Лекция 5
Хемометрични методи – обзор. Структуриране на данните

5.1. Класификацията на хемометричните методи. Тази класификация може да се извърши съобразно различни критерии:

- Според знанието за статистическото разпределение на образите:
  - Параметрични методи (*parametric methods*)
  - Непараметрични методи (*nonparametric methods*)

Представа за параметричните методи може да се добие при разглеждане на статистическите хипотези и тяхната проверка, които се извършват с различните статистически разпределения. Става въпрос за едномерни статистически разпреледеления, т.е. такива, които описват разпределението на <u>една случайна величина</u>. Основа за този начин на проверка на статистическите хипотези е допускането, че (1) данните от измерването на дадена величина, X_1, X_2, … X_N, са нормално разпределени с едни и същи параметри на разпределение – математическо очакване μ и дисперсия σ^2, (2) стойностите на измерванията са независими едно от друго, т.е. имаме N на брой, еднакво, но независимо разпределени случайни величини. Тогава различните критерии, които се изчисляват, са разпределени по определен начин и с определени параметри [1].

Непараметричните методи не предполагат знание за разпределението на многомерните образи. (Има и непараметрични методи за проверки на хипотези при едномерни данни.) Естествено е да очакваме, че образите са разпределени с многомерно нормално разпределение, но това не е така, а и да е така, не може да се провери с голяма достоверност. Причина за невъзможността за проверка обикновено е голямата размерност на образите и недостатъчният техен брой. Съществува правило, че броя на образите трябва да е поне три пъти по-голям от тяхната размерност [2], и въпреки че

обикновено това съотношение е много по-голямо от 3, например 1000, то многомерността на пространството довежда до "разреждане" на образите в него и от там до недостоверност на всяка една хипотеза за принадлежност на данните към някакво разпределение.

Всички методи, разгледани в дисциплината, могат да се прилагат без да имаме информация за разпределението на признаците на образите.

- Според зависимостта, която се търси между обектите и техните признаци:
  - Методи за изобразяване (display methods): това е методът анализ на главните компоненти (PCA, principle component analysis).
  - Методи за кластерифициране (analysis of clusters): това е методът на йерархичния кластерен анализ.
  - Регресионни методи (regression analysis): такива методи от изучаваните в този курс са методи линейна многопроменлива регресия и изкуствените невронни мрежи (ИНМ) с право разпространение на сигналите и обратно разпространение на грешките.
  - Методи за класифициране (classification methods): това практика са почти всички методи, които се разглеждат в тези лекции.
- Съобразно знанията за класовете на обектите
  - **Методи с известни класове при обучението (*supervised learning methods*).** Към тази **категория** спадат линейната обучаваща машина, методът на центроидите и линейната многопроменлива регресия. Не всички от моделите на ИНМ спадат към тази категория, но някои от моделите могат да се използват само при известни класове – например ИНМ с *право разпространение на сигнала*.

- **Методи с неизвестни класове при обучението (*unsupervised learning methods*)**: това са методите кластерен анализ и PCA и част от моделите на ИНМ – например ИНМ на Кохонен (T. Kohonen) [3].

5.2. Прилагане на хемометричните методи. Основното допускане в хемометрията, което се подразбира при прилагане на нейните методи гласи:

Ако два химични обекта са еднакви или близки по отношение на някакво тяхно свойство (характеристика), то от техните други свойства (характеристики) могат да се съставят образи, които са близки в пространството на образите

Ето защо при повечето приложения целта е да предскажат едно или повече химични свойства въз основа на набор от други характеристики.

В науката връзката между свойствата на химичните обекти се описва като връзка между стойностите на еднозначно дефинирани (физико)химични величини, и тази връзка се определя експериментално при анализ на набор от (физико)химични измервания, които се обработват математически в рамките на някакъв предварително възприет модел. Пример за такава връзка между физикохимични величини е зависимостта между обема, V, температурата, T [K], и налягането, p, на идеалните газове:

$$pV = nRT$$

където R е така наречената универсална газова константа, R = 8.314 $Jmol^{-1}K^{-1}$.

Много често, обаче, между свойствата на химичните обекти няма еднозначна количествена връзка поради няколко причини.

- Не всички свойства (характеристики) могат да се изразят еднозначно: най-добър пример за това е степента на структурно подобие на една молекула към друга молекула.

