

Zajęcia 4. Dyskryminacja oparta na regresji liniowej i logistycznej.

Ćwiczenia

Plik `choroby_serca.txt` zawiera dane dotyczące zachorowalności na choroby serca mężczyzn z obszaru Afryki Południowej. W danych znajdują się następujące zmienne:

<code>sbp</code>	skurczowe ciśnienie krwi
<code>tobacco</code>	całkowite spożycie tytoniu (kg)
<code>ldl</code>	cholesterol niskobiałkowy
<code>adiposity</code>	otyłość
<code>famhist</code>	obecność/brak chorób serca w historii rodziny (Present, Absent)
<code>typea</code>	zachowanie typu A
<code>obesity</code>	otyłość
<code>alcohol</code>	obecne spożycie alkoholu
<code>age</code>	wiek
<code>chd</code>	zawał serca : 1 - wystąpił, 0 - nie wystąpił

Dla tych danych wykonaj następujące polecenia:

- i) Dokonaj podziału obserwacji na podzbiór uczący (pierwsze 324 obserwacje) oraz zbiór testowy (pozostałe).
- ii) Dopasuj model regresji logistycznej bazując na podzbiorze uczącym.
- iii) Stosując krokową procedurę wyboru modelu zdecyduj, które zmienne powinny pozostać w modelu.
- iv) Skonstruuj klasyfikator w oparciu o próbę uczącą. Przedstaw na rysunku klasyfikację dokonaną przez klasyfikator. Wyznacz błąd klasyfikacji na próbie testowej.
- v) Skonstruuj klasyfikator w oparciu o metodę LDA bazując na próbie uczącej oraz porównaj jego błąd klasyfikacji na próbie testowej z błędem klasyfikacji klasyfikatora opartego na regresji logistycznej.

Zadania

Zadanie 7. Plik `urine.txt` zawiera dane dotyczące obecności kryształów oraz innych właściwości moczu. W danych znajdują się następujące zmienne:

<code>pres</code>	obecność kryształów (0- brak, 1- obecność)
<code>gravity</code>	ciężar właściwy
<code>ph</code>	wartość współczynnika pH
<code>osmo</code>	osmolarność moczu
<code>conduct</code>	przewodnictwo
<code>urea</code>	stężenie mocznika
<code>calc</code>	stężenie wapnia

Dla tych danych wykonaj polecenia z powyższego ćwiczenia. Dla wyznaczenia zbioru uczącego wybierz losowo 50 obserwacji ze zbioru danych, pozostałe traktując jako zbiór testowy.

Zadanie 8.

Plik `puls.txt` zawiera dane dotyczące poziomów tętna spoczynkowego (niskie/wysokie) w grupie 92 pacjentów. W danych znajdują się następujące zmienne:

<code>TetnoSpocz</code>	poziom tętna spoczynkowego (Niskie/Wysokie)
<code>Palacz</code>	czy pacjent pali papierosy (0 - nie pali, 1 - pali)
<code>Waga</code>	waga w kilogramach

Dla tych danych wykonaj następujące polecenia:

- i) Dokonaj podziału obserwacji na podzbiór uczący (75% obserwacji, wybranych w sposób losowy) oraz zbiór testowy (pozostałe).
- ii) Dopasuj model regresji logistycznej bazując na podzbiorze uczącym, przyjmując za zmienną objaśnianą `TetnoSpocz`.
- iii) Stosując krokową procedurę wyboru modelu zdecyduj, które zmienne powinny pozostać w modelu.
- iv) Dopasuj model regresji logistycznej bazując na podzbiorze uczącym, uwzględniający interakcję pomiędzy zmiennymi objaśniającymi.
- v) Przeanalizuj modele zawierające jako zmienne objaśniające następujące zestawy zmiennych:

M1: Palacz, Waga, Palacz*Waga

M2: Waga, Palacz*Waga

M3: Palacz, Palacz*Waga

M4: Palacz*Waga

Który z tych modeli zawiera tylko istotne zmienne objaśniające?

vi) Dla pełnego modelu bez interakcji i modelu wskazanego w punkcie **(v)** skonstruuj klasyfikatory w oparciu o próbę uczącą. Przedstaw na rysunku klasyfikację dokonaną przez oba klasyfikatory. Wyznacz błąd klasyfikacji na próbie testowej.

vii) Skonstruuj klasyfikatory w oparciu o metodę LDA bazując na próbie uczącej oraz porównaj ich błędy klasyfikacji na próbie testowej z błędami klasyfikacji klasyfikatorów opartych na regresji logistycznej. Skomentuj uzyskane wyniki.