Inteligencja obliczeniowa – stud. niestac.

Laboratorium 3: Zadanie klasyfikacji – poznanie trzech algorytmów klasyfikujących: kNN, NaiveBayes, drzewo decyzyjne.

Przy pomnijmy sobie bazę danych z irysami. Na poprzednich laboratoriach zmniejszaliśmy wymiarowość bazy danych, by lepiej zobrazować przynależność do gatunków na wykresie. Jednak nie napisaliśmy programu, który rozstrzyga czy irys z takimi a nie innymi parametrami należy do tego gatunku czy innego.

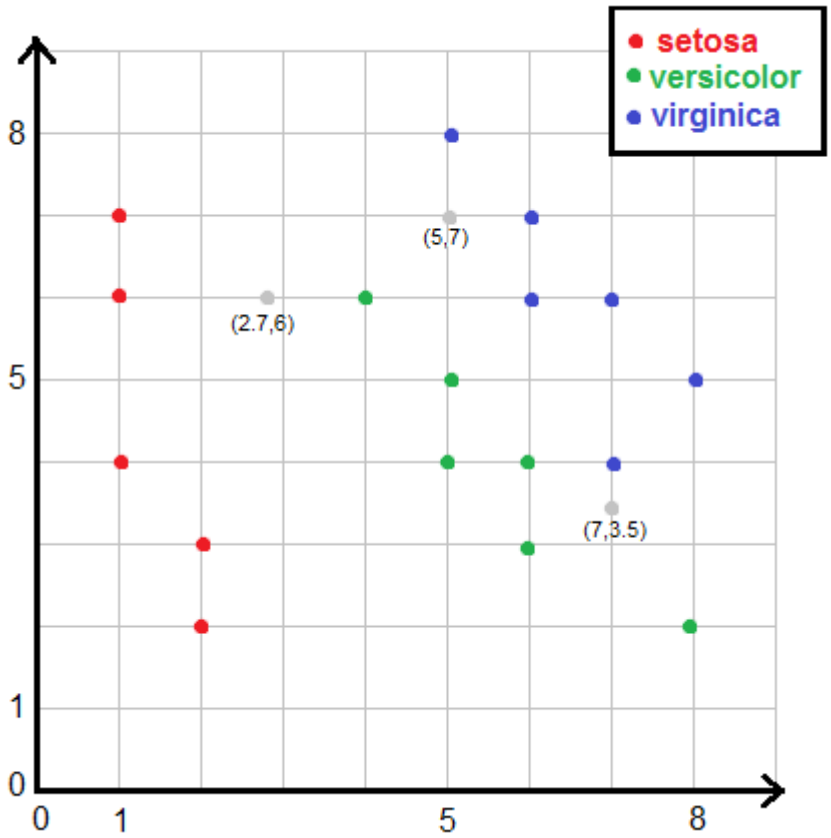
Zadanie rozstrzygania do jakiej klasy należy dana instancja (rekord) zwane jest zadaniem klasyfikacji. Istnieje wiele algorytmów, które klasyfikują rekordy. My poznamy (mniej lub bardziej) trzy z nich.

Zadanie 1 (1 pkt)
A) **Algorytm k-najbliższych sąsiadów (kNN)**
Klasyfikator ten przydziela rekordowi (np. irysowi) odpowiednią klasę (czli w przypadku irysów: gatunek) szukając n (na przykład 3) najbardziej podobnych instancji do niego i przydzielając mu klasę co większość z tych rekordów (trochę jakby w drodze głosowania).

Spójrz na rysunek i powiedz, jaki gatunek będą miały nowe irysy (3 szare kropki) jeśli klasyfikacji dokonamy algorytmem:

- 1-najbliższego sąsiada
- 3-najbliższych sąsiadów

Miarą podobieństwa będzie odległość euklidesowa.



Uwaga: szarych kropek po pokolorowaniu nie uwzględniaj w klasyfikowaniu kolejnych szarych.

B) **Algorytm NaiveBayes**
NaiveBayes to algorytm oparty na prawdopodobieństwie, wykorzystujący wzór Bayesa. Idea jest taka: spójrz co charakteryzuje poszczególne klasy (prawdop. apriori). Gdy pojawi się nowy rekord, wykorzystaj prawodpodobieństwa warunkowe i wzór Bayesa do obliczenia prawdopdobieństwa do jakiej klasy należy ten rekord (aposteriori).

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 (nazwijmy 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) = 0.02679`
`P(buys=no|X)=P(X|buys=no)*P(buys=no) =(1/81)*(3/7) = 0.00529`

Z obu wartości większa jest 0.02679, więc nasz rekord przyjmuje klasę yes.

