Validacija modelov

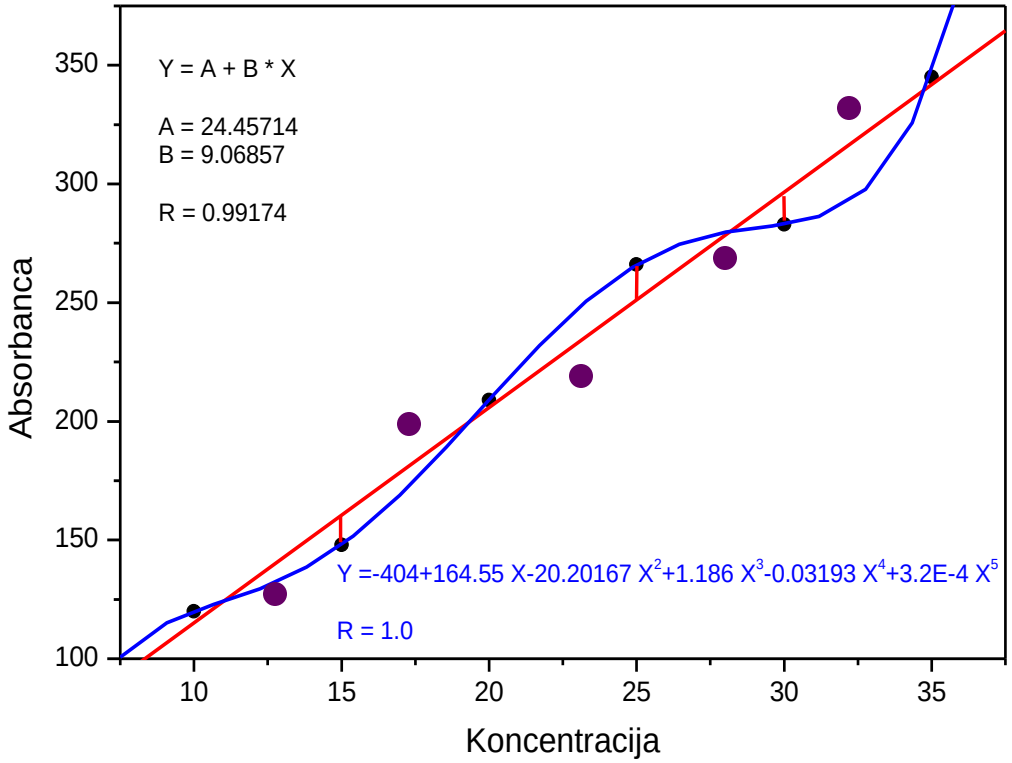
Validacija ali ovrednotenje modelov je izjemno pomembna faza v procesu modeliranja. Pri **deterministični modelih**, kjer je modelna funkcija **f(X),** ($X= x_1, x_2 \ldots x_i \ldots x_k$) vnaprej poznana (npr. premica, polinom, eksponentna funkcija), in imamo ustrezne eksperimentalne podatke oziroma imamo možnost vplivati na njihovo pridobivanje, ponavadi za ovrednotenje modelov uporabimo alanizo varianc (ANOVA). To validacijsko metodo smo že obdelali, nazoren primer je uporaba ANOVE pri kalibracijskih premicah.

Povsem drugačen način ovrednotenja pa je poteben pri uporabi **izkustvenih modelov**, ki ne temeljijo na vnaprej podani modelni funkciji, ampak se "naučijo" razpoznavati relacijo med neodvisnimi in odvisnimi več-dimenzionalnimi spremenljivkami **{X,Y}** na osnovi **učnega niza**, to je podatkovnega niza objektov, za katere poznamo vrednosti parov {X,Y}.

Med izkustvene modele prištevamo modele na osnovi umetnih nevronskih mrež, kot so nevronske mreže z vzratnim širjenjem napake (error back propagation), protitočne nevronske mreže (counterpropagation), mreže na osnovi radialnih baznih funkcij (RBF networks) in razne variacije le-teh. Izkustveni se ti modeli imenujejo zaradi postopnega vnašanja informacij o povezavi med neodvismimi in odvisnimi spremenljivkami X in Y glede na objekte učnega niza. Z drugo besedo tak postopek imenujemo "učenje", kar pomeni pri umetnih nevronskih mrežah iterativno (postopno) prilagajanje uteži v nevronih pri nekem določenem nizu podatkov (objektov). Če je učenje "nadzorovano", se uteži prilagajajo odgovorom oziroma tarčam, ki jih poznamo za vsak podatek (objekt).

