Praca domowa nr 2

Klasyfikacja na wybranej bazie danych

Zadanie

Celem pracy domowej nr 2 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 dec, Sieic neuronowe).
- Alternatywnie, baza danych może być **obrazkowa** i testujemy klasyfikatory radzące sobie na obrazkach (głównie tu chodzi o konwolucyjne sieci neuronowe). Patrz: wykład 6.
- Bardzo alternatywnie: rezygnujemy z klasyfikacji obrazków i **generujemy obrazki** za pomocą przeciwstawnej sieci generatywnej (GAN). To dla osób z mocniejszymi komputerami i cierpliwością. 😉 Patrz wykład 7.

Jak w poprzedniej pracy domowej, całe rozwiązanie należy umieścić w notebooku jupyterowym, który w przejrzysty sposób będzie prezentował rozwiązanie. Notebook powinien zawierać wstawki kodu, ich outputy, Twoje komentarze do kodu i wyników. Notebook powinien być sensownie podzielony na rozdziały: wstęp z prezentacją bazy danych, rozdziały o preprocessingu, rozdział o klasyfikacji (może być podzielony na podrozdziały), podsumowanie i bibliografię.

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.

Gdy wybierasz bazę obrazkową do uczenia GANa, obrazki powinny być do siebie w jakiś sposób podobne (z tej samej dziedziny). Ciężko wyuczyć GAN jednocześnie generować obrazki samolotów, ludzi i zwierząt.

2) **Potwierdzenie i rezerwacja wybranego problemu.** W obrębie grupy laboratoryjnej, każdy powinien mieć unikalny temat (inna baza danych). Po wyborze tematu, zarezerwuj go publicznie, tak by inni wiedzieli, że temat jest zajęty. Robimy to, dodając komentarz w konwersacji na Teams, założonej przez prowadzącego zajęcia. Komentarz powinien zawierać nazwę bazy danych (skopiowaną ze strony).
Obowiązuje zasada: kto pierwszy ten lepszy.

3) **Preprocessing bazy danych i przygotowanie dwóch wersji datasetu**. Bazę danych należy odpowiednio przygotować do klasyfikacji. Pomocna może być tu wiedza z wykładu 8 (12.04).
   - 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 bazy obrazków, również warto zajrzeć do wykładu 8. 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.
   - W przypadku pracy nad GANem, również warto rozpatrzyć różnie przetworzone bazy danych obrazków. Zrób wersję datasetu: jedną mniej przetworzoną, a drugą bardziej.

4) **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).

W przypadku generacji obrazków GANem należy rozpatrzyć różne struktury GANa. Warto tu zwłaszcza poeksperymentować z wewnętrzną strukturą generatora i dyskryminatora. Podaj wyniki generowania obrazków dla trzech różnych GANów.

Terminy i ocenianie

Czas na zrobienie zadania (termin oddania) zostanie napisany przez prowadzącego na Teams. Z uwagi na skomplikowanie zadania, zalecany czas to 2 tygodnie czasu na rozwiązanie go.

Jeśli prowadzący stworzy zadanie na Teams, z możliwością załączenia plików, to należy załączyć notebooki Jupyter w podanym terminie. Jeśli to możliwe, to proszę załączyć też wersję HTML lub PDF notebooka.

Praca domowa oceniana jest na **maks 5 punktów**. Oceniane będą takie aspekty jak:

- Czy baza danych jest odpowiednio duża/ciekawa/skomplikowana?
- Czy sprawozdanie jest przejrzyste i kompletne?
- Jak przeprowadzono preprocessing bazy danych?
- Czy przeprowadzono wymagane eksperymenty z klasyfikacją? (lub generowaniem obrazków)

Praca domowa może być oceniona na trzy sposoby. Dla części osób PD może być sprawdzona jednym sposobem, dla części innym.

- Notebook będzie pobrany przez prowadzącego zajęcia i sprawdzony zdalnie/osobiście.
- Prowadzący poprosi o odpalenie notebooka na zajęciach i krótką prezentację pracy domowej.
- Prowadzący poprosi o prezentacje notebooka na projektorze i prezentację rozwiązania przed całą grupą (w przypadku wybranych osób, lub szczególnie ciekawych prac).