Eksploracja danych
Lista nr 3
29 IV 2025

Zadanie 1 Klasyfikacja na bazie modelu regresji liniowej

a) **Analizowane dane**
Ramka danych `iris`{pakiet `datasets`}. Mamy $K = 3$ klasy oraz $p = 4$ zmienne objaśniające $\mathbf{X} = (PL, PW, SL, SW)$.

b) **Podział danych na zbiór uczący i testowy**
Podziel losowo dane na część uczącą (ang. learning set) i testową (ang. test set), np. w proporcji: 2/3 – zbiór uczący, 1/3 – zbiór testowy.

c) **Konstrukcja klasyfikatora i wyznaczenie prognoz**
Zastosuj model regresji liniowej do konstrukcji klasyfikatora na podstawie danych uczących i wyznacz prognozowane etykietki klas dla przypadków ze zbioru uczącego oraz przypadków ze zbioru testowego.

d) **Ocena jakości modelu**
Wyznacz macierz pomyłek (ang. confusion matrix) oraz błąd klasyfikacji na zbiorze uczącym oraz zbiorze testowym. Spróbuj zilustrować/zbadać czy w tym przypadku obserwujemy zjawisko maskowania klas, o którym była mowa na wykładzie.

e) **Budowa modelu liniowego dla rozszerzonej przestrzeni cech**
Powtórz konstrukcję modelu i ocenę jego dokładności (kroki c)-d)), tym razem budując model regresji po uzupełnieniu wyjściowych cech o składniki wielomianowe stopnia 2 (tzn.: PL^2, PW^2, SL^2, SW^2, $PL \cdot PW$, $PL \cdot SW$, $PL \cdot SL$, $PW \cdot SL$, $PW \cdot SW$, $SL \cdot SW$). Porównaj dokładność klasyfikacji dla obu przypadków i zinterpretuj otrzymane wyniki.

Zadanie 2 Porównanie metod klasyfikacji

a) **Wybór i zapoznanie się z danymi**
Zadanie wykonujemy dla jednego (wybranego) zbioru danych. Do wyboru są następujące dane:

- `Glass` {biblioteka `mlbench`}
- `Wine` {biblioteka `HDclassif`}
- `PimaIndiansDiabetes2` {biblioteka `mlbench`}
- `Vehicle` {biblioteka `mlbench`}
- `Sonar` {biblioteka `mlbench`}

Na wstępie proszę uważnie zapoznać się z opisem wybranego zbioru danych.[1] Należy m.in. ustalić:

- Ile mamy zmiennych (cech) i przypadków?
- Ile mamy klas (grup) i która zmienna zawiera informacje o przynależności obiektu do konkretnej klasy (tzw. etykietki klas)?
- Czy występują jakieś specyficzne/nietypowe własności danych (np. brakujące lub nietypowe wartości, niestandardowe kodowanie niektórych wartości, itp.)?
- Czy wszystkie zmienne mają prawidłowo przypisane typy (w szczególności, czy zmienna zawierająca etykietki klas jest zmienną typu `factor`)?

[1] Opis danych jest dostępny w dokumentacji dołączonej do odpowiedniego R-pakietu. Bardziej szczegółowy opis można znaleźć również na stronie http://archive.ics.uci.edu/ml/ (repozytorium UCI).

b) **Cel analizy**
Celem analizy jest zastosowanie poznanych algorytmów klasyfikacji i szczegółowe porównanie ich dokładności. Porównanie to powinno uwzględniać przynajmniej następujące algorytmy:

- metoda k-najbliższych sąsiadów (*k-Nearest Neighbors*),
- drzewa klasyfikacyjne (*classification trees*),
- naiwny klasyfikator bayesowski (*naïve Bayes classifier*).

Poniżej umieszczone są informacje nt. zagadnień, które powinny zostać uwzględnione w analizie.

c) **Wstępna analiza danych**
Wykorzystując poznane metody analizy opisowej, zbadaj najważniejsze własności danych, w tym:

- Jak wygląda rozkład klas w analizowanym zbiorze i czy obserwujemy istotne dysproporcje w liczebności poszczególnych grup? Jaki błąd klasyfikacji otrzymalibyśmy przypisując wszystkie obiekty do (jednej) najczęściej występującej klasy?
- Czy obserwujemy istotne różnice w zmienności (wariancji) poszczególnych cech, co może oznaczać konieczność zastosowania standaryzacji w przypadku niektórych algorytmów klasyfikacyjnych?
- Które zmienne charakteryzują się najlepszymi zdolnościami dyskryminacyjnymi/predykcyjnymi (tzn. najlepiej separują obiekty należące do różnych klas)?

d) **Ocena dokładności klasyfikacji i porównanie metod**

- Porównaj dokładność klasyfikacji rozważanych algorytmów, dzieląc oryginalny zbiór danych (w odpowiedniej proporcji) na zbiór uczący (*learning set*) i zbiór testowy (*test set*). Do oceny dokładności wykorzystaj macierz pomyłek *(confusion matrix)* oraz błąd klasyfikacji. Szczegółowo zinterpretuj otrzymane wyniki. W szczególności porównaj błędy klasyfikacji na zbiorze uczącym i testowym.
- Do porównania metod wykorzystaj następnie wybrany, bardziej zaawansowany schemat oceny dokładności, np. wielokrotny podział na zbiór uczący i testowy, metodę cross-validation lub schemat typu bootstrap[2]. Porównaj wnioski dotyczące skuteczności metod z tymi otrzymanymi na podstawie pojedynczego podziału na zbiór uczący i testowy.

e) **Różne parametry i różne podzbiory cech**
Przeprowadzając porównanie dokładności klasyfikacji (patrz podpunkt d), uwzględnij:

- różne kombinacje (podzbiory) zmiennych wykorzystanych do konstrukcji klasyfikatorów, w szczególności wszystkie zmienne oraz wybrany podzbiór zmiennych o najlepszej zdolności dyskryminacyjnej. Wskazówka: do wyboru „najbardziej obiecujących" kombinacji zmiennych można wykorzystać wnioski otrzymane w podpunkcie c).
- różny dobór parametrów dla poszczególnych metod (np. różną liczbę sąsiadów w metodzie k-NN, różne parametry dla algorytmu konstrukcji drzew klasyfikacyjnych itp.)

f) **Końcowe wnioski – podsumowanie**
Na podstawie przeprowadzonej analizy, spróbuj odpowiedzieć na następujące pytania:

- Dla jakiego podzbioru zmiennych predykcyjnych i dla jakich parametrów poszczególnych metod otrzymujemy najlepsze wyniki?
- Która z metod klasyfikacyjnych daje lepsze, a które gorsze rezultaty w przypadku analizowanych danych?
- Czy wybór schematu oceny dokładności miał istotny wpływ na wnioski dotyczące skuteczności metod?

[2]Mile widziane jest porównanie wyników dla różnych schematów