Zadanie domowe nr 2
Zgłębianie danych

Rzeczy organizacyjne

Maks. punktów: **9 pkt** (spóźnienie o tydzień = -2 pkt)

Termin oddania: **9 grudnia**

Język programowania: **dowolny** (choć R i Python się do tego nadają chyba lepiej niż inne),

Zasady: dozwolone jest korzystanie z paczek z algorytmami ogólnego zastosowania do wykonania własnych eksperymentów na bazie danych, niedozwolone jest ściąganie już gotowych rozwiązań / wykresów czy ściąganie od siebie nawzajem. Osoby pracujące na tej samej bazie danych będą brane u mnie pod lupę.

Ocenianie: Krótko opowiedzieć na zajęciach, a przed zajęciami wysłać wiadomość na mój email o treści:

PDZ-2-datamin
Jan Kowalski, nr indeksu: 123456
http://tutaj-link-do-rozwiazanej-pracy-domowej.com
(zamiast linka może być załącznik jeśli jest mały, lub udostepnienie mi na Githubie: gmadejsk)

Opis zadania

Celem pracy domowej jest przećwiczenie poznanych na laboratoriach technik zgłębiania danych na wybranym zbiorze danych (najlepiej z kolumną „class” z dwoma możliwymi wartościami).

Wybór bazy danych

Bardzo fajnymi źródłami do wyboru baz danych jest strona Kaggle (https://www.kaggle.com/datasets) oraz UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html). Mają Państwo wolną rękę w wyborze bazy danych, ale proszę zwrócić uwagę na kilka rzeczy:

- Proszę nie wybierać baz danych z obrazkami, długimi tekstami, tweetami lub dźwiękami (to będzie temat innego projektu). Naszym zadaniem jest analiza danych numerycznych (liczby integer, float, binarne) lub kategorialnych (kilka stringów do wyboru). Minimalna liczba kolumn: 6.

- Baza danych powinna mieć łatwo wybieralną klasę. U irysów był to gatunek. W bazach danych medycznych pewnie to pewnie postawienie diagnozy („zdrowy”/”chory”). Niektóre bazy danych nie mają naturalnie wyodrębnionej klasy i takich raczej proszę unikać.
- Proszę wybrać taką bazę danych, której kolumna z klasą ma jedynie dwie wartości „tak”/”nie”, „zdrowy”/”chory”, „mały”/”duży” itp. (irysy byłyby złe bo mają trzy wartości: virginica, setosa, versicolor).
- Baza bazie nierówna. Niektóre są małe lub przejrzyste lub dobrze wypełnione, natomiast inne duże, niejasne i z mnóstwem błędów. Jeśli wybiorą Państwo łatwiejszą bazę danych to projekt pewnie zrobią Państwo znacznie szybciej, ale stracą Państwo 1 punkt za obróbkę i rozgryzanie czy uzupełnianie bazy danych.
- Proszę raczej unikać baz danych z tzw. „time series” czyli zmieniająca się na przestrzeni czasu sytuacją np. badanie indeksów giełdowych czy pogody. To wymaga pracy z innymi technikami i algorytmami.

Poniżej zamieszczam kilka moich propozycji. Zachęcam jednak Państwa do własnych poszukiwań. Może znajdą Państwo coś bardziej interesującego.

Proponowane bazy danych do wyboru (sporo tutaj medycznych propozycji)

a) Pima Indians Diabetes, dostępny na wielu stronach np. tutaj https://www.kaggle.com/uciml/pima-indians-diabetes-database (łatwa)
b) Breast Cancer Wisconsin , https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original) (łatwa)
c) Adult Census Income, https://www.kaggle.com/uciml/adult-census-income lub https://archive.ics.uci.edu/ml/datasets/Adult (trochę trudniejsza)
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). (chyba trochę trudniejsza)
e) Drug Consumption, https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29 (trzeba by zminić klasę z kilku wartości, na dwie wartości – potrzebna sensowna obróbka, albo wykonanie kilku eksperymentów z różnymi klasami) (trudniejsze)
f) Thyroid Disease, https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease (trudniejsze, sporo brakujących danych)
g) Speed Dating Experiment, https://www.kaggle.com/annavictoria/speed-dating-experiment (trudniejsze, co jest klasą?)
h) Suicide in 21st Century, https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/ (może warto połączyć z danymi państw z innych baz danych, żeby mieć szersze pojęcie o warunkach panujących w danym państwie, np. https://www.kaggle.com/fernandol/countries-of-the-world, https://www.kaggle.com/gsutters/economic-freedom - wówczas trudniejsze).
i) Telco Churn, https://www.kaggle.com/blastchar/telco-customer-churn, (średnie)

Wykonanie eksperymentów i sprawozdania

W projekcie należy wykonać kilka zadań. Każde z nich w miarę starannie opisać w sprawozdaniu.

1. Wstęp: objaśnienie wybranej przez Państwa bazy danych. Jakie dane zawiera? Co oznaczają nagłówki kolumn? Która kolumna nadaje się wg Państwa na klasę? Co będzie celem Państwa projektu – odpowiedzi na jakie pytania Państwo szukają?

