Zadanie projektowe nr 2

Porównanie klasyfikatorów na wybranej bazie danych

Ten projekt można zrealizować wg poniższych wytycznych lub wybrać inny temat (patrz na końcu pliku).

Celem projektu jest wybranie bazy danych i zastosowanie technik zgłębiania danych / uczenia maszynowego, ze szczególnym wskazaniem na algorytmy klasyfikujące. Eksperymenty muszą być w czytelny sposób opisane w sprawozdaniu PDF, a projekt (kod) przechowujemy na githubie / bitbucket i podajemy link w sprawozdaniu (można mnie dodać: gmadejsk @ github, grzesiekm @ bitbucket).

Wybór bazy danych

Bazy danych powinny być duże (tysiące, setki tysięcy danych). Im bardziej skomplikowane tym lepsze rokowania na dobrą ocenę. Wiele baz danych można znaleźć na stronach:

a) https://www.kaggle.com/datasets
b) https://archive.ics.uci.edu/ml/datasets.php
c) https://data.world/datasets/data-mining
d) https://datahub.io/search

Proponowane bazy danych do wyboru:

a) Breast Cancer Wisconsin https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original) (łatwa, niższa ocena)
b) Student Performance, https://www.kaggle.com/spscientist/students-performance-in-exams (co uczynić klasą? Baza dość łatwa)
c) Adult Census Income, https://www.kaggle.com/uciml/adult-census-income lub https://archive.ics.uci.edu/ml/datasets/Adult
d) Heart Disease, https://archive.ics.uci.edu/ml/datasets/Heart+Disease (gdy ktoś wybierze, proszę zamienić klasę na dwie wartości 0 = zdrowy, 1 = chory (w stopniu 1,2,3,4).
e) Drug Consumption, https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29 (trzeba by zmienić klasę z kilku wartości, na dwie wartości – potrzebna sensowna obróbka, albo wykonanie kilku eksperymentów z różnymi klasami)
f) Thyroid Disease, https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease (uwaga: sporo brakujących danych)

g) Speed Dating Experiment, https://www.kaggle.com/annavictoria/speed-dating-experiment (co jest klasą?)
h) Telco Churn, https://www.kaggle.com/blastchar/telco-customer-churn
i) COVID-19, diagnoza, https://www.kaggle.com/einsteindata4u/covid19
j) Health Insurance, https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction
k) Red Wine Quality, https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
l) Student Alcohol Consumption, https://www.kaggle.com/uciml/student-alcohol-consumption

Badanie i obróbka bazy danych (wymagane na ocenę 3.0)

W sprawozdaniu należy uwzględnić rozdział wstępny, w którym:

- Opiszesz, jakie dane zawiera wybrana baza danych i jakiej klasyfikacji dokonujemy (która kolumna, jakie wartości).
- Dla każdej z kolumny podasz podstawowe informacje: min, max, średnia i częstość występowania poszczególnych odpowiedzi (np. na wykresie kołowym lub słupkowym). W przypadku kolumn numerycznych można podzielić to na przedziały. Wskażesz też na procent brakujących danych.
- Wyjaśnisz jak baza danych została przygotowana do klasyfikacji. Czy jakieś kolumny zostały zmodyfikowane? Usunięte? Czy wykryto jakieś błędne dane?

Porównanie poznanych klasyfikatorów (wymagane na ocenę 3.0)

Główny rozdział sprawozdania to porównanie skuteczności klasyfikatorów na bazie danych.

- Podziel bazę danych na zbiór testowy i treningowy. Ewaluacji wszystkich klasyfikatorów dokonuj na jednym zbiorze testowym.
- Przetestuj klasyfikatory poznane na zajęciach
    - Naive Bayes
    - Drzewa decyzyjne
    - k Najbliższych sąsiadów (dla wybranego k)
    - Sieci neuronowe (dla wybranej topologii)
- Ewaluacja powinna zawierać dokładność klasyfikatora i macierz błędu.
- Wskaż najlepiej działający klasyfikator.

Rozszerzona klasyfikacja i inne techniki (wymagane na ocenę 4.0 i 5.0)

Tak jak w wymaganiach na ocenę 3, ale wymagania są rozszerzone. Im więcej poniższych podpunktów uwzględnisz, tym wyższa będzie ocena.

