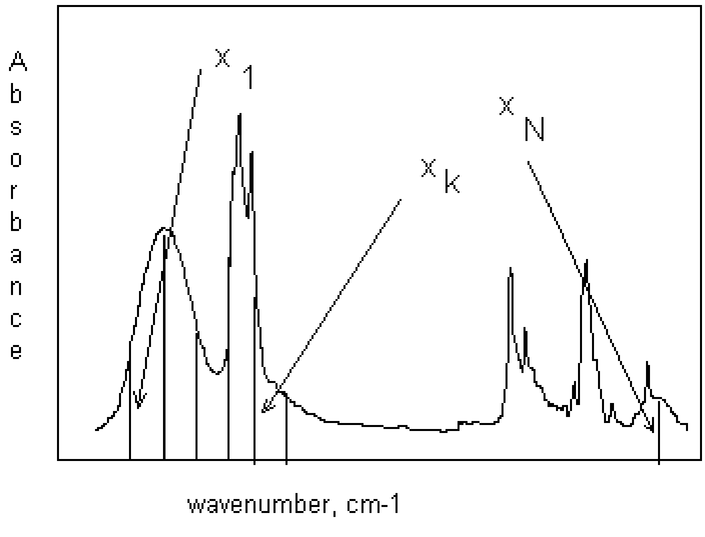
Лекция 3

Многомерно пространство на образите и разстояние в него

3.1. Химични обекти и събития, многомерни образи. Понятието *химичен обект* е една абстракция, която подобно на понятието модел включва най-важните характеристики (параметри) на даден реален обект или обектът е съвкупност от параметри, които обикновено се получават при поредица от измервания на различни величини, които характеризират обекта. Химичното събитие също се описва с параметри, но те се променят във времето.

Пример за *химичен обект* е ИЧ спектър на едно вещество. На фигура 3.1 е даден ИЧ спектър на *1-нонанол*. Ако целият спектрален интервал се раздели на N - 1 интервала, то стойностите на абсорбцията при техните граници (N на брой) представляват отделните признаци. Тези интервали могат да бъдат с еднаква големина (еквидистантни интервали) или с различна големина.

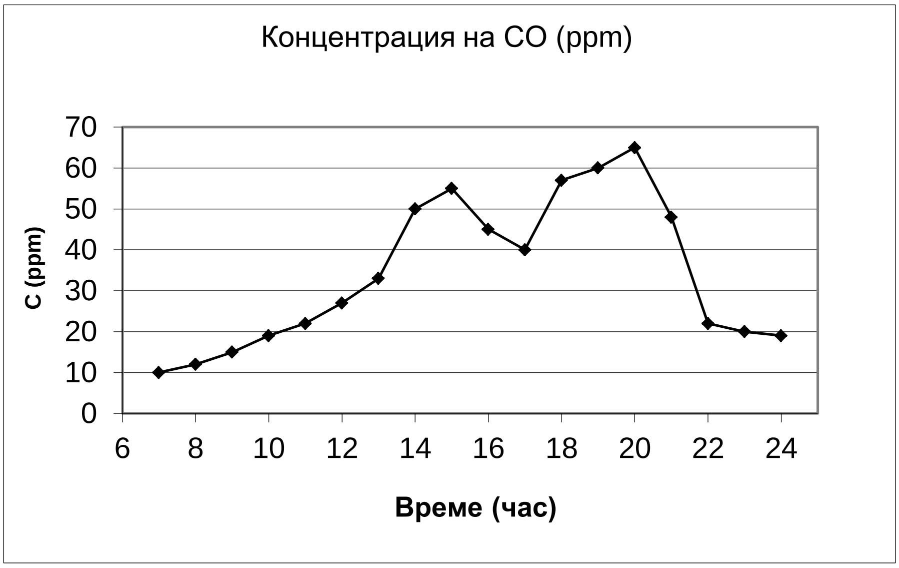


Фигура 3.1. ИЧ спектър на 1-нонанол. Височините на вертикалните линии, които отчитат величината на абсорбцията при определени вълнови числа се взимат за признаци на образа.

Един *химичен обект* или *събитие* се представя в *пространството на образите* като наредена енторка от числа, (x_1, x_2, … x_N), която се нарича ***образ*** (pattern), а пространството (най-общо то е N-мерно) се нарича ***пространство***

на образите (pattern space). Отделните числа, x_k, се наричат ***характеристики***, ***променливи*** или ***признаци*** на образа (***features, variables***). Всяко едно от тях представлява резултат от измерването на дадена величина, която характеризира химичния обект или химичното събитие.