2. Przetwarzanie / obróbka / łączenie / dzielenie baz danych
   Baza danych może mieć błędy, brakujące wartości, dane ze złego zakresu (np. kolumna z klasą ma za dużo wartości). Być może złączymy dwie bazy w jedną, albo z jednej dużej musimy pozbyć się części niepotrzebnych danych. Wszystkie te kroki należy wykonać uzasadniając swoje postepowanie w sprawozdaniu i dodając wstawki z kodem lub komendami, które do tego służyły.

3. Klasyfikatory i ich ewaluacja
   Baza danych powinna mieć kolumnę klasy (najlepiej z dwoma odpowiedziami np. „tak”/”nie” lub „chory”/”zdrowy” itp.).
   a. Przed przystąpieniem do klasyfikacji należy podzielić bazę danych na zbiór treningowy (na którym trenujemy model) oraz zbiór testowy (na którym ewaluujemy każdy klasyfikator). Proszę w sprawozdaniu zawrzeć komendy lub fragmenty kodu to robiące.
   b. Następnie testujemy jak z naszą bazą danych poradzą sobie poznane na zajęciach klasyfikatory: C4.5/ID3 (drzewo), Naive Bayes, kNN oraz przynajmniej jeden inny wybrany klasyfikator nie poznany na zajęciach (dla każdego z nich proszę w sprawozdaniu podać komendy lub fragmenty kodu uruchamiające). Dla każdego z nich należy też podać macierz błędu (confusion matrix) oraz dokładność (accuracy). Dokładności należy zestawić na wykresie słupkowym.
   c. Dodatkowo, należy zapoznać się jak dokładniej ewaluować klasyfikatory wykorzystując macierz błędu. Linki:
      https://en.wikipedia.org/wiki/Sensitivity_and_specificity (ang.)
      https://pl.wikipedia.org/wiki/Tablica_pomy%C5%82ek (pl.)
      Następnie proszę ująć w sprawozdaniu:
      - Odpowiedź na pytanie: Co oznaczają u mnie wartości TP, FP, TN, FN?
      - Dla każdego klasyfikatora obliczyć TPR (recall, sensitivity) i FPR (fall-out, false alarm).
      - Rozszyfrować w jakiej zależności od TPR i FPR są miary FNR i TNR. Podać te wzory w sprawozdaniu.
      - Udzielić odpowiedzi czym jest u Państwa błąd pierwszego i drugiego rodzaju. Jak mają się oba rodzaje błędów do TPR, FPR, TNR, FNR. Im więcej błędów pierwszego rodzaju tym większe jest co? Co z błędami drugiego rodzaju?

- Pytanie filozoficzne, na które również proszę udzielić odpowiedzi: który z błędów w Państwa bazie jest gorszy do popełnienia: pierwszego czy drugiego rodzaju? Odpowiedź uzasadnić.
- Dla każdego z czterech klasyfikatorów obliczyć parę (FPR,TPR) i zaznaczyć jako punkt na wykresie (wykres typ „ROC Space" tzn. Oś X będzie odpowiadała False Positive Rate, a oś Y True Positive Rate). Następnie udzielić odpowiedzi na pytania: gdzie leżałby punkt dla idealnego klasyfikatora? Który z klasyfikator jest najbliżej idealnego? Który klasyfikator popełnia najmniej błędów gorszego, wg Państwa, rodzaju (patrz poprzedni punkt) i jak to się ma do jego pozycji na wykresie?

4. Grupowanie metodą k-średnich
   Grupujemy wszystkie rekordy na 2, 3 i być może 4 klastry. Następnie przyglądamy się jakie rekordy siedzą w danych klastrach. Czy każdy klaster zawiera w sobie inne rekordy? Proszę opisać w kilkunastu zdaniach Państwa wyniki. Czy przy podziale na dwa klastry podział oddzielił jedną klasę od drugiej?
   Do sprawozdania dołączamy też fragmenty kodu lub komendy, które dokonały grupowania.

5. Reguły asocjacyjne.
   Proszę znaleźć reguły asocjacyjne w Państwa zbiorze, a następnie wypisać najciekawsze z nich (chętnie zawierające tez kolumnę klasa). Czy były jakieś zaskakujące reguły?
   Do sprawozdania dołączamy też fragmenty kodu lub komendy, które szukały reguł asocjacyjnych.

6. Być może zechcą Państwo wykorzystać jeszcze inne metody badania danych, nieuwzględnione w poprzednich punktach (np. wykresy słupkowe i kołowe, badające jak klasa zmienia się od płci czy grupy wiekowej itp). Jest to nieobowiązkowe, ale mile widziane.

7. Podsumowanie: czy udało się znaleźć odpowiedzi na nurtujące nas pytania? Jakie ze stosowanych technik były najciekawsze lub najskuteczniejsze w badaniu problemu? Czy było coś zaskakującego w wynikach?