- не съществува строга количествена връзка между величините, които описват свойствата на химичните обекти, въпреки видимата корелация (връзка) между тях – например заснетият масспектър отразява (зависи от) броя на въглеродните атоми на съединението, но няма еднозначна математическа зависимост, която количествено да изразява тази връзка.
- количественото изразяване на някои от свойствата на химичните обекти зависи от редица условия, които ученият не може или му е трудно да контролира или даже не подозира за тях – пример за това са спектрите, чийто вид зависи от условията при които, и апаратурата на която, се заснимат.

Липсата на споменатата еднозначна връзка между свойствата води до необходимостта от получаване на приблизителна връзка по следния начин:

- определят се всички интересуващи ни химични обекти, които са част от множеството на всички възможни (или всички налични) обекти.
- за тях се определят признаците, които изграждат химичните образи на тези обекти

По този начин се съставя една ***извадка*** (набор) от химични образи, която може да се представи като матрица: на всеки ред отговаря един химичен образ, а на всяка колона – един признак. Тази извадка се нарича на английски ***data set – извадка от данни***. Част от обектите (образите) се избират (най-вече случайно) за да може да се установят зависимостите между изследвания признак (който интересува изследователя) и другите признаци (които на практика съставят образа), а останалата част от обектите се използва за проверка на получените емпирични зависимости. С първия набор от образи се извършва установяване на приблизителната връзка и този процес обикновенно се нарича ***обучение*** (*learning*), а образите използвани за обучението съставят така наречената ***обучаваща извадка*** (*learning set*).

Останалите образи се използват за проверка на валидността (надеждността) на получената зависимост. Понякога втората извадка допълнително се разделя на две части: (1) ***тестваща извадка*** (*test set*) и (2) ***валидираща извадка*** (*validation set*).

Първата от тези извадки се използва за изчисляване на някои статистически величини, които характеризират разпределението на изчислената по получения модел числова характеристика, която описва интересуващото ни свойство. Валидиращата извадка проверява общовалидността на получените оценки на статистическите величини.

5.3. За данните и не само за тях. Измерванията струват пари и това води до ограничен набор от експериментални данни. Освен това наличните данни, за съжаление, имат редица неприятни характеристики, като най-неприятните сред тях са:

- наличие на липсващи и непълни данни, а в някои случаи те са финансово недостъпни или просто ограничени по брой;
- наличие на сгрешени данни;
- наличие на бегълци (outliers) сред данните, т.е. данни, чийто признаци съществено се отличават от признаците на другите данни;
- наличие на корелация между различните величини (променливи), които описват данните;
- данните са подредени системно, което е следствие от системния подход при тяхното генериране (експериментално получаване): по същата причина и обхватът на данните е силно ограничен и неравномерно разпределен по класове;
- признаците (променливите) са с различна размерност и/или с различен динамичен обхват (различават се с порядъци в тяхната стойност)

Всички тези недостатъци в извадката от реални експериментални данни за химичните обекти водят до факта, че изборът на обучаващата, тестващата и валидиращата извадка, както и на признаците (променливите) играе голяма роля в установяване на различните количествени зависимости в Хемометрията. В науката няма абсолютна истина, освен може би, фундаменталните физични закони и фактите за наличие на атоми, полета, материя, заряди и подобни представи. В науката решаваща роля играе моделът, който описва зависимостите между избраните от учените величини. А "работещият" модел има три основни характеристики – (1) обяснява много голям брой наблюдавани факти, (2) предсказва нови факти, и (3) има ясно физическо тълкуване (физически смисъл). Читателят трябва да разбере, че науката изгражда и използва модели, които се проверяват с *експериментални наблюдения*, които са два вида – контролирани от учените, наречени *активен експеримент* и неконтролирани от учените – т.н. *пасивен експеримент*. Има и научни модели, например теорията за Еволюцията, които са следствие от провеждането на пасивни и активни експерименти. Съществува схащане сред научната общност, че само с пасивни наблюдения не може да се установи причинно-следствена връзка, т.е. наблюдаваната корелация между стойностите на две променливи, може да не е следствие от тяхната явна (причинно-следствена) зависимост, а следствие от зависимостта им от друга величина или величини или просто да е намесена случайността.

