Eksploracja danych

Projekt zaliczeniowy

W schemacie MIKOLAJ znajdują się tabele CENSUS i CENSUS_TEST. Tabela CENSUS zawiera dane statystyczne potrzebne do trenowania i uczenia klasyfikatorów. Tabela CENSUS_TEST służy do testowania jakości zbudowanych modeli klasyfikacji. Przekopiuj zawartość obu tablic do własnego schematu. Atrybutem decyzyjnym jest atrybut INCOME (problem klasyfikacji binarnej).

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

- Ocena **dostateczna**: zadania 1, 4, 6
- Ocena **dostateczna+**: zadania 1, 4, 5, 6
- Ocena **dobra**: zadania 1, 2, 3, 4, 5, 6
- Ocena **dobra+**: zadania 1, 2, 3, 4, 5, 6, 7, 8
- Ocena **bardzo dobra**: zadania 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

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. Tam gdzie jest to wymagane, proszę także dołączyć kod programu. Termin nadsyłania prac upływa 27.01.2012 (niedziela) o północy (decyduje godzina przybycia listu do mojej skrzynki pocztowej). Żadne wpisy nie będą antydatowane, uzyskanie wpisu za projekt dostarczony po terminie jest możliwe tylko i wyłącznie w oparciu o aktualne przedłużenie sesji. Proszę się również zapoznać z polityką walki z plagiatami umieszczoną na końcu opisu projektu. 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)
    - Podaj kod perspektywy która umożliwia taką operację
6. Określ ważność atrybutów względem atrybutu decyzyjnego INCOME. W analizie pomiń atrybut MARSUPWT (instance weight) określający względną wagę instancji reprezentowanej przez dany wiersz.
    - 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. Wykorzystaj algorytm k-Means do znalezienia najbardziej charakterystycznych skupień cech. Dobierz eksperymentalnie wartość parametru *k* w taki sposób, aby instancje rozkładały się w miarę równomiernie pośród znalezionych skupień.
    - Podaj wybraną przez siebie wartość parametru *k* i rozkład instancji w ramach skupień.
    - Wybierz jedno skupienie i starannie je przeanalizuj (obejrzyj histogramy rozkładów wartości poszczególnych atrybutów w skupieniu). Opisz językiem naturalnym instancje przypisane do danego skupienia. Jaka, Twoim zdaniem, grupa/warstwa społeczna jest opisana za pomocą wybranego skupienia?
8. 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%). Pamiętaj, żeby zbudowane modele testować z wykorzystaniem tabeli CENSUS_TEST!
    - Które atrybuty weszły w skład modelu?
    - Jak wygląda najlepsza znaleziona przez Ciebie macierz kosztów?

- Jakie parametry *singleton_threshold* i *pairwise_threshold* zostały wykorzystane przy budowie modelu?
- Jak wyglądała macierz pomyłek?

9. Zbuduj drzewo decyzyjne służące do przewidywania wartości atrybutu decyzyjnego INCOME. Jako preferowaną klasę wybierz „50 000+”. Dobierz eksperymentalnie optymalne parametry klasyfikatora (głębokość drzewa, miarę jednorodności, rozmiary węzłów wewnętrznych i liści). Pamiętaj, żeby zbudowane modele testować z wykorzystaniem tabeli CENSUS_TEST. Najważniejszym kryterium jest poprawność przewidywania preferowanej klasy a nie ogólna poprawność klasyfikatora.
   - Które atrybuty weszły w skład modelu?
   - Jak wygląda najlepsza znaleziona przez Ciebie macierz kosztów?
   - Jakie wartości parametrów zostały wykorzystane przy budowie modelu?
   - Jak wyglądała macierz pomyłek?
10. 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 $\wedge$ FATHER=INDIA $\Rightarrow$ SELF=CANADA.
    - Podaj kod polecenia SQL tworzącego perspektywę potrzebną do znalezienia reguł asocjacyjnych o podanym formacie
    - Podaj 5 odkrytych reguł o najwyższym wsparciu i ufności
    - Napisz program w PL/SQL odkrywający i wyświetlający powyższe reguły (prześlij kod programu dołączony w osobnym pliku o nazwie **infxxxxx.sql**)

UWAGA!!! W przypadku wykrycia plagiatu **obie** osoby otrzymują ocenę 2.0. Zwracam uwagę, że kwestia rzeczywistego autorstwa projektu nie ma dla mnie najmniejszego znaczenia – oddanie cudzego projektu jako własnego traktuję jak oszustwo, a udostępnienie swojego projektu jak współudział w oszustwie. Wymagania dotyczące uzyskania oceny dostatecznej są tak proste, że nie ma potrzeby kraść cudzych projektów.