Пример за поредица от *химични събития* е стойността на концентрацията на дадено вещество във въздуха, като функция от времето. На фигура 3.2 е дадена концентрацията на въглеродния моноксид на определено кръстовище в Пловдив през деня като функция от времето.

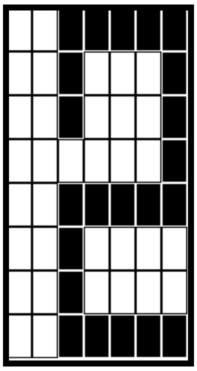


Фигура 3.2. Концентрацията на въглеродния моноксид като функция от времето.

В този случай отделният химичен образ може да се състави по няколко начина – например стойността на концентрацията в дадено време представлява едномерен образ, x_1. Но могат да се образуват и тримерни химични образи, съставени от три съседни концентрации – например химичния образ, който отговаря на 14 часа е съставен от концентрациите на CO, измерени в 13, 14 и 15 часа. Разбира се, че цялата графика, съставена от 18 стойности, може да се приеме за едно събитие, в този случай 18-мерно събитие, което характеризира замърсяването на кръстовището през този ден с въглероден моноксид.

Когато в пространството на образите се дефинира разстояние (вижте следващите точки), то химичният образ напълно се припокрива по свойства с понятието многомерен вектор, т.е. при преобразуване на координатната система, химичният образ се преобразува подобно на многомерен вектор. Във физиката многомерните вектори обикновено имат координати с размерност разстояние, докато химичните образи по правило са съставени от различни по размерност величини.

3.2. Двоични образи. Това са образи съставени от признаци, които могат да заемат само две стойности, например 0 и 1, или -1 и 1. Пример за такъв образ е графичното числено изображение да дадена чернобяла фигура, наречено ***цифрово двоично изображение или образ*** (digital binary image). На фигура 3.3 е дадено изображението на цифрата 2.



Фигура 3.3. *Цифрово изображение на цифрата 2 с разделителна способност 8 x 7.*

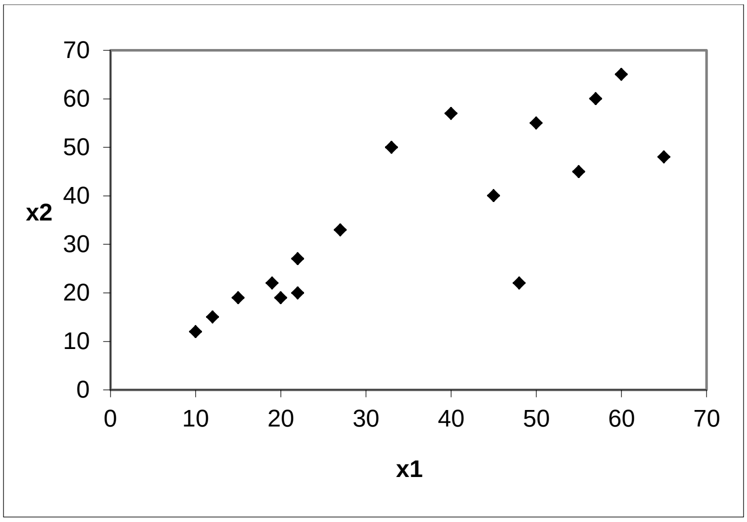
Изображението е с разделителна способност 8 на 7 точки (pixels), т.е. реално в цифрова форма то представлява една матрица с размери 8 реда (rows) на 7 колони (columns), чиито елементи $d_{r,c}$ са 1 (черен цвят) или 0 (бял цвят). В този случай това изображение може да се превърне по няколко начина в образ, като най-елементарният е признаците да се получат от матричните елементи чрез взимането на последните по редове. Общата формула в този случай е $x_{7(r-1)+c} = d_{r,c}$, т.е. $x_1 = d_{1,1}$, $x_2 = d_{1,2}$, ..., $x_{32} = d_{5,4}$, и т.н до $x_{56} = d_{8,7}$. За изображението на фигура 3 се получава следния двоичен образ, представен на фигура 3.4.

```
0011111
0010001
0010001
0000001
0011111
0010000
0010000
0011111
```

Фигура 3.4. *Двоичният образ на цифрата 2; векторът е представен в осем реда по 7 цифри като са изпуснати запетаите между отделните признаци, x_k.*

При някои от използваните методи нулата като стойност на признак е неудобно число, защото умножението с нея дава нула, и тогава тя се заменя с -1: в този случай двоичният образ е съставен от признаци, които приемат стойности само -1 и +1; в този случай образът се нарича *биполярен*, а не двоичен.

