Case study…….pkt, **Imię i nazwisko…….**
Pytania zamknięte…….pkt
Razem……. pkt

Ocena……

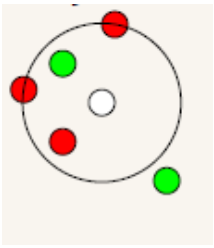
K O L O K W I U M 1

15 marca 2019 r.

Statystyczna Eksploracja Danych

Pytania zamknięte (20 pytań po 0.5 pkt każde, jedna odpowiedź prawidłowa)

Zad 1. Ile punktów zostanie wziętych pod uwagę podczas wyznaczania przynależności do klasy punktu środkowego metodą 3-nn (rysunek po prawej)?



(a) 3
(b) 4
(c) 2

Zad 2. Ten sam rysunek, co w poprzednim pytaniu. Czy w przypadku metody (4,3)-nn punkt centralny zostanie
(a) zaliczony do klasy czerwonej,
(b) zaliczony do klasy zielonej,
(c) decyzja nie zostanie podjęta.

Zad 3. Jakie założenia prowadzą do różnych postaci reguły klasyfikacyjnej w metodach LDA i QDA?
(a) jednowymiarowy rozkład Gaussa (LDA), dwuwymiarowy rozkład Gaussa (QDA),
(b) różne macierze kowariancji (QDA), równe macierze kowariancji (LDA),
(c) stosowalność dla dwóch klas (LDA), stosowalność dla trzech klas (QDA).

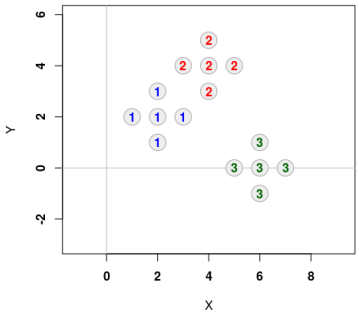
Zad 4. Wybierz **prawdziwą** cechę naiwnego klasyfikatora Bayesowskiego:
(a) multiplikatywny wpływ kolejnych składowych wektora obserwacji,
(b) Gaussowski rozkład obserwacji p(k|**x**),
(c) niezależność składowych wektora obserwacji.

Zad 5. W metodzie bootstrap:
(a) losujemy bez zwracania
(b) losujemy ze zwracaniem
(c) tworzymy tylko jedną pseudopróbę.

Case study (5 pkt)

Dane są trzy serie danych (x,y): $\mathbf{X_1}$, $\mathbf{X_2}$ oraz $\mathbf{X_3}$, każda pochodząca z innej klasy (czyli **g=3**), w każdej z nich jest po **5 punktów:**

$$X_1=\begin{pmatrix}1 & 2\\ 2 & 3\\ 2 & 2\\ 2 & 1\\ 3 & 2\end{pmatrix} \quad X_2=\begin{pmatrix}3 & 4\\ 4 & 5\\ 4 & 4\\ 4 & 3\\ 5 & 4\end{pmatrix} \quad X_3=\begin{pmatrix}5 & 0\\ 6 & 1\\ 6 & 0\\ 6 & -1\\ 7 & 0\end{pmatrix}$$



Oblicz optymalny kierunek rozdzielający **a** za pomocą metody Fishera. Oceń, czy metoda umożliwia rozdzielenie klas. Dlaczego drugi kierunek (odpowiadający drugiej wartości własnej) jest gorszy?

Wzory:

Wartość oczekiwana w każdej klasie k ($\mathbf{n_k}$ to liczba punktów w danej klasie):

$$\boldsymbol{m}_k=\frac{1}{n_k}\sum_{i=1}^{i=n_k}\boldsymbol{x}_i$$

Macierz kowariancji w każdej klasie k:

$$\boldsymbol{S}_k=\frac{1}{n_k-1}\sum_{i=1}^{i=n_k}(\boldsymbol{x}_i-\boldsymbol{x}_k)(\boldsymbol{x}_i-\boldsymbol{x}_k)^T$$

Macierz kowariancji wewnątrzgrupowej **W (n** to całkowita suma punktów **n = $\mathbf{n_1}$ + $\mathbf{n_2}$ + $\mathbf{n_3}$)**:

$$\boldsymbol{W}=\frac{1}{n-g}\sum_{k=1}^{k=g}(n_k-1)\boldsymbol{S}_k$$

Macierz kowariancji międzygrupowej **B (m** to średnia ze wszystkich punktów**)**:

$$\boldsymbol{B}=\frac{1}{g-1}\sum_{k=1}^{k=g}n_k(\boldsymbol{m}_k-\boldsymbol{m})(\boldsymbol{m}_k-\boldsymbol{m})^T$$

Optymalny kierunek **a** otrzymujemy maksymalizując wyrażenie

$$J=\frac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a}}$$

wyznaczając wektor własny odpowiadający największej wartości własnej równania

$$\boldsymbol{A} = \boldsymbol{W}^{-1}\boldsymbol{B}$$

Wartości własne λ można otrzymać wyznaczając pierwiastki wielomianu charakterystycznego

$$det\left(\boldsymbol{A} - \lambda \boldsymbol{I}\right) \quad ,$$

gdzie **I** to macierz jednostkowa (same zera oprócz jedynek na diagonali), a odpowiadające poszczególnym wartościom własnym λ_i wektory własne $\mathbf{x_i}$ rozwiązując równania

$$\left(\boldsymbol{A} - \lambda_i \boldsymbol{I}\right) \cdot \boldsymbol{x_i} = 0$$