Лекция 4

Класификация по разстоянието до центроидите на извадката

Класификацията по разстоянието до центроидите на съответните извадки от образи е най-елементарният метод. Въпреки това методът е изключително мощен и надежден, когато се сравни с другите линейни методи за класификация и продължава да се използва в научните и приложни изследвания.

4.1. Центроид. Средният вектор на група от вектори се нарича *центроид*. Той се изчислява като се усреднят съответните координати на векторите в групата и очевидно има същата размерност като тях. В едномерното пространство центроидът е средната стойност. Ако имаме тримерни образи и те се представят като точки в тримерното пространство, то центроидът е точката, която е геометричен център на групата и формулата на центроида (4.1) съвпада с центъра на масите на група от точкови маси, които са равни – ето защо центроидът се нарича още *център на масите*, или неправилно *център на тежестта*, а съответно методът има в името си тези понятия.

Координатите (признаците), c_n, на съответните центроиди се изчисляват по формулите:

$$c_n^{(k)} = \frac{\sum_{m=1}^{M_k} x_{m,n}^{(k)}}{M_k}; n = 1, 2 \ldots N \qquad (4.1)$$

където с k е означен k^{тият} клас (от общо K на брой класове), с m – m^{тият} обект от този клас, M_k е съответния брой образи в този клас, n е номерът на признака, а N е броят признаци (размерността на образите).

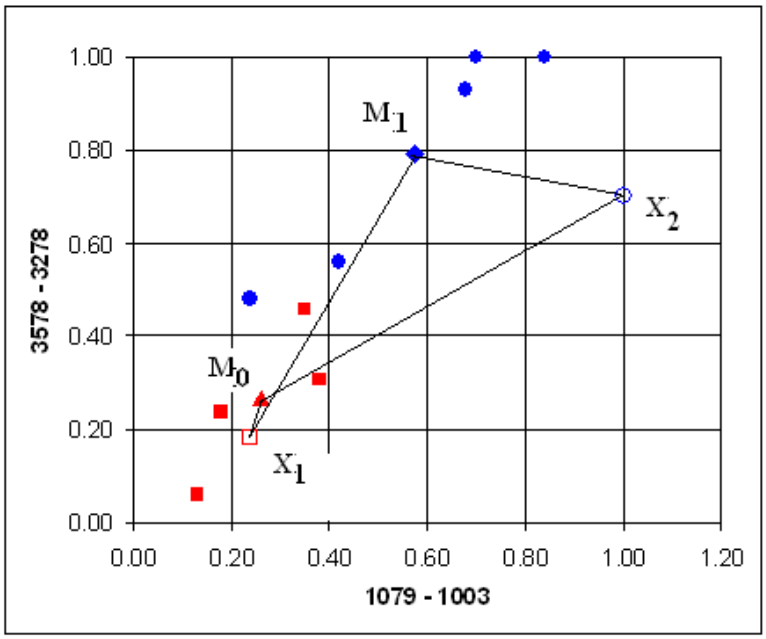
4.2. Метод на центроидите. Идеята на този метод е изключително елементарна [1] – изчислява се разстоянието в пространството на образите между класифицирания образ и центроидите на класовете. Образът се избира от този клас, до чийто центроид има най-малко разстояние.

Изчисляват се всички разстояния между непознатия образ, X, и всичките K центроида, $C^{(k)}$, $d_k = D(C^{(k)}, X)$; k = 1, 2, … K. Непознатият образ се класифицира към този клас p, за който имаме $d_p = \min(d_1, d_2, \ldots d_K)$.

Видът на разстоянието, което се използва, се избира от изследователя и обикновено зависи от типа на признаците, които съставят хемометричните образи. Оптималната за дадена класификация мярка за разстояние зависи и от типа на хемометричните обекти, чийто образи се класифицират – ето защо е правилно да се проверят няколко мярки с помощта на обучаваща и тестваща извадка. Обучаващата извадка се използва да се изчислят центроидите на класовете, а образите от тестващата извадка се използват като непознати образи, чиято класификация се оптимизира.

4.3. Нагледно представяне на метода. На фигура 4.1 е показана класификацията по два класа на два "непознати" образа, $\mathbf{X_1}$ и $\mathbf{X_2}$ в двумерното пространство: на фигурата са изобразени съответните Евклидови разстояния. Признаците на образите представляват максимумите на ивиците в ИЧ спектри в интервалите 1079-1003 cm^{-1} (x_1) и 3578-3278 cm^{-1} (x_2). В първия клас са съединения, които са първични алкохоли (кръговете), а в нулевия клас – които не са (квадратите). Съответните им центроиди са изобразени като ромб (клас 1) и триъгълник (клас 0), а двата "непознати" образа са изобразени като празен квадрат ($\mathbf{X_1}$) и празен кръг ($\mathbf{X_2}$).

Вижда се, че образ $\mathbf{X_1}$ е по-близо до центроида на нулевия клас, а $\mathbf{X_2}$ – до центроида на първи клас. В таблица 4.1 са дадени съответните образи от фигура 4.1.



