Семинар 3
Класификация по разстоянието до центроидите на извадката

Задача 4.1. От лекциите по Аналитична геометрия си припомнете уравнението за равнина в тримерното пространство. Обобщете уравнението за случая на N-мерно пространство. Намирате ли прилика с уравнение (4.2)? Ако да, за коя равнина става въпрос в уравнение (1)?

Решение: Уравнението на равнина е:

$$ax + by + cz = d$$

$$y(X) = \sum_{n=1}^{N} w_n x_n + w_{n+1} \qquad (4.2)$$

Ако в уравнение (4.2) заместим y(X) = 0, което отговаря на разделящата повърхност, ще получим за тримерен образ

$$w_1x_1 + w_2x_2 + w_3x_3 = - w4,$$

което на практика е същото уравнение.

Задача 4.2. Пречи ли на класификацията припокриването на образите от двата класа в пространството на образите? Как влияят бегълците (*outliers*) на положението на центроидите? А на класификацията?

Отговор: Образ, който е беглец, променя в много голяма степен центроида на съответната извадка – затова пречи. Класификацията също ще се промени значително, ако се сравни с тази, получени при премахване на беглеца.

Задача 4.3. Имате два четиримерни образа

$X_1 = (0.7, 0.2, -0.1, 0.8)$ $X_2 = (-0.6, 0.1, 0.2, 0.5)$

а) Изчислете сумата им $S = X_1 + X_2$

$S = X_1 + X_2 = (0.7, 0.2, -0.1, 0.8) + (-0.6, 0.1, 0.2, 0.5) = (0.1, 0.3, 0.1, 1.3)$

b) Изчислете разликата им D = X_1 - X_2

S = X_1 - X_2 = (0.7, 0.2, -0.1, 0.8) - (-0.6, 0.1, 0.2, 0.5) = (1.3, 0.1, -0.3, 0.3)

c) Изчислете произведението на първия с числото 0.4: P = 0.4*X_1

$$P = 0.4*X_1 = 0.4x(0.7, 0.2, -0.1, 0.8) = (0.28, 0.08, -0.04, 0.32)$$

Задача 4.4. Имате пет четиримерни образа

X_1 = (0.7, -0.2, 0.1, 0.8) X_2 = (0.6, 0.1, -0.2, 0.5) X_3 = (0.7, 0.6, 0.3, 0.8)

X_4 = (0.5, 0.5, -0.4, 0.5) X_5 = (0.3, 0.4, 0.4, -0.5)

Изчислете образа M, който отговаря на центроида на тази извадка.

Решение: Първо изчисляваме тяхната сума, като сумираме съответните им признаци:

$$S = (2.8, 1.4, 0.2, 2.1)$$

След това делим всеки признак на сумата на броя на образите, 5. Получаваме:

$$M = (0.56, 0.28, 0.04, 0.42)$$

Задача 4.5. Имате два четиримерни образа

$$M_1 = (0.7, 0.2, 0.1, -0.8) \text{ и } M_2 = (-0.6, 0.1, -0.2, 0.5)$$

Изчислете разстоянието в Манхатан от тях до образа X.

$$X = (0.3, -0.4, 0.4, 0.5)$$

Ако M_1 и M_2 са центроиди, от кой клас е образът X?

Решение: За първото разстояние, |X - M_1|, получаваме:

$$d_1 = |0.7- 0.3| + |0.2 – (-0.4)| + |0.1 - 0.4| + |-0.8 - 0.5| =$$

$$= |0.4| + |0.6| + |-0.3| + |-1.3 | = 0.4 + 0.6 + 0.3 + 1.3 = 2.6$$

За второто разстояние, $|M_2 - X|$, получаваме:

$$d_2 = |0.3 - (-0.6)| + |-0.4 - 0.1| + |0.4 - (-0.2)| + |0.5 - 0.5| =$$

$$= |0.9| + |-0.5| + |-0.6| + |0 | = 0.9 + 0.5 + 0.6 + 0 = 2.0$$

Ако M_1 и M_2 са центроиди, образът X е втория клас, защото разстоянието от X до втория центроид е по-малко от разстоянието от X до първия центроид.

Практически задачи

Задача C1. Отворете файла **seminar04_centroids.xls**. Разгледайте таблицата (sheet) "**Spectra**" в която са дадени девет двумерни образа, чиито признаци са спектрални признаци, изчислени от ИЧ спектри. Първите четири съединения нямат първична алкохолна група, докато вторите пет са първични алкохоли. Двата центроида са изчислени на редове 11 и 12, а след тях има две „неизвестни" съединения, които трябва да се класифицират.

