Laboratorium

Wykorzystanie technik przetwarzania danych w uczeniu maszynowym

Zadanie 1

Celem zadania jest wybranie większej i ciekawszej bazy danych, odpowiedniej obróbki danych, a następnie przetestowanie, jak działają na niej poznane algorytmy klasyfikujące.

- Baza danych może być **numeryczno-kategoryczna** i testujemy podstawowe klasyfikatory (kNN, Naive Bayes, Drzewo decyzyjne, Sieci neuronowe).
- Alternatywnie, baza danych może być **obrazkowa** i testujemy klasyfikatory radzące sobie na obrazkach (głównie tu chodzi o konwolucyjne sieci neuronowe). Patrz: wykład 6.

Co należy zrobić?

1) **Wybieramy bazę danych do klasyfikacji.** Ważny krok to wybór odpowiedniej bazy danych. Działaliśmy na prostych bazach danych (iris.csv i diabetes.csv), teraz pora wybrać coś bardziej skomplikowanego. Zachęcam do poszukania pod linkiem

   https://www.kaggle.com/datasets

   Można na tej stronie fajnie filtrować, np. zaznaczyć „classification" i pliki csv: https://www.kaggle.com/datasets?fileType=csv&sizeStart=20%2CKB&sizeEnd=50%2CMB&tags=13302-Classification

   Alternatywna strona:

   https://archive-beta.ics.uci.edu/datasets

   Pod powyższymi linkami można też znaleźć bazy danych obrazkowe. Wówczas warto poszukać „image classification".

   Jakie cechy powinna posiadać baza danych?

   - Powinna być odpowiednio duża. Minimum parę tysięcy rekordów, najlepiej powyżej 7 kolumn. Mile widziane jednak są jeszcze większe (parędziesiąt tysięcy kolumn, kilkanaście/kilkadziesiąt kolumn).
   - Powinna być przeznaczona do klasyfikacji. Tzn. łatwo w niej znaleźć kolumnę/zmienną, którą należy odgadywać i jest to kolumna z danymi kategorycznymi (lub numeryczna, którą można zamienić na kategorie).
   - Dobrze będzie, jeśli baza będzie trochę „popsuta" 😊 Jeśli będzie z błędami, brakującymi danymi lub będzie wymagała innych technik preprocessingu to zawsze plus dla rozwiązania.
   - Spróbuj znaleźć bazę, która choć trochę Cię zainteresuje. Tematyka tych datasetów jest bardzo szeroka 😊

   Gdy wybierasz bazę danych obrazkową, wymagania są podobne:

   - Obrazków powinno być dużo. Minimum parę tysięcy, ale lepiej więcej.
   - Jeśli chcesz je klasyfikować to powinny być oznaczone nazwami klas.

- Plusem jest, jeśli baza jest trochę nieobrobiona np. obrazki są różnych rozmiarów i trzeba ją trochę przetworzyć.

2) **Preprocessing bazy danych i przygotowanie dwóch wersji datasetu**. Bazę danych należy odpowiednio przygotować do klasyfikacji.
   - W przypadku baz danych numeryczno-kategorycznych na pewno warto sprawdzić czy są błędy i brakujące dane. Jeśli tak, to usunąć je w sensowny sposób. Należy przeprowadzić inne operacje, które są niezbędne do korzystania z datasetu. Następnie warto zastanowić się nad dalszą obróbką danych (PCA, normalizacja, itp.). Przygotuj dwie wersje bazy danych: jedną mniej przetworzoną, a drugą bardziej. Na obu przetestujesz klasyfikację w dalszej części zadania.
   - W przypadku obrazków, obowiązkowym krokiem wydaje się być dopasowanie rozmiaru zdjęć do jednego formatu. Inne techniki warte rozpatrzenia: normalizacji wartości liczbowych w pikselach, skalowanie obrazków, konwersja do skali szarości, augmentacja danych.
     Przygotuj dwie wersje bazy danych: jedną mniej przetworzoną, a drugą bardziej. Na obu przetestujesz klasyfikację w dalszej części zadania.

3) **Trenujemy i testujemy klasyfikatory na obu wersjach bazy danych.** Dla obu wersji bazy danych trenujemy i testujemy klasyfikatory. Na początku oczywiście dzielimy bazę danych na zbiór testowy i treningowy (i ewentualnie walidacyjny). Następnie trenujemy po kolei klasyfikatory na zbiorze treningowym. Każdy z klasyfikatorów testujemy na zbiorze testowym. Podajemy jego dokładność (accuracy) oraz macierz błędów (najlepiej w formie graficznej), a w przypadku sieci neuronowych również krzywą uczenia się (learning curve) uwzględniająca zbiór treningowy i walidacyjny. Na koniec robimy podsumowanie klasyfikatorów dla obu wersji bazy danych. Który zadziałał najlepiej?

   Dla bazy danych numeryczno-kategorycznej, klasyfikatory do testowania to:
   - Drzewo decyzyjne (w wersji mniejsze z przyciętymi gałęziami i większej).
   - Naiwny Bayes.
   - K-Najbliższych Sąsiadów (dla paru różnych k)
   - Sieć neuronowa (dla paru topologii i być można dla paru konfiguracji uczenia).

   Dla bazy danych obrazkowej:
   - K-Najbliższych Sąsiadów (każdy piksel obrazka to liczba).
   - Sieć neuronowa o sensownej strukturze (każdy piksel obrazka to neuron wejściowy)
   - Konwolucyjna sieć neuronowa (najlepiej z różnymi konfiguracjami lub topologiami).

Alternatywną wersją powyższego zadania jest:

- Wygenerowanie danych (logów) do gry w czołgi (za pomocą grania ręcznego lub naszych botów).
- Wytrenowanie sieci neuronowej na tych danych (w Pythonie). Sieć na podstawie algorytmów zgaduje jakie z 7 klawiszy ma naciskać (0/1). Czyli mamy 7 kolumn z klasą binarną.
- Przepisanie tej sieci neuronowej z Pythona do Javascript, tak aby można było wkleić do naszej gry przeglądarkowej.

- Kompresja logów (np. za pomocą PCA) i powtórzenie powyższego eksperymentu z uczeniem. Czy sieć wytrenowana na skompresowanych logach działa lepiej?

Termin oddania: zajęcia po majówce (9 maja).