładzie lub własną. Zobacz jak wpływa to na wyniki (zarówno na danych uczących jak i na danych testowych).
e) Zrób cross validation wszystkich tworzonych w tym zadaniu klasyfikatorów na zestawie danych Titanic.

Zadanie 3. *(2 punkty)*
Zapoznaj się z implementacją algorytmów Random Forest i Extremely Randomized Trees w pakiecie SciKit (http://scikit-learn.org/stable/modules/ensemble.html#forest). Użyj ich do klasyfikacji danych IRIS oraz Titanic. Oceń ich skuteczność dzieląc dane na dwa zestawy lub robiąc cross validation.

Zadanie 4. *(2 punkty)*
Zbiór danych Mushroom (http://archive.ics.uci.edu/ml/datasets/Mushroom) zawiera informacje o grzybach i o tym czy są one jadalne, trujące, nie polecane do jedzenia lub nieznane. Zapoznaj się z tym zbiorem danych i stwórz drzewo decyzyjne określające jadalność grzybów na podstawie ich cech określonych w zestawie danych.

Zadanie 5. *(2 punkty)*
Zbiór danych Car Evaluation (http://archive.ics.uci.edu/ml/datasets/Car+Evaluation) zawiera informacje o samochodach i rekomendacje dotyczącą ich ewentualnego kupna. Zapoznaj się z tym zbiorem danych i stwórz dla niego drzewo decyzyjne.

Zadanie 6. *(2 punkty)*
Zbiór danych Bank Marketing (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) zawiera informacje o klientach banku i ich zainteresowaniu lokatami bankowymi. Zapoznaj się z tym zbiorem danych i na jego podstawie stwórz drzewo decyzyjne pozwalające przewidywać czy klient banku będzie zainteresowany lokatami czy nie na podstawie zgromadzonych o nim informacji.

Zadanie 7. *(2 punkty bonusowe)*
Zapoznaj się ze zbiorem danych Human Activity Recognition Using Smartphones (http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones). Spróbuj stworzyć klasyfikator, który będzie rozpoznawał czynność wykonywaną przez człowieka na podstawie odczytów sensorów określonych w zestawie danych.

UWAGA: W zadaniach 4, 5, 6 i 7 należy zastanowić się także nad preprocessingiem danych: ewentualną normalizacją, standaryzacją, zmianą kodowania danych czy sposobem traktowania wybrakowanych rekordów danych, jeśli takie występują. Należy także rozważyć różne algorytmy tworzenia drzew klasyfikacyjnych (a w zadaniu 7 także innych poznanych klasyfikatorów, niekoniecznie drzew klasyfikacyjnych), różne ich parametry i ostatecznie wybrać klasyfikator najlepszy (uzasadniając swój wybór w oparciu o wyniki przeprowadzonych testów).