3.3. Пространство на образите. Ако на всеки признак, x_k, на хемометричния образ (x_1, x_2, … x_N) се съпостави една от координатите в N-мерното пространство, то хемометричният образ представлява един **N-мерен вектор**, а самите признаци са ***координатите*** на този вектор. Това пространство се нарича ***пространство на образи***те (pattern space). Например от концентрациите на въглеродния моноксид от фигура 3.2 могат да се съставят двумерни образи, които да бъдат изобразени в равнината. Един от двата разумни начина е следният: първият признак е концентрацията на моноксида един час преди събитието, а вторият признак е концентрацията на моноксида в часа на събитието. Така от поредицата от 18 концентрации се съставят 17 двумерни образи, които са изобразени в двумерното пространство на фигура 3.5.



Фигура 3.5. Двумерното пространство на хемометричните образи, съставени от събитията от фигура 3.2.

3.4. Мерки за разстояние в пространството на образите. Ако признаците на хемометричния образ описват химичния обект (или събитие) химически състоятелно, то образите на подобни в химично отношетние обекти ще се намират близко в съответното пространство. Образите от фигура 3.5 са изобразени в така нареченото *Евклидово пространство*, където мярката за разстояние, D_E, се изчислява по позната формула (3.1), в която разстоянието е равно на корен квадратен от сумата от квадратите на разликите между съответните координати на образите X и W.

$$D_E = \left(\sum_{k=1}^{d} \left(x_k - w_k \right)^2 \right)^{1/2} \tag{3.1}$$

Евклидовото разстояние е това разстояние, с което Еволюцията е дарила Човек (а може би и другите животни) като способност да го оценява между него и недалечни обекти – разбира се, това е разстояние в тримерния свят: d = 3 в уравнение (3.1).

В едно *линейно пространство* математиците могат да дефинират различни мерки за разстояние между точките на това пространство. Но всички тези мерки трябва да отговарят на четири изисквания, които читателят може нагледно да си представи за Евклидовото разстояние: то е неотрицателно, не зависи от посоката, то е нула тогава и само тогава, когато точките съвпадат и се подчинява на правилото на триъгълника.

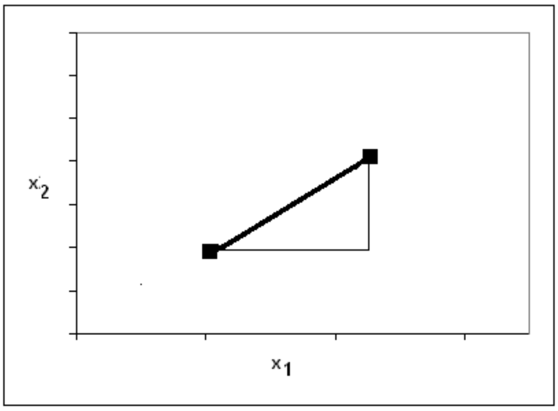
Строго дефинирани горните изисквания за едно N-мерно векторно пространство са следните: Ако $\boldsymbol{X_K}$, $\boldsymbol{X_L}$ и $\boldsymbol{X_M}$ са три кои да е вектора в пространството, то всяка реална функция от тях D($\boldsymbol{X_K}$,$\boldsymbol{X_L}$), която изпълнява следните отношения, може да се приеме като мярка за разстояние:

1. D($\boldsymbol{X_K}$, $\boldsymbol{X_L}$) ≥ 0; (мярката е неотрицателна).
2. D($\boldsymbol{X_K}$, $\boldsymbol{X_L}$) = D($\boldsymbol{X_L}$, $\boldsymbol{X_M}$); (мярката не зависи от посоката на преминаване на разстоянието).
3. D($\boldsymbol{X_K}$, $\boldsymbol{X_L}$) = 0 само ако $\boldsymbol{X_K}$ = $\boldsymbol{X_L}$; (тя е нула само за съвпадащи вектори).
4. D($\boldsymbol{X_K}$, $\boldsymbol{X_L}$) + D($\boldsymbol{X_L}$, $\boldsymbol{X_M}$) ≥ D($\boldsymbol{X_K}$, $\boldsymbol{X_M}$). (неравенство на триъгълника – сумата от две от страните му е винаги по-голяма от третата или равна ако трите точки лежат на една права).