На редове 17-29 са изчислени разстоянията до центроидите на двата класа – в колона **F** до центроида на клас 0, а в колона **J** до центроида на клас 1 (първични алкохоли). Каква мярка за разстояние е използвана?

Разгледайте колона C в редове 17-29. Какво означава формулата "**=IF(F17<J17, 0, 1)**" в клетка C17? Какво точно се изчислява от C17 до C29?

Задача C2. Във файла **seminar04_centroids.xls** в таблицата (sheet) **Euclidean** повторете изчисленията от таблицата **Spectra**. В таблицата (sheet) **Manhatten** във файла **seminar04_centroids.xls** проведете изчисленията с използване на разстояние в Манхатън.

Задача C3. В таблицата **Manhatten** във файла **seminar04_centroids.xls** повторете изчисленията от таблицата **Spectra** като използвате разстояние в Манхатън.

Задача C4. В таблицата (sheet) **Plot** във файла **seminar04_centroids.xls** е изчислена разделящата права и начертана на графика, заедно с 13^те образа – 9 образа от обучаващата извадка, 2 центроида и 2 образа от тестващата извадка. Следващите факти в т. 1 до т. 7 са приложени, за да се начертае тази права.

1. От геометрията е известно, че ако прекараме през средата на отсечката между две точки (в случая, тези на центроидите) права, която е перпендикулярна на отсечката (в случая, разделящата права), то разстоянията от коя да е точка на разделящата права до двата центроида са равни.

2. Уравнението на права, която минава през две точки (x_1,y_1) и (x_2,y_2) се дава с елегантна формула, чийто елементарен геометричен смисъл преподавателят обяснява на лекциите:

$$(y - y_1)/(x - x_1) = (y_2 - y_1)/(x_2 - x_1)$$

Наклонът на тази права очевидно е $(y_2 - y_1)/(x_2 - x_1)$. За правата, която минава през двата центроида $(m1_1,m1_2)$ и $(m2_1,m2_2)$ наклонът и ще е

$$a_1 = (m2_2 - m1_2)/(m2_1 - m1_1)$$

3. От аналитичната геометрия е известно, че ако две прави с уравнения $y = a_1x + b_1$ и $y = a_2x + b_2$ са перпендикулярни, то $a_1a_2 = -1$. Тогава наклонът на разделящата права (която да повторим – е перпендикулярна на отсечката между двата центроида) ще е

$$a_2 = -1/a_1 = -(m2_1 - m1_1)/(m2_2 - m1_2)$$

4. Разделящата права минава през средата на отсечката между центроидите, която е своеобразен техен центроид и има координати

$$\tfrac{1}{2}\,(m2_1 + m1_1,\ m2_2 + m1_2)$$

5. Отрезът b_2 на права с наклон a_2, която минава през точка (x_2,y_2) е

$$b_2 = y_2 - a_2 x_2$$

6. Тогава отрезът на правата, която е перпендикулярна на отсечката между двата центроида и през средата на отсечката между центроидите ще е

$$b_2 = \tfrac{1}{2} (m2_2 + m1_2) - [-(m2_1 - m1_1)/(m2_2 - m1_2)]\tfrac{1}{2} (m2_1 + m1_1)$$

(използвайте цветовете в последните две уравнения)

или [използвахме че $(a + b)(a - b) = a^2 - b^2$]

$$b_2 = \tfrac{1}{2} (m2_2 + m1_2) + \tfrac{1}{2} (m2_1^2 - m1_1^2)/(m2_2 - m1_2)$$

или когато приведем под общ знаменател $(m2_2 - m1_2)$

$$b_2 = \tfrac{1}{2} [(m2_2^2 - m1_2^2) + (m2_1^2 - m1_1^2)]/(m2_2 - m1_2)$$

или

$$b_2 = \tfrac{1}{2} [m2_1^2 + m2_2^2 - m1_1^2 - m1_2^2)]/(m2_2 - m1_2)$$

7. По последната формула в клетка `D21` от таблицата "`Plot`" във файла `seminar04_centroids.xls` е изчислен отрезът, а в клетка `D22` - наклонът на разделящата права. В региона `B24:B34` са дадени „хиксовете" на 11 точки от правата, а „игреците" им са изчислени в региона `C24:C34`, за да се нарисува разделящата права.

Задачата Ви е да проследите горните формули и да разберете изчисленията и рисунката в Excel.