- Dodaj kilka innych klasyfikatorów:
    - k Najbliższych sąsiadów (dla kilku wybranych k)
    - Sieci neuronowe (dla kilku wybranych topologii)
    - Support Vector Machines
    - Random Forest
    - Metody typu Ensemble
    - Inne?

    Dla nowych klasyfikatorów zrób krótki wstęp teoretyczny. Wyjaśnij na jakiej zasadzie działają.
- Rozszerzona ewaluacja klasyfikatorów. Jak inaczej można oceniać klasyfikatory? Które miary będą miały sens w Twoich badaniach? Na początek można rzucić okiem na: https://en.wikipedia.org/wiki/Sensitivity_and_specificity i poszukać innych źródeł rozwijających temat.
- Szukanie reguł asocjacyjnych. Czy ma sens? Jakich najlepiej szukać? Podaj te dla nas szczególnie interesujące.
- Porównanie skuteczności klasyfikatorów na jakichś wykresach. Słupkowy? ROC?
- Czy w naszej bazie danych jakieś rodzaje błędów są ważniejsze / poważniejsze niż inne? Dlaczego?

Prezentacja i ocenianie

- Termin oddania na portalu edukacyjnym:
    - **18.01.2021, godz. 23:59** (dla studiów dziennych, osoby wybierające projekt „typu standard")
    - **25.01.2021, godz. 23:59** (dla studiów dziennych, osoby wybierające projekt z części „inne typy projektów")
    - **16.01.2021, godz. 23:59** (dla studiów zaocznych, osoby wybierające projekt „typu standard")
    - **23.01.2021, godz. 23:59** (dla studiów zaocznych, osoby wybierające projekt z części „inne typy projektów")

- Na portalu załączamy jedynie sprawozdanie (link do kodu źródłowego na repozytorium powinien być w sprawozdaniu).
- Ocena wystawiana jest głównie w oparciu o zakres materiału pracy (czy spełnione są wymagania na ocenę 3.0, 4.0, 5.0?).
- Sprawozdanie w formie PDF powinno być czytelne, bo posłuży też jako prezentacja na zajęciach. Można je zrobić jako dokument ze stronami, lub prezentacja ze slajdami. Prezentacje będą odbywały się na ostatnich i przedostatnich zajęciach.

Inne typy projektów

Zamiast powyższego projektu dotyczącego klasycznej klasyfikacji dla baz danych, można wybrać jeden z poniższych projektów, a nawet zaproponować własny.

Pomysł 1: Analiza tekstów

Zadanie ma na celu przetestowanie technik analizy tekstów, ze szczególnym naciskiem na analizę opinii (sentyment analysis) i ze szczególnym naciskiem na bazy pobierane z twittera. Przykładowe tematy:

a) **Andrzej Duda i Rafał Trzaskowski**. Jak zmieniały się opinie na temat dwóch kandydatów na prezydenta w okresie kwiecień-lipiec 2020? Pobierz bazę danych tweetów z tego okresu, oceń tweety jako pozytywne/neutralne/negatywne i sprawdź, który z kandydatów miał większy procent pozytywnych opinii w poszczególnych tygodniach tego okresu. Jak zmieniała się liczba tweetów ogółem, pozytywnych, negatywnych dla obu kandydatów na przestrzeni miesięcy? Baza danych musi mieć przynajmniej parę tysięcy tweetów. Czy w badaniach można uwzględnić retweety i serduszka pod postem?

b) **Szczepionka na COVID-19**. Wyszukaj tweety z odpowiednimi hasztagami (#szczepionka #covid19 #koronawirus lub angielskie #vaccine #covid19 #coronavirus, i tym podobne). Sprawdź jak zmieniały się emocje związane ze szczepionką na przestrzeni ostatnich kilku miesięcy. Możesz wziąć pod uwagę różne emocje np.: strach/złość/radość/smutek lub wprowadzić prostszy podział pozytywne/negatywne. Policz tweety z tymi emocjami w kolejnych tygodniach (podobnie jak w podpunkcie (a)). Czy były jakieś szczególne wahania? Z czego wynikają? Czy w badaniach można uwzględnić retweety i serduszka pod postem?