Poznan problem pri izkustvenih modelih je **pretreniranje**, kar pomeni, da se **preveč** dobro **prilagodijo** podatkom iz učnega podatkovnega niza. Podatki iz učnega niza, pridobljeni na realnih vzorcih oziroma problemih, nikoli niso idealni in vsebujejo večje ali manjše odmike od pravih vrednosti. Nanje vplivajo razni dejavniki, ki jih ne moremo kontrolirati (ozadje, matriks, nestabilnost instrumenta) in jih opišemo kot **šum**. Pretreniran model torej ne popisuje le relacij parov **{X,Y}**, ampak tudi posnema šum v podatkih. Zato zelo dobro napoveduje lastnost (**Y**) objektov iz učnega niza, medtem ko so napovedi neznanih objektov (testnih) slabe.

Na spodnji sliki je prikazan primer, kako bi zgledal model, ki preveč dosledno izračunava vrednosti nenatančnih meritev. Namesto dveh parametrov v primeru linearnega modela (rdeč graf) potrebujemo za model pete stopnje šest parametrov (moder graf).



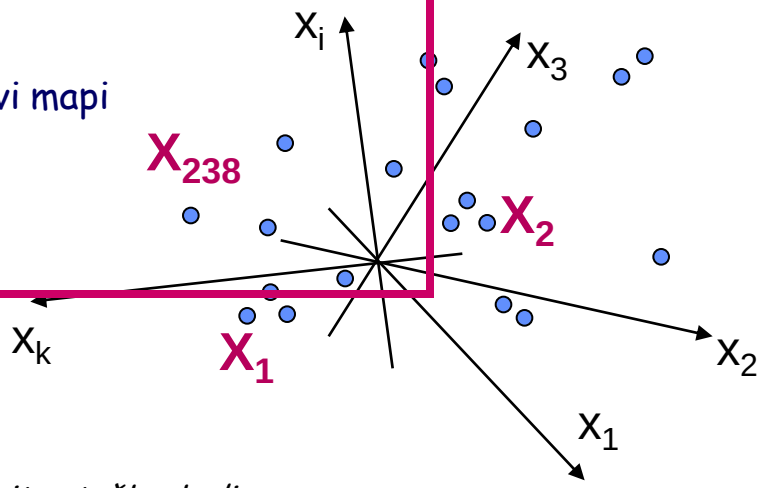
X	Y
10	120
15	148
20	209
25	266
30	283
35	345

Za validacijo napovedne zmožnosti modelov, ki jo v določeni meri izvajamo že v samem procesu izgradnje modelov, imamo na voljo več metod, ki večinoma temeljijo na izločanju dela objektov (testnega niza) iz učnega niza:

- Delitev niza podatkov na učni/testni/validacijski niz
- Navzkrižni test z izpuščanjem po en objekt (leave-one-out, boot-strap)
- Navzkrižni test z izpuščanjem po več objektov hkrati (leave-many-out)
- Naključnostni test (premešanje lastnosti (Y) v parih $\{X_s, Y_s\}$

Delitev podatkovnega niza na 2 (3) nize izvedemo lahko na različne načine:

- Razdelitev po metodi Kennard-Stone
- Razdelitev glede na porazdelitev objektov v Kohonenovi mapi
- Razdelitev po metodi Sphere-excluder
- Naključna razdelitev



Predstavitev točk v k-dimen-zionalnem (merskem) *prostoru*

Kennard-Stonova delitev podatkov na dve skupini

Kennard-Stonova metoda razdeli poljubno skupino N objektov na dva dela. Objekti v prvi skupini so pomebni za predstavitev merskega prostora, objekti druge skupne pa so vedno v bližini objektov prve skupine. Čim je v merskem prostoru nek objekt osamljen, ga Kennard-Stonov postopek zelo hitro uvrsti med "izbrance" prve skupine, ki predstavljajo merski prostor. Z drugimi besedami povedano, to pomeni, da so objekti v drugi skupni nekakšno dopolnilo prve skupine. Zaradi tega so zelo dobri za različna testiranja modelov, ki so bili narejeni s pomočjo objektov prve skupine.