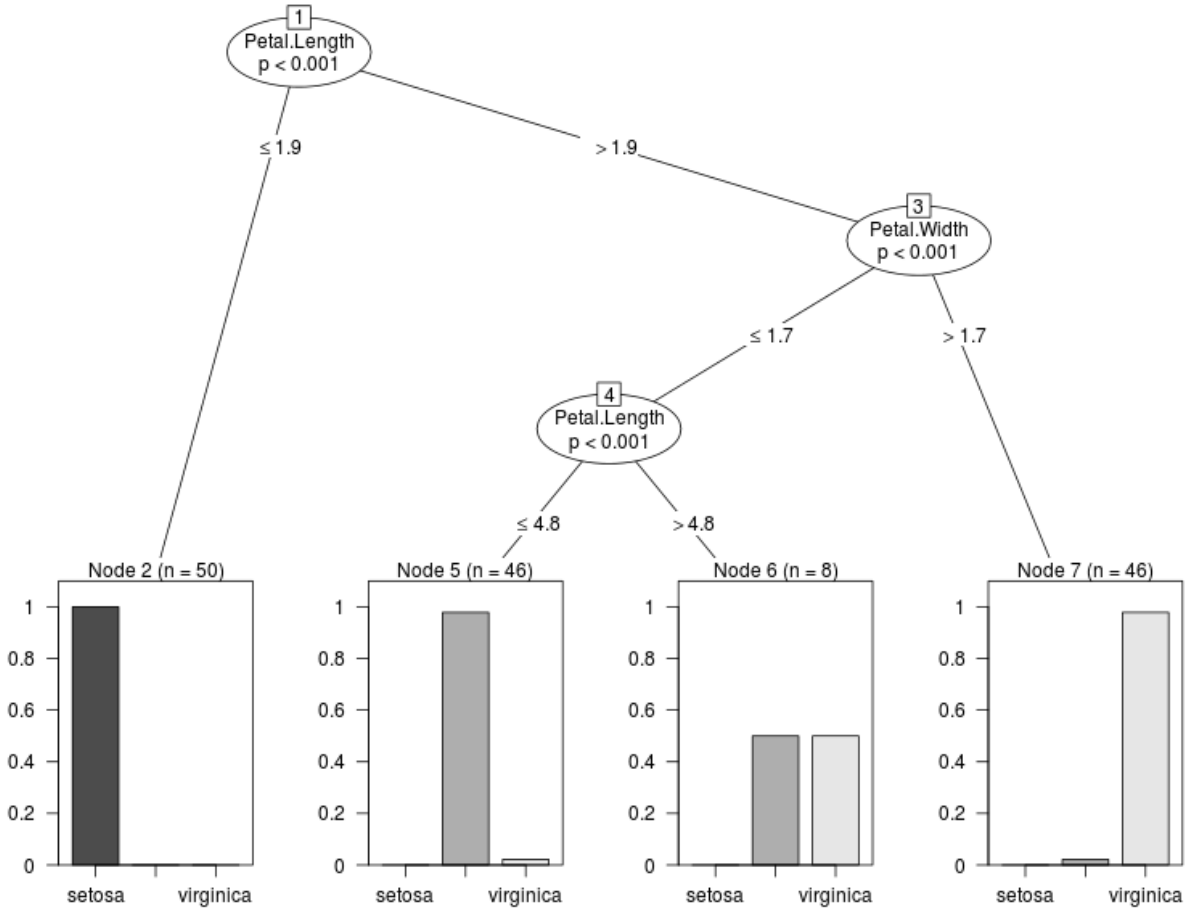
Mamy kolejną osobę Y:

>40	low	no	fair	???

Przyporzadkuj jej klasę obliczając z pomocą kalkulatora Google, Excela lub R wartości z czterech kroków. Obliczeń dokonaj na bazie siedmiu rekordów w tabeli (nie dołączaj do tabeli osoby X, poprzednio obliczanej).

C) **Drzewa decyzyjne (np. oparte na algorytmie C4.5)**
Klasyfikować rekordy można też na podstawie drzewa decyzyjnego. Budowanie drzewa jest stosunkowo skomplikowane (dużo logarytmowania i długich wzorów: patrz wykład), dlatego ograniczymy się do analizowania gotowego drzewa.

Załóżmy, że dla bazy z irysami algorytm wygenerował poniższe drzewo. Zauważmy, że drzewo bierze pod uwagę tylko pomiary petal (być może sepal nie były potrzebne / miarodajne). W liściach drzewa podana jest liczba irysów (n) oraz na wykresie proporcje gatunków. Można przyjąć, że gatunek, który ma najwyższy słupek wygrywa i jest przyporządkowany do badanego rekordu.



Jakie sklasyfikowane będą następujące trzy irysy?