Като цяло Науката се справя успешно с поставените и от обществото задачи, т.е. нейните модели "работят", това ще рече, че моделите на Науката предсказват, обясняват, спестяват пари с тези предсказания, достоверни са с някаква статистическа сигурност, постоянни са в предсказанията си, ясни са (макар и понякога за ограничен брой учени) и като цяло винаги са

приближени и не претендират за вездесъщност (приложимост при всички случаи). Добрият учен, инженер или техник винаги знае кога моделът може да се приложи и кога не, каква е неговата точност, възпроизводимост и устойчивост. Всички други подходи извън Науката са чисто и просто шарлатания, която често се оправдава с изрази от рода на "науката е безсилна в този случай". Да, Науката е безсилна в много случаи, и ако се открие *нещо* по-добро, *което* отговаря на гореописаните характеристики, *което* дава по-добри резултати от досегашните модели, *което* експериментално може да се провери, то *това* ще е пак ... *научен модел*.

5.4. Конструиране на обучаваща, тестваща и валидираща извадки. Първият важен етап в тази дейност е изборът на химичните обекти. Желателно е те да са много на брой, да са разнообразни по природа, да са представителни за проблема, който се решава, и не на последно място, данните за тях да са взети от достоверни източници. Съвременните компютри, с тяхната оперативна и дискова памет и скорост на изчисление, доведоха до отпадането на изискването за предварително ограничаване на обема на извадката (броя данни) и нейната размерност (броя признаци) – това е така поне в повечето от поставените съвременни задачи.

Изборът на обекти се влияе и от претенциите на модела – дали е *глобален модел* (например, отнася се за всички химични съединения) или *локален* (например, прилага се само за даден клас от химични съединения). И при двата вида модели трябва да се вземе представителна извадка от обекти, в първия случай – представителна за „всички" обекти, във втория – представителна за ограничения клас от обекти. Понятието „представителна извадка" интуитивно се разбира като случайно избрана от всички представители на класа: при едно такова случайно избиране отношението на броя на различните подкласове се запазва – например ако в цялата извадка

има два пъти повече бензени отколкото алкани, то и в избраната извадка това отношение трябва да е приблизително две. Един от проблемите при глобалните модели е определянето на *това*, което би трябвало да бъде *генералната съвкупност*. Например, в Chemical Abstract Service (CAS) са регистрирани (към януари 2014 г.) около 73 милиона химични съединения и техният брой постоянно нараства. Почти никоя научна група няма пълни данни за техните (физико)химични свойства, а това естествено води до ограничен избор. Друг проблем е, че повечето от тези изолирани от природата съединения са подобни по структура, защото така „работи" Природата - впечатляващо е как Еволюцията е избирала винаги "лесния път". Синтезираните от Човек сред тях са още по-подобни, защото лесно се синтезират аналози, защото много изследователски групи работят по една или две тематики и т.н. Тогава, до колко даже и всички регистрирани химични съединения са представителни е спорно! Нахвърляните по-горе проблеми оставаме без решение, не само защото то е извън обсега на тази дисциплина, а и защото и авторът не знае най-доброто такова. Но остава препоръката – избират се колкото се може повече химични обекти, за които има достоверни данни.

Предполага се, че за химичните обекти могат да се намерят количествени характеристики, които ще бъдат превърнати в признаци (променливи). Тези променливи, въпреки че са случайни величини, трябва еднозначно да са определени (с метода за измерването им, в начина им на тълкуване и пр.) – в противен случай никакви реално-действащи зависимости няма да могат да се открият. За всеки един обект от извадката по един и същ начин се съставя набор от признаци. Например, не може концентрацията на олово в пробата да е първи признак в едни обекти, а в други на първо място да стои концентрацията на магнезий. Тези признаци на обекта представляват числа

и изграждат образа му: това е наредена ен-торка от числа. В много от случаите тези променливи имат различна размерност и/или различен динамичен обхват, и затова трябва да се нормират по определен начин (вижте т. 5.5 от настоящата лекция).

Тъй като Науката се занимава не само с обясняване на съществуващите факти и зависимости, а предсказва нови такива, то моделите в хемометрията се изграждат с една извадка, а се проверяват с друга. Както споменахме по-горе, първата е т.н. обучаваща извадка, а втората – тестваща извадка. И двете извадки трябва представително да отразяват състава на първичната извадка, т.е те трябва да са избрани случайно.

Най-грешният избор на образи в тях, е да се избере първата половина от наличните образи в едната извадка, а втората половина да състави другата извадка. Обикновено базите от данни, пък и колекциите от данни, са съставени от цели блокове от данни на подобни съединения, да не говорим че понякога те са сортирани предварително (например по молекулна маса или молекулна формула).