Opisani test namreč izpolnjujejo dva zelo pomembna pogoja:

a) dobro pokrivajo merski prostor (ker ležijo v bližini točk prve skupine) in
b) niso identični objektom iz prve skupine.

To bi bilo povsem res, če bi imel vsak objekt iz prve skupine v bližini najmanj enega, če že ne več objektov iz druge skupine. Ta ugotovitev nakazuje, da je zelo primerna izbira, če je število objektov v prvi skupini manjše od števila objektov v drugi skupini, oziroma da je manjše od polovice vseh objektov ($p < N/2$).

Rezultat Kennard-Stonove razvrstitve objektov je **lista** z zaporednimi številkami objektov ki si sledijo od najbolj reprezentativnega (edinstvenega v danem merskem prostoru) do najmanj reprezentativnega - najbolj podobnega kateremu koli od ostalih objektov.

Podobnost merimo z **razdaljo** med objekti (evklidska razdalja).

Delitveno mejo potegnemo glede na zahteve, ki jih pri obdelavi imamo. Ena od zahtev je lahko velikost testne množice. Navadno želimo imeti vsaj toliko testnih objektov, kot imamo učnih. Lahko se seveda zgodi, posebej pri sorazmeroma majhnem številu objektov, da bomo morali izbrati za test samo tretjino, četrtino, ali še manj objektov. Druga možna zahteva je vnaprej predpisano števila objektov *p*, ki morajo merski prostor predstavljati. V tem primeru preprosto vzamemo prvih *p* objektov v ureditvi, ki jo določi Kennard-Stonov postopek. Najpogostejši kriterij pri delitvi je minimalna razdalja med dvema objektom, za katero menimo (na podlagi predpisov, eksperimentov ali izkušnje), da se dva objekta med seboj še signifikantno razlikujeta.

Pri Kennard-Stonovi delitvi je izhodiščni podatek **matrika razdalj** med vsemi objekti. Postopek pričnemo s tem, da v celotni množici objektov poščemo tista dva, ki sta med seboj najbolj oddaljena. To sta objekta, ki sta prva in najpomembnejša objekta skupine. Ker je metoda odvisna od razdalj, ki so komutativne *(d(Xi, Xj) = d(Xj, Xi))*, obeh najbolj oddaljenih objektov ne moremo razvrstiti po pomembnosti. Se pravi, da ju ne rangiramo na prvega in drugega, ampak sta obema pripišemo enako največjo pomembnost v skupini. Vse naslednje objekte izbiramo posamezno in jim glede na to, v katerem koraku so bili izbrani, določimo pomembnost.

Celotni postopek izbiranja poteka ponavljajoče se, v naslednjih štirih korakih:

•korak 1. Določimo oba najbolj oddaljena objekta, ju imenujemo "izbranca", vse ostale pa "kandidati". Po prvem koraku imamo v skupini N objektov dva izbranca $p = 2$ in $N - p$ kandidatov, se pravi, vse ostale.

•korak 2. Pri vsakem od N-p kandidatov pregledamo razdalje med njim in vsemi p izbranci in določimo najmanjšo, t.j., razdaljo od kandidata do najbližjega izbranca. Na koncu drugega koraka imamo toliko najmanjših razdalj, kolikor je kandidatov, to je N-p.