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
4,7	3,2	1,3	0,2
6,6	2,9	4,6	1,3
7,9	3,8	6,4	2

Zadanie 2
Korzystając ze wskazówek w poradniku http://blog.datacamp.com/machine-learning-in-r/ i poniższych informacji dokonaj klasyfikacji irysów metodą k-najbliższych sąsiadów.
B). W owym poradniku zacznij od „Step 4”, choć możesz przeczytać również „Step 1,2,3”. Generalnie warto wczytać się co jest napisane, bo jest to opowiedziane rpzystepnym językiem.

Krok. 1. (przygotowanie i podział danych na zbiór testowy i treningowy)
Żeby algorytm nauczył się rozpoznawać gatunki irysów, trzeba dać mu zbiór treningowy, na którym

dokona „nauki”. Potrzebny jest też zbiór testowy, na którym przetestuje czy dobrze działa (czyli czy jego odpowiedzi, pokrywają się z klasą w tabeli).

a) wczytaj bazę danych irysów i znormalizuj dane liczbowe

b) podziel na zbiór treningowy i testowy

Krok.2 (klasyfikacja)

c) Uruchamiamy nasz algorytm, który pracuje na zbiorze treningowym. Skorzystaj z komendy knn by dokonać klasyfikacji.

d) Patrzymy teraz jakie gatunki przyporzadkowane będą irysom ze zbioru testowego. Można to rozumieć jak wstawianie szarych kropek na wykres składający się tylko z pokolorowanych kropek treningowych.

Krok. 3 (ewaluacja)
Irysy ze zbioru testowego zostały sklasyfikowane – dostałe etykietki przewidywane (predicted) od algorytmu. Czy faktycznie algorytm dobrze działał? Można go zewaluować go porównując klasę przewidzianą z realna (tą z tabeli).

e) Dokonaj ewaluacji. Sklej przewidziane gatunki z prawdziwymi ze zbioru testowego. Wyświetl jaki procent rekordów ma te same etykiety (liczbowo i w procentach). Do tego celu możesz napisać prosta funkcję lub dokonać prostych obliczeń w R.

f) Wyświetl macierz błędu (confusion matrix), tak jak zrobiono to w tutorialu. Sprawdź jakie gatunki pomylono.

Zadanie 3
Klasyfikator kNN działa dobrze. Zobaczymy jak działa klasyfikator NaiveBayes.

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 (z podpkt a), wzięliśmy jako zbiór testowy i treningowy całą bazę irysów. Podziel bazę w proporcjach 67/33 na zbiór treningowy i testowy tak, jak robiliśmy to dla algorytmu k-NN i jeszcze raz uruchom algorytm naiveBayes. Dokonaj ewaluacji (% poprawnych + confusion matrix). Być może konieczne będzie użycie funkcji predict i table na zbiorze testowym w tym podpunkcie.

Zadanie 4
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/
Podziel zbiór irysów na zb. testowy i treningowy. Dokonaj ewaluacji klasyfikatora (% poprawnych + confusion matrix). Być może konieczne będzie użycie funkcji predict i table na zbiorze testowym w tym podpunkcie.

Praca domowa

Zadanie 5
W załączonym zbiorze danych diabetes.csv znajdują się dane kobiet z USA, które zachorowały lub nie zachorowały na cukrzycę. Sprawdź jak działają poznane klasyfikatory na tej bazie danych. Dokonaj porównania:

- kNN, k=1
- kNN, k=3
- kNN, k=5
- kNN, k=11
- NaiveBayes
- Drzewa decyzyjne.

Stwórz dwa wykresy:

1. Wykres słupkowy, w którym każdy słupek odpowiada jednemu z klasyfikatorów, a jego wysokość to procent dobrze sklasyfikowanych instancji ze zbioru testowego (wysokość od 0 do 100%).

2. Wykres punktowy, w którym każdy punkt odpowiada klasyfikatorowi (na wykresie będzie 6 punktów w różnym kolorze i legenda wyjaśniająca jaki punkt odpowiada jakiemu klasyfikatorowi). Obie osie mają skalę od 0 do 1. Oś X będzie odpowiadała False Positive Rate, a oś Y True Positive Rate (Sensitivity), więc każdy z klasyfikatorów to punkt o współrzędnych (FP-Rate, TP-Rate).

TP-Rate i FP-Rate oblicza się na podstawie macierzy błędu (confusion matrix). Jak to obliczyć? Możesz rzucić okiem pod link: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Dodatkowe pytania:

- Gdzie znajdowałby się punkt na wykresie odpowiadający klasyfikatorowi idealnemu?
- Który z badanych klasyfikatorów jest najbardziej zbliżony do klasyfikatora idealnego?