

- a. Enunțați [succint] regula de decizie a algoritmului k-NN pentru o instanță de test x_0 .

Care este bias-ul inductiv al algoritmului k-NN?

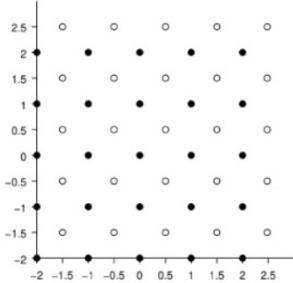
În continuare veți lucra pe datele din figura de mai jos. Veți aplica algoritmul k-NN, folosind distanța euclidiană.

k-NN funcționează bine atunci când instanțele dintr-o aceeași clasă sunt plate într-una sau mai multe zone din spațiu relativ bine delimitate, fără întrepătrunderi puternice.

Obiectivul nostru acum este să analizăm ce se întâmplă atunci când datele sunt puternic mixate. Rezultatele pe care le veți obține la calculul erorilor vor fi exprimate sub formă de numere [fracționare] din intervalul $[0, 1]$.

Observație importantă:

În cazul în care există două sau mai multe instanțe situate exact pe „marginea” [adică, pe conturul circular al] k-NN-vecinătății asociate instanței de clasificat, se va considera că toate aceste instanțe aparțin respectivei vecinătăți, iar fiecare dintre ele dispune de un vot întreg.



- b. Pentru $k = 1$, calculați eroarea la antrenare și eroarea la cross-validation cu metoda “leave-one-out” (CVLOO).

Ce puteți spune comparând cele două rezultate? (Care este legătura între bias-ul inductiv al lui k-NN și puterea de generalizare a lui 1-NN pe astfel de date?)

- c. Pentru $k = 2$, calculați eroarea la cross-validation cu metoda “leave-one-out” (CVLOO).

d. Considerăm $k = 50$. (Remarcăți faptul că în total în setul nostru de date sunt 50 de instanțe.) De această dată, vom impune ca algoritmul k-NN să ia decizia în mod probabilist. Aceasta înseamnă că dacă în vecinătatea k-NN a unei instanțe de test există n vecini pozitivi și m vecini negativi, atunci algoritmul k-NN va returna (pentru instanța respectivă) decizia + cu

- a) Bias-ul inductiv al k-NN este „Cine se ascundă în se adună” sau mai nou: "Spune-mi cu cine vădăzi ca nă-ți spun cine ești".
 b) pentru $k=1$ se obț. din tabel că eroarea la CVLOO nu poate fi de 100%, datorită proprietății simetriei

Eroarea la antrenare pt. 1-NN este 0, întrucât datele de antrenament nu conțin inconistențe.

- c) pentru $k=2$ obținem pt. punctele:

$$\text{Instanță } \begin{bmatrix} -2 \\ -2 \end{bmatrix} = \left\{ \begin{bmatrix} -1,5 \\ -1,5 \end{bmatrix}, \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \end{bmatrix} \right\} = 1 \quad \checkmark$$

Vecinătate

$$\text{Instanță } \begin{bmatrix} 2,5 \\ 2,5 \end{bmatrix} = \left\{ \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2,5 \\ 1,5 \end{bmatrix}, \begin{bmatrix} 1,5 \\ 2,5 \end{bmatrix} \right\} = 0 \quad \checkmark$$

\Rightarrow folosim convenția \Rightarrow
 \Rightarrow lucram cu clasele cu ieșiri intregi.

$$(\text{CVLOO} =) \frac{48}{50}$$

probabilitatea $n/(n+m)$ și decizia – cu probabilitatea $m/(n+m)$. În consecință, pentru întreg setul de date vom putea calcula o eroare medie.

Calculați eroarea medie la antrenare pentru algoritmul 50-NN pe datele de mai sus. Cunoașteți o metodă de clasificare foarte simplă care obține pe aceste date rezultate la fel de bune / proaste precum 50-NN?

d)

Puncte	etichete	50-NN	$\frac{n}{n+m}$	$\frac{m}{n+m}$	Clasificare
$\begin{bmatrix} -2 \\ -2 \end{bmatrix}$	1	{0, 1}	$\frac{25}{50}$	$\frac{25}{50}$	0
$\begin{bmatrix} -1.5 \\ -1.5 \end{bmatrix}$	0	{..}	$\frac{25}{50}$	$\frac{25}{50}$	1
:					

Calc: eroare medie :

$$\frac{\frac{1}{50} + \frac{1}{50} + \dots + \frac{1}{50}}{50} = \frac{1}{2}$$

sus. Cunoașteți o metodă de clasificare foarte simplă care obține pe aceste date rezultate la fel de bune / proaste precum 50-NN?