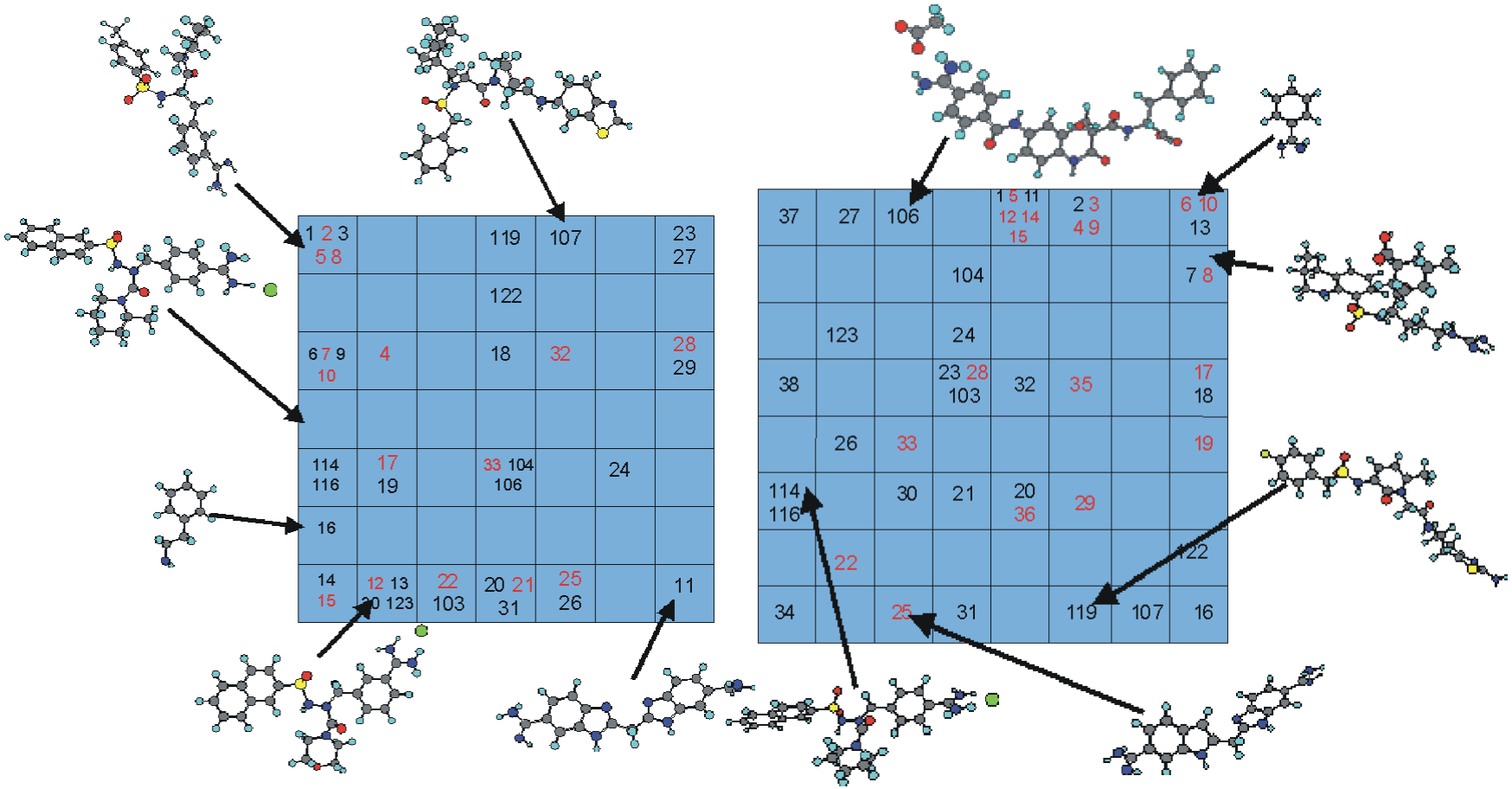
•korak 3. Med vsemi N-p najmanjšimi razdaljami, od katerih je vsaka pripisana enemu kandidatu, poiščemo tisto, ki je med vsemi največja. Kandidat, ki mu pripada največja od vseh najmanjših razdalj, je novi izbranec.

•korak 4. Število izbrancev se za enega poveča, $p \rightarrow p + 1$, število kandidatov pa se za enega zmanjša. Postopek ponovno pričnemo pri koraku 2.

