Inteligencja obliczeniowa

Laboratorium 6: Zadanie klasyfikacji – Naive Bayes. Drzewa decyzyjne.

Na poprzednich laboratoriach poznaliśmy algorytm k-najbliższych sąsiadów, który potrafi odgadnąć przynależnośc do klasy na podstawie kilku najbardziej podobnych rekordów w tabeli. Na tych laborotriach poznamy inne klasyfikatory: naiveBayes (opiera się na twierdzeniu Bayesa) oraz drzewa decyzyjne.

Zadanie 1 *(1 pkt)*
NaiveBayes to algorytm wykorzystujący prawdopodbieństwo warunkowe

Załóżmy, że mamy małą bazę danych osób, które decydują się (lub nie) na kupno komputera. Parametry tych osób to wiek, dochód, bycie studentem, zdolność kredytowa. Klasa „buys" odpowiada na pytanie: „czy osoba kupuje komputer?".

age	income	student	credit.rating	buys
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	high	yes	excellent	yes
>40	low	yes	excellent	no
31..40	low	no	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	no

Pojawia się nowa osoba (nazwijsmy rekord X), której klasa jest niewiadoma. Jak ją obliczyć?

>40	medium	no	excellent	???

Krok.1 Obliczamy prawodopodobieństwo obu klas:
P(buys=yes)=4/7 P(buys=no)=3/7

Krok.2 Obliczamy prawdopodobieństwa warunkowe biorąc pod uwagę dane z niewiadomego rekordu.
P(age>40|buys=yes)=2/4 (wśród osób kupujących komputer liczymy osoby starsze niż 40)
P(age>40|buys=no)=1/3 (jest jedna osoba 40+ wśród 3 osób niekupujących komputera)
P(income=medium|buys=yes)=1/4
P(income=medium|buys=no)=1/3
P(student=no|buys=yes)=3/4
P(student=no|buys=no)=1/3
P(credit.rating=excellent|buys=yes)=2/4
P(credit.rating=excellent|buys=no)=1/3

Krok.3 Mnożymy prawdopodobieństwa warunkowe dla każdej klasy z osobna, otrzymujemy prawdodpodobieństwo apriori (zakładamy prawdziwość hipotezy i patrzymy na obserwacje):
P(X|buys=yes) = (2/4) * (1/4) * (3/4) * (2/4) = 3/64
P(X|buys=no) = (1/3) * (1/3) * (1/3) * (1/3) = 1/81

Krok.4 Obliczamy ze wzoru Bayesa prawdopodobienstwo aposteriori (mamy obserwacje wysnuwamy hipotezę)
P(buys=yes|X)=P(X|buys=yes)*P(buys=yes) = (3/64)*(4/7) = 0.0267857
P(buys=no|X)=P(X|buys=no)*P(buys=no) =(1/81)*(3/7) = 0.00529

Z obu wartości większa jest 0.0267857, więc nasz rekord przyjmuje klasę yes.

Mamy kolejną osobę Y, przyporzadkuj jej klasę obliczając z pomocą kalkulatora Google, Excela lub R wartości z czterech kroków:

>40	low	no	fair	???

Zadanie 2 *(1 pkt)*
a) Korzystając z paczki e1071 przetestuj działanie algorytmu Naive Bayes na bazie danych irysów. Przydatny link: http://ugrad.stat.ubc.ca/R/library/e1071/html/naiveBayes.html (na dole znajduje się nawet odniesienie do irysów)

b) Korzystając z powyższego linku, wzięliśmy jako zbiór testowy i treningowy całą bazę irysów. Podziel bazę w proporcjach 70/30 na zbiór treningowy i testowy tak, jak robiliśmy to dla algorytmu k-NN i jeszcze raz uruchom algorytm naiveBayes (skorzystaj z funkcji predict i ew. table)

Zadanie 3 *(1 pkt)*
Znajdź i wykorzystaj do klasyfikowania irysów paczkę w R do tworzenia drzew decyzyjnych np.
http://www.rdatamining.com/examples/decision-tree
http://www.r-bloggers.com/a-brief-tour-of-the-trees-and-forests/

Podaj przykład dwóch irysów (zestaw parametrów), które będą przyporządkowane do dwóch różnych kgatunków. Skorzystaj z drzewa decyzyjnego.

Podziel zbiór irysów na zb. testowy i treningowy. Dokonaj ewaluacji klasyfikatora (skorzystaj z funkcji predict i ew. table)

Zadanie 4 *(1 pkt)*
Porównaj działanie trzech klasyfikatorów:

- knn dla k=5
- NaiveBayes
- DrzewaDecyzyjne

dla zbioru danych diabetes.csv.

Szczególnie ważne jest zwrócenie uwagi na ewaluację.

- Wyświetl macierz błędu (CrossTable, ConfusionMatrix). Gdzie w macierzy błędu znajdują się wartości FP,TP,FN,TN? Co oznaczają skróty?
- Wyświetl dokładność każdego z algorytmów [% dobrze sklasyfikowanych instancji, czyli ((TP+TN)/liczba rekordów)]
- Z jakich wzorów obliczamy TP-Rate i FP-Rate (co to jest? https://en.wikipedia.org/wiki/Sensitivity_and_specificity).
- Zaznacz dla każdego klasyfikatora jego (FPRate, TPRate) jako punkt na wykresie z legendą. Wykres ma na osi x FPR, na osi y TPR i zakres od 0 do 1. Legenda powinna wyjaśniać nazwy klasyfikatorów.
- Gdzie znajdowałby się punkt na wykresie odpowiadający klasyfikatorowi idealnemu? Który z badanych klasyfikatorów jest najbardziej zbliżony do klasyfikatora idealnego?