Dăm cu banul , eroare 50% , nu zic nimic.

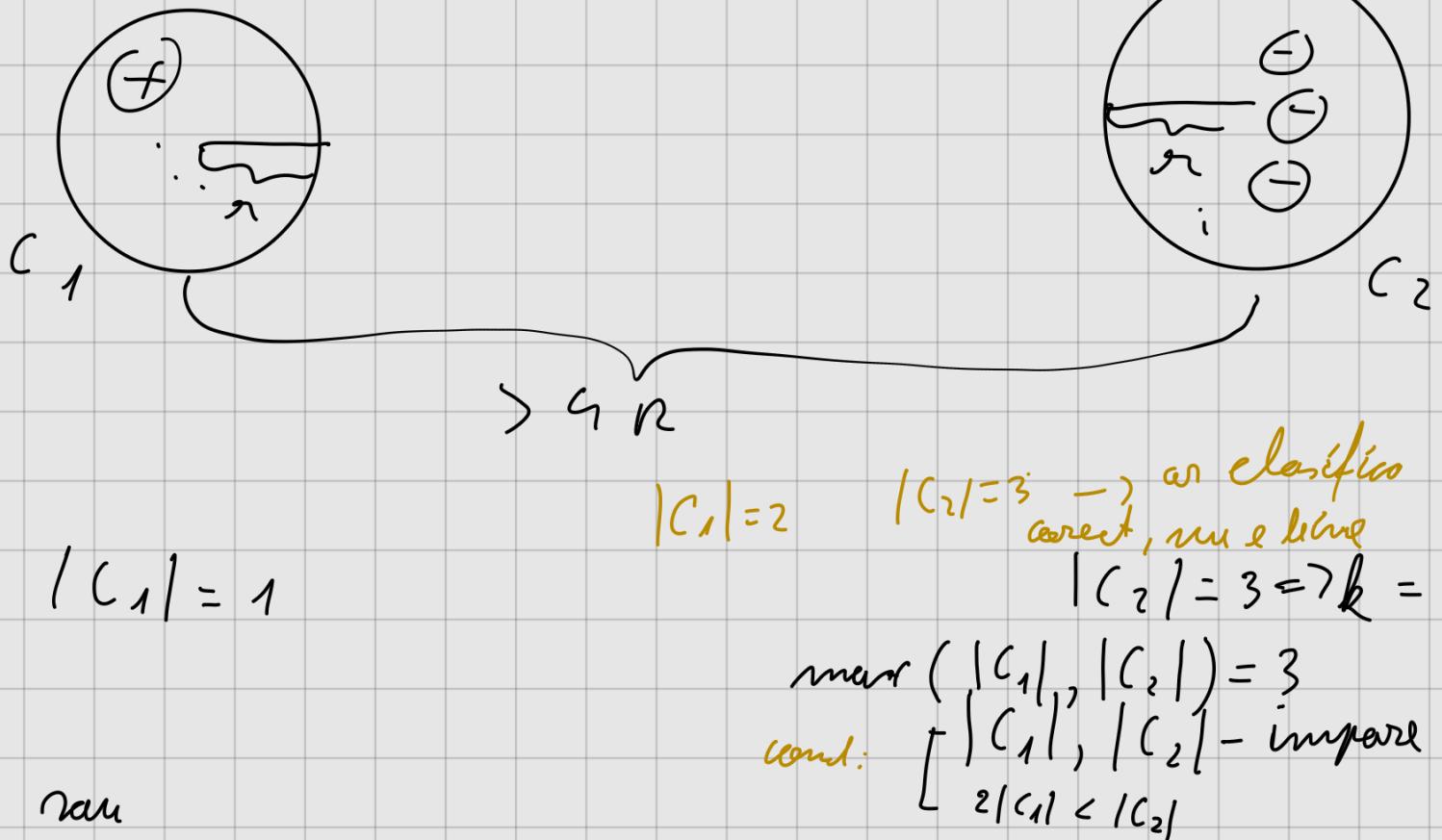
17.