c) **Donald Trump i Joe Biden**. Podobnie jak w podpunkcie (a) ale dla amerykańskich kandydatów na prezydenta. Język tweetów: angielski. Dla tego zadania istnieje właściwie gotowa baza danych: https://www.kaggle.com/manchunhui/us-election-2020-tweets

d) **Aborcja i TK**. W okresie jesieni 2020, które strony były bardziej aktywne na twitterze: zwolennicy zaostrzania prawa aborcyjnego czy ich przeciwnicy (protestujący)? Co można powiedzieć o tweetach obu stron (pozytywnego/negatywne)? Czy któraś ze stron posługiwała się częściej tzw. mową nienawiści? Przeprowadź analizę.
e) **Inne tematy wywołujące emocje**: do wyboru.
f) **Klasyfikacja tweetów ze względu na ocenę**. Należy pobrać bazę danych tweetów https://www.kaggle.com/kazanova/sentiment140 ktore już zawierają ocenę. Następnie zaproponować klasyfikatory, który będą oceniały tweety jako pozytywne, neutralne, negatywne i ewaluować je. Można się podjąć tez innego zadania niż klasyfikacja: sprawdzić jakie słowa zawierają tweety pozytywne a jakie negatywne. Zrobić listę najczęstszych słów w tweetach pozytywnych i negatywnych. Inspiracji do rozwiązania zadania można też szukać w zakładce „Notebooks" (https://www.kaggle.com/kazanova/sentiment140/notebooks)
g) **Klasyfikacja tweetów o zmianach klimatycznych**: https://www.kaggle.com/edqian/twitter-climate-change-sentiment-dataset (podobnie jak w (f)).
h) **Klasyfikacja tweetów z recenzjami wina**. https://www.kaggle.com/zynicide/wine-reviews (podobnie jak w (f)).
i) **Klasyfikacja kickstarterów**. https://www.kaggle.com/kemical/kickstarter-projects To zadanie jest z pogranicza zwykłej klasyfikacji i analizy tekstów. Tesktowe są wyłącznie tytuły. Czy da się wykorzystać tytuły do ciekawej analizy? (np. odpowiedzi na pytanie: jakie słowa pojawiają się najczęściej w udanych Kickstarterach, a jakie w nieudanych?)

(można poszukać innych baz danych: https://www.kaggle.com/datasets?search=twitter – najlepiej ocenionych złotym medalem 🥇)

Co zawrzeć w rozwiązaniu zadania?

1. Tworzenie/opisanie bazy danych.
   Jeśli tworzysz własną bazę danych, powiedz o sposobie w jaki to zrobiłeś(aś). Czy wykorzystane było twitter API, czy narzędzia do automatycznego sczytywania ze stron (Selenium?). Ile tweetów udało się pobrać? W Jakim języku pracujemy? Jeśli korzystasz z gotowej bazy danych, omów krótko jakie informacje są w niej zawarte. Własnoręcznie tworzona baza danych i język polski zawyżają ocenę z projektu (w stronę 5.0), pobrana ze strony baza danych i język angielski trochę zaniżają (w stronę 4.0-4.5).
2. Zastosowanie rożnych technik obróbki danych.
   Należy zastanowić się jak wydobyć najważniejsze informacje z tekstu. Przydatne techniki to tokenizacja, lemmatyzacja, skorzystanie z listy stopwords (można znaleźć je w internecie lub przygotować własną listę). W wydobywaniu emocji/opinii przydatne będą listy ze słowami pozytywnymi/negatywnymi/złości/smutku/radości/ itd. Im ciekawsze techniki tym większa szansa na ocenę 5.0.

3. Poszukiwanie odpowiedzi na pytania w projekcie.
   Jeśli analizowaliśmy tweety na przestrzeni czasu: zróbmy wykres porównujący ich liczbę / procentowy odsetek w kolejnych okresach czasu. Można też stworzyć tzw. chmury tagów (np. osobną dla tweetów pozytywnych i osobną dla negatywnych). Warto przeanalizować też nagłe skoki emocji w czasie – jakie wydarzenia mogły wpłynąć na takie skoki? Jeśli mowa o klasyfikacji, to przetestuj przynajmniej parę klasyfikatorów – najlepiej takich, które nadają się do przefiltrowanych danych tekstowych.