Една обща мярка за разстояние (която се подчинява на горните изисквания) е разстоянието на Минковски, формула (3.2), в която M може да заема цели положителни числа – 1, 2, 3 и т.н.

$$D_M = \left(\sum_{k=1}^{d} \left| x_k - w_k \right|^M \right)^{1/M} \qquad (3.2)$$

За M = 2 получаваме Евклидовото разстояние, а при M = 1 разстоянието се нарича разстояние в Манхатан (*Manhatten distance* или *city block distance*). На фигура 3.6 са дадени двете разстояния, Евклидовото е показано с удебелена линия, а това в Манхатън с две тънки линии, сумата от чиито дължини дава разстоянието.



Фигура 3.6. Евклидово разстояние (удебелената линия) и разстояние в Манхатън (сума от другите две линии) в двумерното пространство.

За образи, двоично кодирани с нула и единица се използват две други мерки за разстояние, разстояние по Хеминг (Hamming distance), формула (3.3) и разстояние по Танимото (Tanimoto distance), формула (3.4).

$$D_H = \sum_{k=1}^{d} x_k \ xor \ w_k \tag{3.3}$$

$$D_T = 1 - (\sum_{k=1}^{d} x_k \ and \ w_k)/(\sum_{k=1}^{d} x_k \ or \ w_k) \tag{3.4}$$

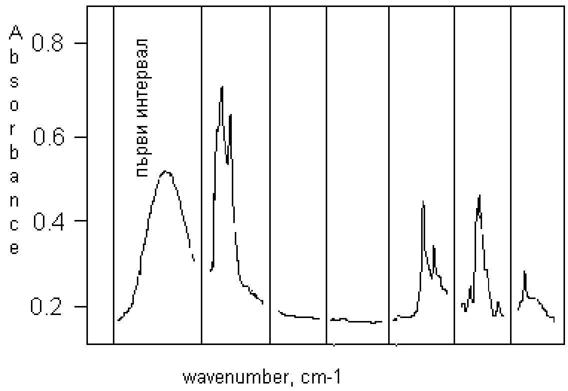
където x_k и w_k са к-тите координати на двоичните образи X и W, а функциите ***and***, ***or*** и ***xor*** са съответните логически функции, чиито значения са дадени в таблица 3.1. Стойността 0 отговаря на лъжа, а 1 – на истина. Тогава може лесно да се запомни, че само ако двете твърдения са истина, съставното твърдение с оператора „и“ ще е истина. Също така - само ако двете твърдения са лъжа, съставното твърдение с оператора „или“ ще е лъжа.

Таблица 3.1. Значение на логическите функции ***and***, ***or*** и ***xor***.

x_k	w_k	x_k and w_k	x_k or w_k	x_k xor w_k
0	**0**	**0**	**0**	**0**
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

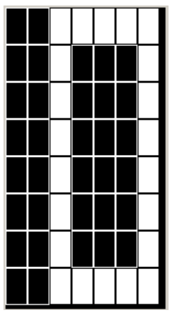
Задачи

Задача 3.1. Спектърът от фигура 1 е разделен на 7 интервала с различна дължина. Определете на око приблизителната стойност на седемте признака, които представляват абсорбцията на максималния пик в интервала.



Задача 3.2. Колко петмерни образа биха се получили от 18-те данни от фигура 3.2, ако се взимат четирите измервания предхождащи часа и измерването в часа. Кои са другите възможни начина при взимане на пет поредни измервания на концентрацията? Съответно колко образа се получават при всеки начин на съставяне на образите?

Задача 3.3. Кодирайте двоично цифрата 0, дадена на изображението по-долу. Колко е размерността на изображението? Колко е размерността на образа? Колко е стойността на x_{12}, x_{23} и x_{49}?



Задача 3.4. Какъв е химическият смисъл на диагоналната линия (минаваща през точката с координати (0, 0) и точката (70, 70) на фигура 3.5? Каква е

разликата между точките над нея и под нея? Какъв най-общ признак обединява точките в близост до началото на координатната система? Каква е разликата между тях и точките в близост до точката (70, 70)?

