

Zajęcia 3. Liniowe metody klasyfikacji. Fisherowska dyskryminacja liniowa

Ćwiczenia

Plik `dane_kosiarki.txt` zawiera dane dotyczące posiadaczy kosiarek jeżdżących do trawy.

X_1	wysokość dochodu w tysiącach dolarów
X_2	wielkość działki w tysiącach metrów kwadratowych
Y	informacja, czy rodzina posiada, czy nie posiada kosiarkę 1- oznacza, że posiada, 2 - że nie posiada kosiarki.

Dla tych danych wykonaj następujące polecenia:

- i) Podziel dane na dwie grupy zależne od wartości zmiennej Y .
- ii) Przedstaw te dane na rysunku, używając różnych symboli klas.
- ii) Uzypełnij rysunek liniową funkcją klasyfikującą (prostą dyskryminacyjną) otrzymaną przy założeniu o rozkładach normalnych obserwacji w obu grupach. W tym celu :

1.) Stwórz macierze złożone ze zmiennych X_1 i X_2 dla rodzin posiadających i nie posiadających kosiarki.

2.) Oblicz wektory średnich wartości w klasach, próbkowe macierze kowariancji (s_1 oraz s_2) oraz próbkową macierz kowariancji wspólną dla obu klas

$$W = \frac{1}{n-2}((n_1-1) * s_1 + (n_2-1) * s_2),$$

gdzie n jest licznością próbki a n_1 i n_2 odpowiednimi licznościami klas.

3.) Oblicz macierz odwrotną do wspólnej macierzy kowariancji (polecenie `solve(W)`).

4.) Można wykazać, że wektor maksymalizujący iloraz

$$\frac{(\mathbf{a}^T \bar{x}_1 - \mathbf{a}^T \bar{x}_2)^2}{\mathbf{a}^T W \mathbf{a}}$$

jest proporcjonalny do wektora $W^{-1}(\bar{x}_1 - \bar{x}_2)$, natomiast regułę klasyfikacyjną można sformułować jako:

Przypisz punkt x do klasy 1, jeśli

$$(\bar{x}_1 - \bar{x}_2)^T W^{-1} x > \frac{1}{2} (\bar{x}_1 - \bar{x}_2)^T W^{-1} (\bar{x}_1 + \bar{x}_2)$$

Korzystając z tych informacji wyznacz prostą dyskryminacyjną najlepiej rozdzielającą klasy.

5.) Przedstaw wyznaczoną prostą dyskryminacyjną na rysunku wspólnie z danymi.

Zadania

Zadanie 5. Dla danych z powyższego ćwiczenia przeprowadź analizę dyskryminacyjną korzystając z funkcji `lda` w pakiecie **MASS**. Prawdopodobieństwa a priori przynależności do klas przyjmij jako równe 0.5.

- i) Jakie są średnie wartości w grupach?
- ii) Jakie są wartości współczynników funkcji dyskryminacyjnych?
- iii) Dla każdej z funkcji dyskryminacyjnych określ która ze zmiennych objaśniających jest najbardziej istotna.
- iv) Jakie są średnie wartości funkcji dyskryminacyjnych w każdej z grup?
- v) Przeprowadź predykcję przynależności do klas na zbiorze testowym.
- vi) Jakie są prawdopodobieństwa przynależności do klas dla dwudziestej obserwacji?
- vii) Do której z klas zostanie zaklasyfikowana ta obserwacja?
- viii) Wyznacz procent właściwie zaklasyfikowanych obiektów.
- ix) Przedstaw i skomentuj tabelę predykcji.

Zadanie 6. Plik `dane_wino.txt` zawiera dane dotyczące koncentracji trzynastu różnych substancji chemicznych zawartych w winie pochodzącym z jednego regionu Włoch, ale z trzech różnych upraw. Zmienna `V1` jest indeksem ze zbioru $\{1, 2, 3\}$ oznaczającym trzy typy uprawy. Zmienne `V2` do `V14` oznaczają poziomy koncentracji kolejnych substancji chemicznych. Dla tych danych przyjmij, że prawdopodobieństwa a priori przynależności do klas są równe $\frac{1}{3}$. Odpowiedz na pytania **i)-ix)** z powyższego zadania.