Normalzirane koordinate		ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0.50	0.10	0.38	1	0.00	0.68	0.45	0.35	0.71	0.61	0.72	1.12	0.61	0.31	0.42	0.50	0.46	0.88	0.71	0.85	0.90
0.13	0.00	0.94	2	0.68	0.00	0.25	0.41	0.79	0.69	0.77	**1.01**	0.87	0.74	0.84	**1.013**	0.89	**1.35**	1.17	1.18	1.23
0.25	0.10	0.75	3	0.45	0.25	0.00	0.17	0.64	0.52	0.61	0.93	0.69	0.53	0.65	0.81	0.68	1.12	0.94	0.98	1.03
0.38	0.20	0.69	4	0.35	0.41	0.17	0.00	0.60	0.42	0.50	0.87	0.54	0.41	0.53	0.68	0.53	0.99	0.77	0.82	0.87
0.00	0.60	0.44	5	0.71	0.79	0.64	0.60	0.00	0.31	0.50	0.73	0.95	0.91	1.02	1.12	0.93	0.73	0.88	0.81	0.93
0.25	0.60	0.63	6	0.61	0.69	0.52	0.42	0.31	0.00	0.20	0.55	0.67	0.71	0.80	0.92	0.70	0.74	0.68	0.60	0.70
0.38	0.70	0.75	7	0.72	0.77	0.61	0.50	0.50	0.20	0.00	0.41	0.58	0.73	0.80	0.92	0.67	0.79	0.63	0.49	0.57
0.25	1.00	1.00	8	1.12	1.01	0.93	0.87	0.73	0.55	0.41	0.00	0.90	1.12	1.18	1.30	1.04	1.04	0.93	0.71	0.77
0.88	0.40	0.75	9	0.61	0.87	0.69	0.54	0.95	0.67	0.58	0.90	0.00	0.38	0.35	0.45	0.21	0.98	0.48	0.57	0.53
0.75	0.10	0.56	10	0.31	0.74	0.53	0.41	0.91	0.71	0.73	1.12	0.38	0.00	0.13	0.28	0.24	1.01	0.64	0.80	0.81
0.88	0.10	0.56	11	0.42	0.84	0.65	0.53	1.02	0.80	0.80	1.18	0.35	0.13	0.00	0.18	0.20	1.05	0.63	0.81	0.80
1.00	0.10	0.44	12	0.50	1.01	0.81	0.68	1.12	0.92	0.92	1.30	0.45	0.28	0.18	0.00	0.27	1.04	0.62	0.84	0.82
0.88	0.30	0.56	13	0.46	0.89	0.68	0.53	0.93	0.70	0.67	1.04	0.21	0.24	0.20	0.27	0.00	0.90	0.44	0.62	0.60
0.50	0.90	0.00	14	0.88	1.35	1.12	0.99	0.73	0.74	0.79	1.04	0.98	1.01	1.05	1.04	0.90	0.00	0.57	0.56	0.68
0.88	0.70	0.38	15	0.71	1.17	0.94	0.77	0.88	0.68	0.63	0.93	0.48	0.64	0.63	0.62	0.44	0.57	0.00	0.27	0.27
0.75	0.90	0.50	16	0.85	1.18	0.98	0.82	0.81	0.60	0.49	0.71	0.57	0.80	0.81	0.84	0.62	0.56	0.27	0.00	0.14
0.88	0.90	0.56	17	0.90	1.23	1.03	0.87	0.93	0.70	0.57	0.77	0.53	0.81	0.80	0.82	0.60	0.68	0.27	0.14	0.00

Zaporedje izbrancev	Št. izbranca	najbližji izbranec	razdalja do najbližjega izbranca
1-2	2	14	1.353
3	12	2	1.013
4	8	2	1.010
5	5	14	0.729
6	17	14	0.676
7	1	12	0.504
8	9	12	0.451
9	7	8	0.410
10	4	1	0.351
11	10	12	0.280
12	15	17	0.274
13	13	9	0.213
14	6	7	0.203
15	3	4	0.172
16	16	17	0.140
17	11	10	0.125

Delitev podatkov na dve skupini s pomočjo Kohonenove nevronske mreže



Delitev podatkov na dve skupini s pomočjo Kohonenove nevronske mreže

6	6	6	6	6	5	5	5	5	5	5		8	8	8	8	8	8	8	
6	6	6	6			5	5		5	5	5		7		8	8	8		8
6	6	6	6		5	5	5	5	5	5			7		8	8	8	8	8
6	6	6	6	5		5	5	5	5	5		7	7	7	7	8	8	8	8
			6	6	5	5	5	5		7		7	7	7	7	7		8	8
3	3				5	5	5		7	7	7		9	9	9	9		8	
3					5	5	5		7	7	7		9	9	9	9			
3	3					5	5		7	7	7	7	9	9	9	9	9		
3	3				3			7	7	7	7		9	9	9				1
3	3	3	3	3	3	3			7	7	7	9					1	4	1
3	3	3	3	3	3	3	2	2					4		2	1	1	1	1
3	3	3	3		3	3	3	3		2	2	2	2	2	2	1		1	1
3	3	3	3	3	3	3	3	3					2	2			1	1	1
3		3			3	3	3			3	2	2	2		2		2	4	4
3	3	3	3	3	3	3	3	3	3	3		2		2		2	2		
	3		3	3	3	3	3	3	3	3	3	2		2	2	4		4	4
3	3	3		3	3	3	3	3	3	3	3	3	4	2	2	4	4	4	4
3	3	3	3		3		3	3	3	3		4		2			2	2	2
3	3		3		3	3	3	3		3	4	4	3	4	4	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3		3	4	4	4	2	2	2	4