Задача 3.5. Имате три двоични петмерни образа, X_1 = (1, 1, 0, 1, 0), X_2 = (0, 0, 1, 1, 0) и X_3 = (0, 1, 1, 1, 0). Изчислете следните разстояния между тях: D_1, D_2, D_H и D_T. [за означенията вижте формули (3.1) – (3.4)]

Задача 3.6. Кое от разстоянията на Минковски (M = ?) отговаря на разстоянието по Хеминг? Какво показва разстоянието по Хеминг?

Задача 3.7. Съществуват още логически функции, освен тези дадени в таблица 3.1. Двете по-известни от тях са *импликация*[1] (x → w) и *еквивалентност*[2] (x = w). Техните стойности са дадени в таблица 3.2. Всички логически функции могат да се изразят с трите функции **and**, **or** и **not**: тези три функции е прието да се записват[3] с означенията Λ, V и ⌉. Последната функция действа на един операнд и ако той е 1 го променя на 0, ако той е 0 – на 1: ⌉1 = 0 и ⌉0 = 1.

Таблица 3.2. Значение на логическите функции ***импликация*** и ***еквивалентност***.

x_k	w_k	$x_k \rightarrow w_k$	$x_k = w_k$	⌉ x_k	⌉ x_k or w_k	⌉(x_k xor w_k)
0	0	1	1			
0	1	1	0			
1	0	0	0			
1	1	1	1			

[1] „Импликация" може да се преведе на български като „следствие" - изразът **x** → **w** се чете „от **x** следва **w**". Ясно е, че този израз няма да е верен само, ако **x** се е случило, а **w** не. За студентите е объркващо защо стойността на 0 → 1 е 1, но това може да се тълкува като „въпреки че **x** не се е случило, то други предпоставки са довели до **w**".

[2] Еквивалентност между две твърдения (x = w) имаме, ако логическите стойности на двете твърдения са равни. Ще видим от задача 3.8 и другото значение на тази операция.

[3] Изключващото или, **xor**, се означава като **or** с буквата ве, но с точка отгоре - $\dot{\vee}$. На български логическата функция **not** се нарича *отрицание*.

Вашата задача: Изчислете в пета, шеста и седма колона съответните стойности на логическите операции – използвайте таблица 3.1. Сравнете шеста с трета и седма с четвърта колона и вижте, че $x \rightarrow w$ е еквивалентно[4] на $\rceil \mathbf{x} \vee \mathbf{w}$ и че $\mathbf{x} = \mathbf{w}$ има същите стойности като $\rceil(\mathbf{x} \dot{\vee} \mathbf{w})$.

Задача 3.8. Попълнете четвърта, пета и шеста колона на таблица 3.3. Сравнете стойностите в трета и шеста колона и покажете, че $\mathbf{x}_k = \mathbf{w}_k$ е еквивалентно на $(\mathbf{x}_k \rightarrow \mathbf{w}_k) \wedge (\mathbf{w}_k \rightarrow \mathbf{x}_k)$. Тази еквивалентност[5] илюстрира изучаваното в училище твърдение „**x** е изпълнено тогава и само тогава, ако **w** е изпълнено". Това твърдение се нарича още „изпълнението на **w** е необходимо и достатъчно условие за изпълнението на **x**". (Ясно е, че в предните две изречения **w** и **x** могат да разменят местата си.)

Таблица 3.3. Значение на някои логически функции.

$\mathbf{x}_k$	$\mathbf{w}_k$	$\mathbf{x}_k = \mathbf{w}_k$	$\mathbf{x}_k \rightarrow \mathbf{w}_k$	$\mathbf{w}_k \rightarrow \mathbf{x}_k$	$(\mathbf{x}_k \rightarrow \mathbf{w}_k) \wedge (\mathbf{w}_k \rightarrow \mathbf{x}_k)$
0	0	1			
0	1	0			
1	0	0			
1	1	1			

[4] Подобно на събирането и умножението, операцията „или" има по-нисък приоритет пред операцията „и". Най-първо, обаче, се извършва „отрицанието", ⌉, което има най-висок приоритет.

[5] Очевидно, еквивалентността на две твърдения води до това че ако е изпълнено едното твърдение, то е изпълнено и второто, както и ако е изпълнено второто твърдение, то е изпълнено и първото.