Фигура 4.1. Класификация по разстоянието до центроидите. Обучаващата извадка се състои от девет образа, съответно 4 от клас 0 и 5 от клас 1.

Таблица 4.1. Хемометричните обекти и образи, които са изобразени на фигура 4.1.

Име на съединението	**Библ. номер**	**Първичен алкохол**	**A 3578-3278 cm^{-1}**	**A 1079-1003 cm^{-1}**
Бутиламин	1	0	0.13	0.06
2-Метилхексанол	44	0	0.38	0.31
Холестерол	8902	0	0.35	0.46
2-Метил-2-(метиламино)пропиофенон	3105	0	0.18	0.24
2-Метилен-1-бутанол	788	1	0.84	1.00
4-Фенокси-1-бутанол	1119	1	0.24	0.48
3,5-Дихлоробензенетанол	3124	1	0.42	0.56
2,6-Диметилбензенетанол	3171	1	0.70	1.00
Изобутанол	12307	1	0.68	0.93
Центроид на клас 0	-	0	0.26	0.27
Центроид на клас 1	-	1	0.58	0.79
X_1: 2-Пропиламин	5	0	0.24	0.18
X_2: Глецерин (1,2,3-пропантриол)	9265	1	1.00	0.70

Приведеният пример бе нарочно избран да е в двумерното пространство, за да може да се визуализира концепцията на метода, но класификацията обикновено се провежда в многомерното пространство. Интересно е да се

отбележи, че образ 6 от обучаващата извадка, *4-Феноксu-1-бутанол*, е по-близо до центроида на нулевия клас, въпреки че образът е от първи клас.

Както видяхме методът е приложим и за класификация към няколко класа – непознатият образ се избира от този клас, до чийто центроид разстоянието му е най-малко. В най-елементарните приложения всеки клас може да е съставен от само един образ. Съвсем нелишено от достоверност е твърдението, че човек интуитивно използва този метод за класификация на обектите в природата – обикновено с ежедневния опит си изграждаме представа за средния по характер обект и един непознат обект се причислява към тази група, чийто среден обект е най-подобен на непознатия.

4.4. Разделяща повърхност. Ако класовете в обучаващата извадка са само два, *то класифициращият критерий* (decision criterion) за непознатия образ **X**, y(**X**), може да се запише по следния начин:

$$y(\mathbf{X}) = \tfrac{1}{2}\,[d(\mathbf{X}, \boldsymbol{C}_1)^2 - d(X, \boldsymbol{C}_2)^2]$$

ако y(**X**) < 0, то **X** е от клас 1, ако y(**X**) > 0, то **X** е от клас 2, и ако y(**X**) = 0, то не може да се вземе решение; *C* **d(·)** е означено разстоянието между образа и двата центроида.

Ако се използва Евклидово разстояние, то *класифициращият критерий* е равен на:

$$y(X) = (1/2)[\sum_{n=1}^{N}(x_n - c_n^{(1)})^2 - \sum_{n=1}^{N}(x_n - c_n^{(2)})^2] = \sum_{n=1}^{N}(c_n^{(2)} - c_n^{(1)})x_n + (1/2)[\sum_{n=1}^{N}c_n^{(1)^2} - \sum_{n=1}^{N}c_n^{(2)^2}]$$

Групираме първата разлика и втората разлика, развиваме квадратите и получаваме, че критерият е равен на

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

където

$$w_n = (c_n^{(2)} - c_n^{(1)}) \quad \text{и} \quad w_{n+1} = (1/2)[\sum_{n=1}^{N} c_n^{(1)^2} - \sum_{n=1}^{N} c_n^{(2)^2}]$$

Ако положим в уравнение (4.2) Y(X) = 0, което е условие точката в хиперпространството да лежи на разделящата равнина, то ще получим уравнението на тази хипер-равнина в многомерното пространство. В тримерното пространство тази хипер-равнина представлява обикновенна равнина, в двумерното – права[1], а в едномерното – точка. Поради тази разделяща равнина, методът се причислява към линейните методи за класификация, към които спада и методът на линейната обучаваща машина, който е разгледан в една от следващите лекции.

Линейните методи се затрудняват с класификацията на линейно-неразделими данни [2], затова много от съвременните приложвния използват нелинейни методи, каквито са изкуствените невронни мрежи [3] (вижте следващите лекции).

ЛИТЕРАТУРА

1. K. Varmuza; *Chemometrics*. Springer Verlag, Berlin, 1980.
2. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michote, L. Kaufman; *Chemometrics: A Textbook*. Elsevier, Amsterdam, 1988.
3. T. Kohonen; *Self-Organization and Associative Memory*. Springer Verlag, Berlin, 1988.

Задачи

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 (1)? Ако да, за коя равнина става въпрос в уравнение (1)?

[1] Вижте задача С4 от семинар 2.

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?

Задача 4.3. Имате два четиримерни образа

$$X_1 = (0.7, 0.2, -0.1, 0.8) \text{ и } X_2 = (-0.6, 0.1, 0.2, 0.5)$$

a) Изчислете сумата им $S = X_1 + X_2$ b) Изчислете разликата им $D = X_1 - X_2$

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

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, който отговаря на центроида на тази извадка.

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?