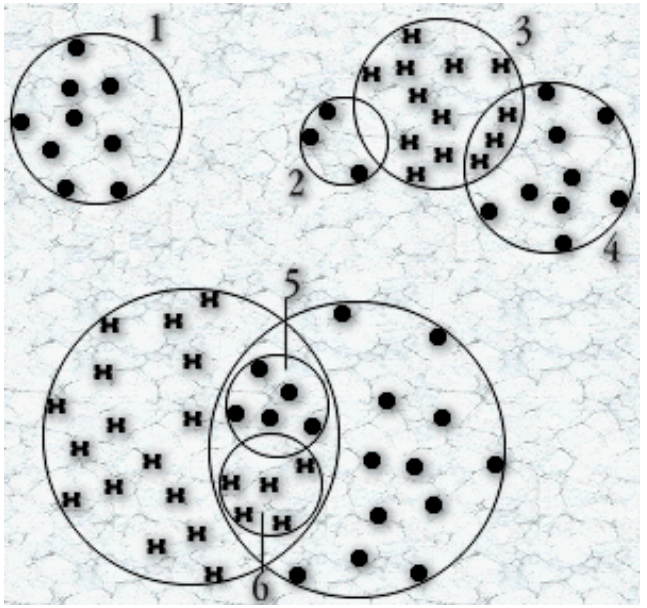
Razred vzorcev	Pokrajina	Število
1	North Apulia	25
2	Calabria	56
3	South Apulia	206
4	Sicily	36
5	Inner Sardinia	65
6	Coastal Sardinia	33
7	East Liguria	50
8	West Liguria	50
9	Umbria	51
	Σ	572

Sphere-exclude delitev podatkov na dve skupini

Algoritem:

1. Izbere spojino z največjo vrednostjo Y in jo da v učni niz "TSET".
2. Naredi krožnico z radijem R okoli te točke (objekte)
3. Vsi objekti znotraj krožnice gredo v testni niz.
4. Izloči objekte iz koraka 3 iz podatkovnega niza.
5. Če ni več spojin, konča.
6. Izračuna razdalje med preostalimi spojinami in centrom.
7. Izbere objekt z največjo razdaljo in gre na korak 2.

C: izbrana konstanta; V/N: volumen; K: dimenzija objektov



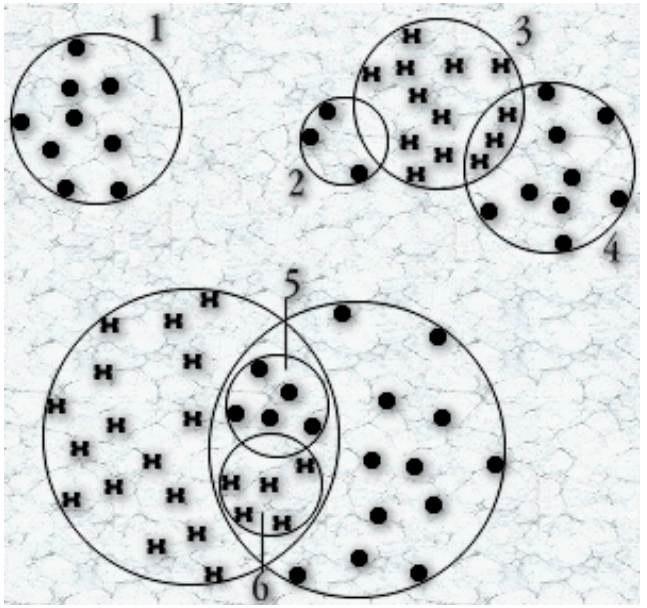
$$R = c\left(\frac{V}{N}\right)^{1/K}$$

Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Golbraikh A, Tropsha A. Journal of Computer-Aided Molecular Design 2002;16(5-6):357-369.