Случайният избор на голям брой от обекти (образи) се извършва с компютърни алгоритми, които пък използват псевдослучайни числа. Нека си представим, че имаме 10 000 образа и трябва да изберем случайно половината в обучаващата, а другата половина в тестващата извадка. Един елементарен алгоритъм е да генерираме случайни числа между 1 и 10 000 и генерираното число да показва номера на образа, който трябва да вземем в обучаващата извадка. След случайната генерация на повече от 5000 числа (защото сред тях ще има повторения), най-накрая ще получим точно 5000 различни числа, които ще показват номерата на образите, които съставят обучаващата извадка: естествено останалите номера са на образите в тестващата извадка. В епохата на „дървените компютри и железните

програмисти"[♦] на първите алгоритъмът ще е доста бавен, а вторите определено ще се ужасят от неговата неефективност, но най-важното е, че на практика в обучаващата извадка ще има цели групи от обекти, последователни в първичната извадка, т.е. ще има образи на много подобни обекти в обучаващата извадка. Много по-просто и надеждно е мислено да разделим на двойки първоначалната извадка (т.е. 1 и 2 образ, 3 и 4 образ и т.н.), да генерираме случайно отново число между 1 и 10 000 и ако то е по-малко от 5 000 да вземем първият образ от съответната двойка в обучаващата извадка. Защо навлязохме в подробности? Защото с втория алгоритъм, който е значително по-бърз от първия, става ясно че един от всеки два съседни образа ще е в обучаващата извадка, а останалият – в тестващата. А ние точно това искаме – съставът на двете извадки да е подобен, защото с първата откриваме зависимост, а с втората извадка проверяваме дали тази зависимост важи за подобни съединения.

Вторият алгоритъм може да се приложи и за три извадки – разглеждаме тогава тройки от образи, поредни в първоначалната извадка. Един от тройката отива случайно в ОИ[♣], един в ТИ и един във ВИ, като се премахва неслучайната подредба на обектите в първоначалната извадка.

Засега тези извадки (ОИ, ТИ и ВИ) ви звучат абстрактно, но при всеки разглеждан метод те ще бъдат конкретизирани и изрично упоменавани.

Остава само да характеризираме що е това *валидираща извадка*. В хемометричната литература няма еднозначно определение, но в повечето случаи под ВИ се разбира извадка от образи (съответно обекти), които не участват в ОИ и ТИ. Ако с ОИ разкриваме вида на зависимостта, то с ТИ тази зависимост може да бъде статистически охарактеризирана – да се определи

[♦] изразът е от Интернет

[♣] ОИ – обучаваща извадка, ТИ – тестваща извадка, ВИ – валидираща извадка

надеждността на класификация чрез редица числени параметри. Един такъв директен числов критерий е процентът на правилно класифицираните обекти от ТИ, който процент се нарича *предсказваща способност (PA, prediction ability)*. Съответно, процентът на правилно класифицираните обекти от ОИ се нарича *разпознаваща способност (RA, recognition ability)*. Желателно е двете да са близки до 100%, но ако са по-ниски, то трябва да са приблизително равни – в противен случай сме открили зависимост, специфична за образите (обектите) в ОИ, и от нея няма полза за никаква друга класификация, която е различна от класификацията на образите в ОИ.

Когато PA близка по стойност до RA можем смело да твърдим, че сме открили зависимост, която е значително по-обща (по-обобщаваща, more generalizing). Дали тази зависимост важи за "всички" обекти? Това най-добре се проверява с трета извадка, наречена валираща извадка. Най-добре е тази извадка да е от данни, взети от източници, различни от източниците на ОИ и ТИ.

И така да обобщим: образите в ОИ, ТИ и, евентуално, ВИ се избират случайно. В никоя от тези извадки няма образ от друга извадка. В ОИ и ТИ може да има образи на подобни по между си обекти, но не и на еднакви. „Съставът" на ОИ и ТИ е по възможност разнообразен вътре в тях и подобен в средно между тях. ВИ е от образи, пожелателно взети от други независими източници.

5.5. Скалиране на данните в извадките. Първоначално признаците на отделните химични образи са взети по определен алгоритъм от различни характеристики на химичните обекти. Поради това тези признаци обикновено са с различна размерност, а тези които са с еднаква размерност може да имат различен динамичен обхват. Ето защо е необходимо признаците да се скалират в отделните извадки.