Pomysł 2: Przetwarzanie obrazów

Dostajemy bazę danych obrazów / zdjęć. Naszym celem jest wytrenowanie algorytmów rozpoznających, co jest na zdjęciach i klasyfikujących je według kilku etykiet.
Przykładowe bazy danych:

a) Obiekty geograficzne: https://www.kaggle.com/puneet6060/intel-image-classification
b) Zapalenie płuc – rentgen https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (pewnie trudniejsze)
c) Krwinki i choroby https://www.kaggle.com/paultimothymooney/blood-cells
d) Jedzenie https://www.kaggle.com/kmader/food41
e) Małpy https://www.kaggle.com/slothkong/10-monkey-species
f) Litery w języku migowym https://www.kaggle.com/datamunge/sign-language-mnist
g) Psy i koty https://www.kaggle.com/jessicali9530/stanford-dogs-dataset
h) Simpsonowie https://www.kaggle.com/alexattia/the-simpsons-characters-dataset
i) Owoce https://www.kaggle.com/moltean/fruits

I wiele innych. Można wygooglać „image dataset classification” - może coś ciekawego się trafi.

Projekt ten w zasadzie nie różni się bardzo od projektu standardowego. Również należy przetestować kilka klasyfikatorów na zbiorze danych. Warto spróbować zastosować jakiś podstawowy klasyfikator np. kNN i porównać jego działanie z sieciami głębokimi, np. kilka wersji sieci konwolucyjnej o różnie ustalonych parametrach.

Problemem może być wstępne przygotowanie obrazków do klasyfikacji. Jeśli zdjęcia są duże, być może trzeba nanieść odpowiednie filtry lub zmniejszyć rozmiar.

Pomysł 3: Analiza szeregów czasowych

Analizę szeregów czasowych (time series analysis) przeprowadza się, by symulować i przewidywać pewne zjawiska występujące w czasie (prognoza pogody, giełda, ceny nieruchomości, itp.).

Celem projektu może być porównanie podstawowych technik do prognozy zachowania szeregów czasowych np. ARIMA z technikami z gatunku uczenia głębokiego np. sieci LSTM (w różnych wersjach).

Bazę danych można wygooglać wpisując „time series dataset". Przykłady:

a) Giełda https://www.kaggle.com/szrlee/stock-time-series-20050101-to-20171231
b) COVID-19 https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
c) Przewidywanie wyniku meczu piłku nożnej.
   https://www.kaggle.com/hugomathien/soccer
   Tutaj właściwie mamy taki projekt hybrydowy, w którym jest trochę klasyfikacji, trochę badania szeregów czasowych. Dane są rozproszone i trzeba je sprytnie połączyć. Właściwie można spróbować napisać narzędzie, które będzie przewidywało jaki będzie wynik meczu, jak skomponujemy dwie drużyny z zawodników.
d) Globalna temperatura https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data (można zbadać jak się zmieniała i jak się będzie zmieniać w różnych miejscach na Ziemi)

I inne: https://www.kaggle.com/datasets?search=time+series&sizeStart=2%2CMB

Co może zawierać rozwiązanie:

1. Narysowanie szeregów czasowych w formie wykresu liniowego (dla dostępnych danych). Można dodatkowo przeanalizować i zinterpretować jego przebieg. Skąd się biorą wahnięcia w górę i dół?
2. Jeśli baza danych oferuje wiele plików z osobnymi szeregami czasowymi, to być może warto przeanalizować każdy z osobna? Lub połączyć w jedną bazę danych. Co ma sens?
3. Wykorzystanie modelu ARIMA (i innych podobnych) oraz technik uczenia głębokiego do szeregów czasowych (sieci LSTM?), do przewidywania dalszego przebiegu szeregu czasowego.
4. Projekt można rozwinąć o dodatkowe badania, które wydadzą się wam interesujące (mapy? animacje?).