Golbraikh, A., J. Chem. Inf. Comput. Sci. 40 (2000) 414–425.

Sphere-exclude delitev podatkov na dve skupini

Algoritem:



J Chem Inf Comput Sci. 2003 Jan-Feb;43(1):317-23.
DISE: directed sphere exclusion.
Gobbi A, Lee ML.

$$R = c\left(\frac{V}{N}\right)^{1/K}$$

Mol Divers. 2002;5(4):231-43.
Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection.
Golbraikh A, Tropsha A.

Navzkrižni test z izpuščanjem po en objekt (leave-one-out, boot-strap)

N=8

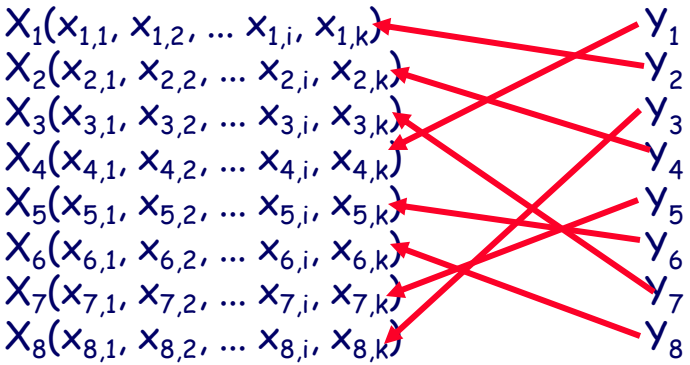
$X_1(x_{1,1}, x_{1,2}, \ldots x_{1,i}, x_{1,k})$	y_1
$X_2(x_{2,1}, x_{2,2}, \ldots x_{2,i}, x_{2,k})$	y_2
$X_3(x_{3,1}, x_{3,2}, \ldots x_{3,i}, x_{3,k})$	y_3
$X_4(x_{4,1}, x_{4,2}, \ldots x_{4,i}, x_{4,k})$	y_4
$X_5(x_{5,1}, x_{5,2}, \ldots x_{5,i}, x_{5,k})$	y_5
$X_6(x_{6,1}, x_{6,2}, \ldots x_{6,i}, x_{6,k})$	y_6
$X_7(x_{7,1}, x_{7,2}, \ldots x_{7,i}, x_{7,k})$	y_7
$X_8(x_{8,1}, x_{8,2}, \ldots x_{8,i}, x_{8,k})$	y_8

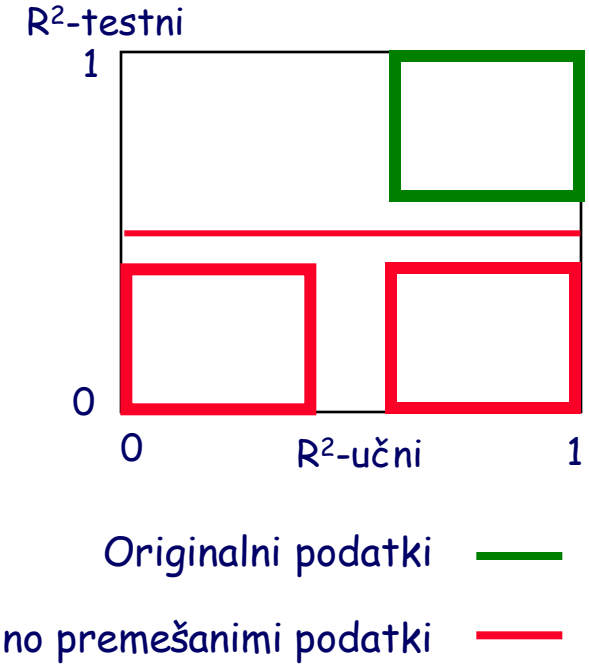
N-1 objektov uporabimo za izgradnjo modela (določitev parametrov modela)
1 objekt ostane za testiranje modela

Za končno ovrednotenje naredimo N modelov z N-1 objekti in vsakič testiramo drug objekt, to je izpuščeni *i*-ti objekt.

Naključnostni test (randomization test)

N=8





N-1 objektov uporabimo za izgradnjo modela (določitev parametrov modela)
1 objekt ostane za testiranje modela

Ovrednotenje modela z naključno premešanimi podatki mora jasno pokazati veliko napako napovedanih vrednosti.