5.5.1. За необходимостта от скалиране на данните. Друга причина за скалирането на признаците е математическата особеност на използваните хемометрични методи. Например, повечето софтуери изчисляват коефициентите на линейна многопроменлива регресия с еднакъв брой значещи цифри. Например ако първи признак заема много малки стойности, а втори признак – много големи, то нормално е да се получат два коефициента на регресия с различна стойност, например a_1 = 1234.4567890 и a_2 = 0.0000870. В този случай вторият коефициент ще бъде само с три значещи цифри, което при прилагане на регресията ще доведе до предсказване на свойството (зависимата променлива У) само с три значещи цифри.

При нелинейните методи, като изкуствените невронни мрежи, обучението на класификатора ще отчете само признака с много голяма стойност, а останалите признаци „ще се загубят" сред различните суми, произведения и повдигания на степен. На практика това означава, че класификаторът ще работи само с признаците, които имат много големи стойности.

Има и други недостатъци, когато се използват признаци с различен динамичен обхват, но те са трудно обясними за начинаещи, които не са запознати с разнообразните методи и изчислителни подходи на хемометрията.

5.5.2. Традиционно скалиране на спектрални данни. В почти всички спектроскопии данните се скалират по определен начин, който е традиционен и който е следствие от желанието на спектроскопистите да сравняват спектри на различни проби (вещества или техни смеси). В спектралните колекции и библиотеки от ИЧ и Раман спектри, спектралните криви се представят нормирани в интервала 0 – 1 по ординатата. Това позволява на екрана на компютъра да могат да се сравняват спектрите на

неизвестната проба с тези на библиотечните съединения. В мас-спектрометрията на органични съединения пиковете при различните масови числа (по-строго казано не масово число, а отношението маса/заряд - m/z) се представят така че максималният пик да е 1.00 (100%). Това, както при ИЧ и Раман спектрите, се постига чрез формулата:

$$x'_{m,n} = (x_{m,n} - min_m)/(max_m - min_m);\ за\ m = 1, 2, \dots M\ и\ n = 1, 2, \dots N, \quad (5.1)$$

където $x'_{m,n}$ и $x_{m,n}$ са новия (скалиран) и стария (нескалиран) признак, m е номерът на образа (обекта), n – номерът на признака (имаме M образа, които са с размерност N), max_m и min_m са максималната стойност на ординатата в m-тия спектър. Обърнете внимание, че скалирането се извършва като се използват максималните и минимални стойности на признаците на даден образ, а не максималните и минимални стойности на даден признак, както е в т. 5.5.3. Тук трябва да се отбележи, че в мас-спектрометрията min_m се приема за нула, защото йонният ток на дадено отношение m/z е много еднозначно определен и неговата „нула" е също определена от апарата – просто няма йонен ток. При ИЧ и Раман min_m се взима от минималната стойност на абсорбцията (разсейването) в спектралния интервал, с който е представен спектърът в компютъра. Причина за това е наличието на т.н. базова линия (back-ground) на спектъра.

5.5.3. Интервално скалиране. При това скалиране признаците се преобразуват по формула, която е за всеки признак, а не за даден образ. Обърнете внимание на индекса на на „min" и „max" в (5.1) и (5.2)

$$x'_{m,n} = a + b(x_{m,n} - min_n)/(max_n - min_n);\ за\ m = 1, 2, \dots M\ и\ n = 1, 2, \dots N, \quad (5.2)$$

Числата a и b осигуряват скалирането на даден признак за цялата извадка в интервала (a, b). Ако a = 0 и b = 1 имаме скалиране в интервала (0, 1), ако a = -1 и b = 1 всеки признак се скалира в интервала (-1, 1). Недостатък на

интервалното скалиране е, че стойностите на min_n и max_n се взимат от бегълците (outliers) в извадката, затова това скалиране рядко се използва или се прилага, ако сме се уверили, че в данните нямаме бегълци. Това скалиране се използва при моделиране на експеримента, защото там изследователя определя интервалите на изменение на входните променливи (които играят един вид ролята на признаци) и така де факто няма бегълци – вижте лекция 9.

5.5.4. Автоскалиране. Това скалиране премахва донякъде влиянието на бегълците - то се извършва по формула (5.3).

$$x'_{m,n} = (x_{m,n} - m_n)/s_n; \text{ за } m = 1, 2, \ldots M \text{ и } n = 1, 2, \ldots N, \qquad (5.3)$$

където m_n и s_n са съответно средната стойност и стандартното отклонение на n-тия признак в извадката. Тези стойности се влияят донякъде от бегълците в извадката и особенно при наличие на голяма група бегълци. Но от формулите за средна стойност и стандартно отклонение е ясно, че ако има един беглец (по този признак), то неговото влияние намалява N пъти в средната стойност и "корен от N" пъти в стандартното отклонение. Вместо средна стойност може да се използва медианата, а вместо стандартното отклонение разликата между 75 квантил и 25 персентил, която казано по-друг начин е разлика между третия и първия квартил, а медианата е втория квартил.

