Семинар 12

Линеен дискриминантен анализ

В този семинар ще се запознаем с линейния дискриминантен анализ (ЛДА), който се използва в статистиката, разпознаването на образи и обучението на машини. От обектите на обучаващата извадка се изчислява тегловен вектор, с който линейно се преобразуват признаците на образите, за да се получи нова променлива, която характеризира много по-добре разделянето на два или повече класа.

В този семинар ще разгледаме пример от компютърната класификация на ИЧ спектри. Ще изчислим тегловния вектор по малка обучаваща извадка. Примерът съдържа девет образа, разделени на два класа. Образите са двумерни и са изчислени от ИЧ спектри на органични съединения. Задачата, която е поставена е да се намери тегловен вектор, който разделя съединенията на два класа: първи клас на съединения, които съдържат третична бутилова група, $-C(CH_3)_3$, и втори клас – на съединения, имащи изопропилова група, $-CH(CH_3)_2$. В извадката всички съединения съдържат поне една от двете групи и няма съединения, които съдържат едновременно двете химични групи.

За ИЧ спектри на органични съединения е известно [1], че ивицата при 1380 cm^{-1} се дължи на симетричните деформационни трептения на метиловата група. Присъствието й показва, че в изследваната молекула има една или повече метилови групи. Когато ивицата при 1380 cm^{-1} е разцепена в дублет с приблизително еднакъв интензитет при около 1390 cm^{-1} и 1370 cm^{-1}, изследваното съединение притежава изопропилова група. Ако ивиците на дублета са при 1395/1365 cm^{-1} и имат различен интензитет (нискочестотната е по-интензивна), в молекулата на съединението има третична бутилова група. Единият признак използва отношението на интензитетите, което за

изопропиловата група е близко до единица, а другият – ширината на разцепване на ивиците, което според литературните данни, дадени по-горе, е по-голямо за третична бутилова група.

Практически задачи

Задача C12.1. Отворете файла `seminar12_lda.xls`. Разгледайте таблицата (sheet) "`Spectra`", в която са дадени девет двумерни образа, чиито спектрални признаци са изчислени от ИЧ спектри. Четири съединения притежават изопропилна група, докато останалите пет имат третична бутилова група. В ИЧ спектрите на деветте съединения ивица при 1380 cm^{-1}, разцепена на две. В колони `D` и `F` са дадени вълновите числа на двете ивици, а в колони `E` и `G` – техните абсорбции: всички спектри са нормирани в интервала 0.0 – 1.0 а.е.

Разгледайте таблицата (sheet) "`LDA`", в която съединенията са сортирани като първите пет са тези с третичната бутилова група. В колона `H` са изчислени отношенията на интензитетите (абсорбциите) на двете ивици, а в колона `I` – разликата между техните вълнови числа. В региона H11:I11 е изчислен центроидът на първия клас, а в региона H12:I12 – центроидът на втория клас.

Проверете формулите и повторете изчисленията в таблицата (sheet) "`WORK`".

Задача C12.2. Разгледайте таблицата (sheet) "`LDA`" във файла `seminar12_lda.xls`. В региона H13:I13 е изчислена разликата между първи и втори центроид, а в региона H14:I14 - средното между двата центроида.

Проверете формулите и повторете изчисленията в таблицата (sheet) "`WORK`".

Задача C12.3. Разгледайте таблицата (sheet) "`LDA`" във файла `seminar12_lda.xls`. В региона B16:C17 е изчислена ковариационната матрица на образите от първия клас. Ковариационната матрица на образите от втория

клас е изчислена в региона F16:G17. По надолу са изчислени техните обратни матрици, сумата от последните и теглата по следната формула.

$$\mathbf{w} = \overline{\Delta}\,(\Sigma_1^{-1} \text{ и } \Sigma_2^{-1})$$

Теглата са изчислени в региона B28:C28. Векторът $\overline{\Delta}$ в тази формула (в случая матрица с размерности 1 на 2) е разликата между двата центроида, която е изчислена в региона H13:I13. По-надолу, в региона B30:C30, е изчислен нормираният тегловен вектор.

Проверете формулите и повторете изчисленията в таблицата (sheet) "WORK".

Задача C12.4. Разгледайте таблицата (sheet) "LDA" във файла seminar12_lda.xls. В региона B30:C30, както споменхаме, е изчислен нормирания тегловен вектор. С този вектор в колона J е изчислена разделящата (дискриминационната) променлива.

Проверете и осмислете формулите и повторете изчисленията в таблицата (sheet) "WORK".

Задача C12.4. Разгледайте таблицата (sheet) "LDA" във файла seminar12_lda.xls. В региона B30:C30, е изчислен нормирания тегловен вектор. С този вектор в колона J е изчислена разделящата (дискриминационната) променлива.

Проверете и осмислете формулите и повторете изчисленията в таблицата (sheet) "WORK".

