

Zajęcia 5. Dyskryminacja oparta na regresji liniowej i logistycznej.

Ćwiczenia

Plik `hsb.txt` zawiera dane dotyczące 200 uczniów różnych typów szkół ich statusu ekonomicznego i wyników testów z różnych dziedzin. Zmienną objaśniającą jest zmienna `program` oznaczająca typ szkoły, za zmienne objaśniające przyjmujemy zmienne `status.ek.` - status ekonomiczny o trzech poziomach (niski/średni/wysoki) oraz `pisanie` - wynik testu językowego z pisania.

Dla tych danych wykonaj następujące polecenia:

- i) Wygeneruj podpróbę uczącą o 150 elementach ze zbioru danych. Pozostałe dane przyjmij jako zbiór testowy.
- ii) Przedstaw tabelę liczości w grupach wyznaczonych przez zmienne `status.ek.` oraz `program` w zbiorze uczącym. Przedstaw tabelę wartości średnich i odchyłeń standardowych w grupach wyznaczone na podstawie zbioru uczącego.
- iii) Stosując funkcję `multinom` dokonaj dopasowania modelu log-liniowego opisującego zależność pomiędzy zmienną `program` a zmiennymi `status.ek.` oraz `pisanie`.
- iv) Zmiennej `pisanie` przypisz jej wartość średnią a następnie wyznacz przewidywane prawdopodobieństwa przynależności do klas. Wskaż klasę przewidywaną przez model wyznaczony w punkcie poprzednim w oparciu o regułę Bayesa.
- v) Przedstaw tabelę klasyfikacji i wyznacz błąd klasyfikacji.
- vi) Przeprowadź analizę LDA dla badanego modelu, wyznacz błąd klasyfikacji i porównaj z wynikiem uzyskanym w iv).
- vii) Powtórz analizę modeli 10 razy, za każdym razem generując na nowo próbę uczącą. W każdym kroku wyznacz proporcję poprawnych klasyfikacji dla obu metod. Wylicz średnią proporcję poprawnych klasyfikacji dla obu metod.
- viii) Wygeneruj nowe wartości dla zmiennych `status.ek.` oraz `pisanie`, każdej wartości zmiennej `status.ek.` przypisując ciąg liczb od 30 do 70 (jako wartości zmiennej `pisanie`). Wyznacz wartości prawdopodobieństw przynależności obserwacji do każdej z trzech grup wyznaczonych przez wartości zmiennej `program` oraz średnie wartości tych prawdopodobieństw w grupach.
- ix) Przedstaw rysunek przewidywanych wartości prawdopodobieństw w zależności od wartości zmiennej `pisanie` dla trzech poziomów zmiennej `status.ek.` dla każdego z poziomów zmiennej `program`.

Zadania

Zadanie 9. Podobnie jak w powyższym ćwiczeniu przeanalizuj zbiór danych `iris` (pakiet `MASS`). Za zmienną objaśnianą przyjmij zmienną `Species`, za zmienne objaśniające przyjmij zmienne `Sepal.Length` oraz `Petal.Width`. Zmienną `Sepal.Length` przedstaw jako zmienną o trzech możliwych poziomach: krótki, średni, długi dzieląc równomiernie zakres możliwych wartości tej zmiennej. Rozmiar próby uczącej ustal na 80% liczby wszystkich obserwacji i wybierz ją w sposób losowy. Nowe dane (punkt **viii**)) wygeneruj przypisując każdemu z poziomów zmiennej `Sepal.Length` 41 wartości z zakresu od minimum do maksimum zmiennej `Petal.Width` o równych odstępach. W punkcie **ix**) przedstaw rysunek przewidywanych wartości prawdopodobieństw w zależności od wartości zmiennej `Petal.Width` dla trzech poziomów zmiennej `Sepal.Length` dla każdego z poziomów zmiennej `Species`.