Обърнете внимание, че формули (5.1), (5.2) и (5.3) правят новите (скалирани) признаци безразмерни!

5.5.5. Други методи за скалиране. Те се използват много често и най-вече поради специфичността на данните или количествените отношения между различните променливи и зависимата променлива.

Ако данните не са нормално разпределени, то е възможно използването на сигмоидално скалиране с крива от вида на уравнение (11.2) в лекция 11. Тази крива просто събира бегълците в двата си края към центъра.

Друго скалиране е логаритмичното, обикновено десетичен или натурален логаритъм, но се използва и двоичен логаритъм. Това скалиране е необходимо, когато се работи с данни в много голям обхват. За да добиете представа за необходимостта от него просто сравнете числата 1, 10, 100 и 1000 с техните десетични логаритми – 0, 1, 2 и 3.

В мас-спектрометрията се използва скалиране от вида (5.4) – обърнете внимание, че по разположението на индексите то прилича на (5.1), т.е. се отнася за даден обект, а не признак.

$$x'_{m,n} = x_{m,n} / \Sigma\, x_{m,n};\ \text{за } m = 1, 2, \ldots M \text{ и } n = 1, 2, \ldots N, \qquad (5.4)$$

където сумата е по n за n = 1, 2, … N (при постоянно m). Това на практика води до това, че сумата на всички признаци в даден обект е единица. Което за спектроскопистите означава, че сумарния йонен ток е еднакъв (равен на 1) за всички спектри в извадката.

В мас-спектрометрията се използва и скалиране, което дава единица дължина на вектора – (5.5).

$$x'_{m,n} = x_{m,n} / \mathrm{Sqrt}(\Sigma\, x_{m,n}^2);\ \text{за } m = 1, 2, \ldots M \text{ и } n = 1, 2, \ldots N, \qquad (5.5)$$

където сумирането е отново по n за n = 1, 2, … N (при постоянно m). Припомнете си, че $\mathrm{Sqrt}(\Sigma\, x_{m,n}^2)$ представлява нормата (дължината) на вектора (образа).

Към скалирането формално може да се отнесе и създаването на двоичен образ от оригиналния образ. Например за мас-спектрите йонният ток на всяко m/z, който е по-голям от, или равен на, даден праг (threshold) се приема за единица, а в обратния случай той се кодира с нула [2].

Литература

1. 1. Футеков Л., Пенчев П., "Теория на експеримента", Пловдив, Изд. ПУ, 1992, 1998.
2. П. Джурс, Т. Айзенауэр; *Распознавание образов в химии*. Мир, Москва, 1977.
3. K. Varmuza, P. Filzmoser; *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.

Въпроси

Задача 5.1. Имаме два класификатора, които от ИЧ спектър определят дали съединението е алкохол или не. Разпознаващите и предсказващи им способности са следните RA_1 = 80% и PA_1 = 60% и съответно RA_2 = 61% и PA_2 = 60%. Кой класификатор според вас е по-надежден?

Задача 5.2. Имаме два класификатора, които от ИЧ спектър определят дали съединението е алкохол или не. Разпознаващите и предсказващи им способности са следните RA_1 = 19% и PA_1 = 20% и съответно RA_2 = 61% и PA_2 = 60%. Кой класификатор според вас е по-надежден?

Задача 5.3. Ако в предишната задача сте определили, че вторият класификатор е по-надежден, то променете класификатор 1 в 1′ така: ако той предскаже от ИЧ спектъра, че съединението не е първичен алкохол, приемете, че то е, и обратно – ако предскаже, че първичен алкохол, приемете, че то не е. Какви ще са $RA_{1'}$ и $PA_{1'}$ в този случай? Кой класификатор според вас е по-надежден[1] сега, 1′ или 2?

[1] Задачата има общочовешко и донякъде комично приложение – всички сме срещали хора, чиито прогнози са почти винаги грешни (особено в спорта, политиката). Какво правим в този случай? Не им вярваме? Не, препоръката е просто обръщайте предсказанието им в противоположното.