Разделя ли новата променлива добре двата класа? А оригиналните две променливи добре ли разделят двата класа? Ако е те се справят с тази задача, защо въобще е необходим ЛДА? Към последния въпрос ще се върнем в следващата задача.

Задача С12.5. Отново разгледайте таблицата (sheet) "LDA" във файла seminar12_lda.xls. В региона L2:R6 са дадени пет спектъра на пет съединения, които не са използвани в обучаващата извадка. Техните образи са изчислени подобно на тези за спектрите в обучаващата извадка – вижте региона S2:T6. За тези образи е приложен полученият тегловен вектор – резултатите са в колона U. Ако приемем за прагова стойност дискриминационната променлива на средното между двата центроида (клетка J14), то добре ли се разделят двата класа? Този въпрос е за тези пет тестови спектъра.

Проверете и осмислете формулите и повторете изчисленията в таблицата (sheet) "WORK".

Задача С12.6. Отново разгледайте резултатите от петте тестови спектъра в таблицата (sheet) "LDA" във файла seminar12_lda.xls. Видяхме, че дискриминационната променлива много добре класифицира образите и от обучаващата извадка и тези пет от тестващата извадка (ТИ). За обучаващата извадка (ОИ) и двете оригинални променливи - отношението на интензитетите и разликата между вълновите числа на двете ивици - работят много добре. Вижда се, че отношението на интензитетите за третичните бутили в ОИ е винаги по-голямо или равно на 1.28 (клетка H3), това за изопропилите по-малко или равно на 1.08 (клетка H7). За разликата между вълновите числа на двете ивици може да се каже същото – тя е по-голяма или равна на 24 (клетки I2 и I6) за третичните бутили и по-малка или равна на 20 за изопропилите.

Проверете с признаците на тестовите образи (колони S и T) можем ли да използваме тези прагови стойности. Към кой клас бихме класифицирали четвърти образ от ТИ, ако работим само с първи признак (колона S)? Към кой

клас бихме класифицирали първи и трети образ от ТИ, ако работим само с втори признак (колона T)?

Ако правилно сте отговорили на тези два въпроса, то изводът който може да се направи е, че дискриминационната променлива е по-надеждна от отделните признаци.

Задача C12.7. Разгледайте графиката в таблицата (sheet) "LDA" във файла seminar12_lda.xls. Координатните оси са двата признака x_1 и x_2, съответно отношението на интензитетите и разликата във вълновите числа. На нея са дадени деветте образа от ОИ (запълнените кръгчета и триъгълници), както и четири от петте тестови образа (празните кръгчета и триъгълници): петият образ, (3.44, 20), е извън рисунката. С червено са означени третичните бутили, а със синьо – изопропилите. Линията е прекарана по уравнението

$$x_2 = (w_1/w_2)\, x_1 + 15$$

Тази права линия е успоредна на правата[1] $x_2 = (w_1/w_2)\, x_1$, върху която се проектират образите[2], за да се получи дискриминационната променлива по уравнението

$$d = w_1 x_1 + w_2 x_2$$

На практика проекцията върху дадената на графиката права линия е еднаква с тази върху правата през началото на координатната система, $x_2 = (w_1/w_2)\, x_1$, върху която се проектират образите, за да се получи дискриминационната променлива по уравнението по-горе.

[1] Тази права линия има отрез, равен на нула, и затова минава през началото на координатната система.

[2] Проекция на даден образ върху дадена права означава от неговата точка да спуснем перпендикуляр към правата. Големината на проекцията се отчита от началото на координатната система, ако тази права минава през началото и.

Като имате предвид, че векторът $\mathbf{w}$ е с единична дължина, то какъв смисъл има d в уравнението по-горе?

Ако не можете да отговорите на този въпрос, помислете на какво е равно скаларното произведение между два вектора? А скаларно произведение, между единичен вектор и друг вектор?

Ако и това не се сещате, ето отговорите:

- Скаларно произведение между векторите $\mathbf{x}$ и $\mathbf{w}$ е равно на:

$$\mathbf{x}\ \mathbf{w} = |\mathbf{x}|\ |\mathbf{w}|\ \cos(\boldsymbol{\theta}) = \Sigma\ x_n\ w_n,$$

където $\boldsymbol{\theta}$ е ъгълът между тях, а величините с долни индекси са техните декартови координати.

- Ако векторът $\mathbf{w}$ е с единица дължина, то скаларно произведение между векторите $\mathbf{x}$ и $\mathbf{w}$ е равно на:

$$\mathbf{x}\ \mathbf{w} = |\mathbf{x}|\ |\mathbf{w}|\ \cos(\boldsymbol{\theta}) = |\mathbf{x}|\ 1\ \cos(\boldsymbol{\theta}) = |\mathbf{x}|\ \cos(\boldsymbol{\theta}) = \text{проекция на } \mathbf{x} \text{ върху } \mathbf{w}.$$