Eksploracja danych

Projekt zaliczeniowy

Dysponujesz zbiorem census-income (w formacie ARFF i CSV). W pliku **`census-income.names`** znajdują się podstawowe statystyki dotyczące pliku oraz wyjaśnienie znaczenia poszczególnych atrybutów. Zbiór danych pochodzi z amerykańskiego spisu powszechnego i zawiera szczegółowe informacje na temat obywateli. Atrybutem decyzyjnym jest atrybut `CLASS` (problem klasyfikacji binarnej) wskazujący, czy dana osoba przekroczyła próg zarobkowy $50 000 rocznie. Zanim zaczniecie Państwo analizę, proszę usunąć ze zbioru danych atrybut `MARSUPWT` reprezentujący wagę instancji: wszystkie rekordy będziemy traktować tak samo.

Poniżej umieściłem listę zadań szczegółowych. Możecie Państwo wykonać wybraną część zadań w zależności od tego, jaką chcecie uzyskać ocenę.

Kryteria oceny są następujące:

- **15 pkt.**: rozwiązanie 3 zadań
- **20 pkt.**: rozwiązanie 5 zadań
- **25 pkt.**: rozwiązanie 6 zadań
- **30 pkt.**: rozwiązanie 7-8 zadań

Zadanie można wykonać za pomocą dowolnego narzędzia. Do pracy z Weką lub RapidMinerem wygodniejszy będzie plik **`census-income.arff`**, podczas gdy do pracy z Pythonem wygodniejszy będzie plik **`census-income.csv`**. Sprawozdanie musi zawierać krótki opis wykonanych czynności i dokładne odpowiedzi na wszystkie pytania umieszczone pod każdym zadaniem (sekcja wypunktowana). Opcjonalnie, do sprawozdania można dołączyć zrzuty ekranów z przeprowadzonych ćwiczeń. Sprawozdanie może mieć postać notatnika Jupyter, dokumentu *.pdf, notatnika Rmarkdown.

Termin nadsyłania prac upływa 30.06.2019 (niedziela). Projekty proszę przesyłać na adres Mikolaj.Morzy@put.poznan.pl

Zadania szczegółowe

1. Obejrzyj histogramy dla wszystkich atrybutów, na podstawie wartości średniej i zakresu wartości oceń, dla których atrybutów należy zidentyfikować osobliwości. Przeprowadź usuwanie osobliwości.
- Które atrybuty wybrałaś(eś) do usuwania osobliwości?
- Jaką metodę oznaczania osobliwości wybrałaś(eś) dla każdego atrybutu? Dlaczego?

2. W bazie danych brakujące dane są oznaczone za pomocą znaku ‘?’. Znajdź atrybuty zawierające brakujące dane. Tam, gdzie to możliwe, zamień brakujące dane na wartość „Not in universe”, w przeciwnym przypadku zamień brakujące dane na dominującą wartość atrybutu.
- Które atrybuty zawierają brakujące wartości?
- Jeśli nie można było zamienić brakującej wartości na „Not in universe”, to jaka była dominująca wartość dla danego atrybutu?

3. Wybierz atrybuty numeryczne, które powinny być Twoim zdaniem znormalizowane. Przeprowadź normalizację atrybutów numerycznych.
- Które atrybuty numeryczne wybrałaś(eś) do normalizacji?
- Jaką metodą znormalizowałaś(eś) każdy z atrybutów? Dlaczego?

4. Wybierz atrybuty numeryczne, które powinny Twoim zdaniem podlegać dyskretyzacji. Dla każdego atrybutu wybierz najwłaściwszą Twoim zdaniem metodę dyskretyzacji i przedziały dyskretyzacji.
- Które atrybuty numeryczne wybrałaś(eś) do dyskretyzacji?
- Jaką metodę, liczbę przedziałów i granice przedziałów wybrałaś(eś) dla każdego atrybutu? Uzasadnij swój wybór.

5. Dokonaj dyskretyzacji atrybutu kategorycznego `AMARITL` (marital status) na dwie kategorie: osoby zamężne/żonate (3 wartości) i pozostałe (4 wartości)

6. Określ ważność atrybutów względem atrybutu decyzyjnego `INCOME`.
- Podaj trzy najbardziej przydatne atrybuty do przewidywania wartości atrybutu decyzyjnego. Wytłumacz uzyskany wynik.
- Podaj trzy najmniej przydatne atrybuty do przewidywania wartości atrybutu decyzyjnego. Wytłumacz uzyskany wynik.

7. Zbuduj naiwny klasyfikator Bayesa służący do przewidywania wartości atrybutu decyzyjnego `INCOME` na podstawie wartości pozostałych atrybutów. Wybierz atrybuty, które powinny być włączone do modelu (pozostawienie wszystkich atrybutów skutkuje zbudowaniem modelu niskiej jakości). Jako preferowaną wartość (positive) wybierz „50 000 +”. Zwróć uwagę, żeby przede wszystkim poprawnie przewidywać preferowaną wartość kosztem ogólnej dokładności modelu (przy sztywnym przewidywaniu klasy „- 50 000” dokładność modelu wyniesie i tak 93%).
- Które atrybuty weszły w skład modelu?
- Jak wygląda najlepsza znaleziona przez Ciebie macierz kosztów?
- Jak wyglądała macierz pomyłek?
- Jaką metodę wybrała(e)ś do weryfikacji klasyfikatora?

8. Powiązania między krajem urodzenia ojca, matki i dziecka (atrybuty `PEFNTVTY`, `PEMNTVTY`, `PENATVTY`, odpowiednio) można przedstawić w postaci reguł asocjacyjnych. Stwórz model reprezentujący powiązania między tymi danymi, przykładowa reguła:

   `MOTHER=Mexico ∩ FATHER=INDIA → SELF=CANADA`.

   Podaj 5 odkrytych reguł o najwyższym wsparciu i ufności, pomijając te reguły, w których kraj urodzenia dziecka jest taki sam jak kraj urodzenia ojca i matki (tj. pomiń sytuacje w których dziecko urodziło się w tym samym kraju co oboje rodzice).