(Algoritmul k -NN: acuratețe; comparații pentru diferite valori ale lui k)

• CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 5.a

Considerăm două clase, notează cu C_1 și C_2 , în spațiul euclidian bidimensional. Datele din clasa C_1 sunt distribuite în mod uniform într-un cerc de rază r . Datele din clasa C_2 sunt distribuite în mod uniform într-un alt cerc de rază r . (Observație: Numărul de date din cele două clase nu este neapărat același.) Centrele celor două cercuri sunt situate la o distanță strict mai mare decât $4r$.

Arătați că este posibil ca [la antrenare] acuratețea algoritmului 1-NN aplicat pe aceste date să fie strict mai mare decât acuratețea algoritmului k -NN, pentru un anumit număr întreg $k \geq 3$, impar, ales în mod convenabil.



$$|C_1| = 10 \quad |C_2| = 50 \quad K = 2 \cdot \min(|C_1|, |C_2|)$$

$K = 21$

trb. co corac. pentru marata
d. n̄ no 3 de err.

18. (Algoritmul 1-NN: calculul erorii la CVLOO)
 * CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm exam, pr. 1.7

Care este eroarea clasificatorului 1-NN la cross-validation de tip "Leave-One-Out" pe setul de date alăturat?



Puncte	Vicini	Efectuata	Clasificare CVLOO	Eroare
1	{23}	-	-	nu
2	{1}	-	-	nu
3	{43}	-	+	da
4	{3}	+	-	da
5	{4}	+	+	nu

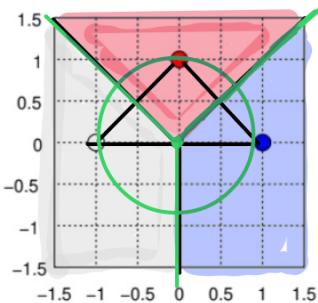
19.

(Algoritmul 1-NN: granițe / suprafețe de decizie)

• CMU, 2013 fall, W. Cohen, E. Xing, final exam, pr. 3.6

D.P. dist. euclidiană

Trasați granițele de decizie (engl., *decision boundaries*) produse de clasificatorul 1-NN la aplicarea pe datele din figura alăturată. Diversele culori ale punctelor reprezintă clase diferite.



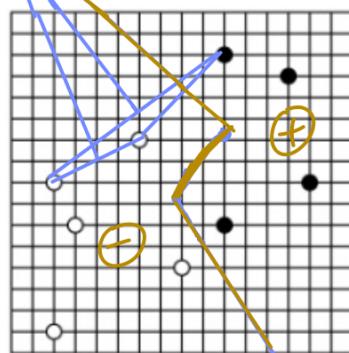
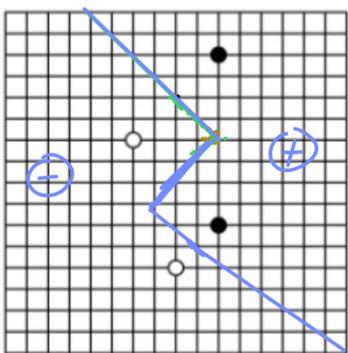
- Se trasează mediul între cele mai apropiate perechi de puncte etichetate diferit.
- Se trasează un triunghiul între 2 puncte etichetate diferit și un alt 3-lea punct a. În interiorul triunghiului nu sunt alte puncte.
- ✓ Ian cercul circumscris însă nu conține alte puncte în interior.

20.

(Algoritmul 1-NN: granițe / suprafețe de decizie)

* CMU, 2008 fall, Eric Xing, HW1, pr. 3

În fiecare din figurile următoare se dau câteva puncte / instanțe în spațiul bidimensional, care sunt etichetate cu • (instanțe pozitive) sau ○ (instanțe negative). Indicați în fiecare caz granițele / suprafețele de decizie pentru algoritmul 1-NN presupunând că se folosește distanța euclidiană.

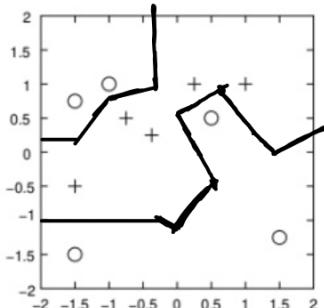
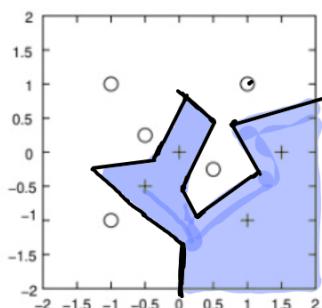
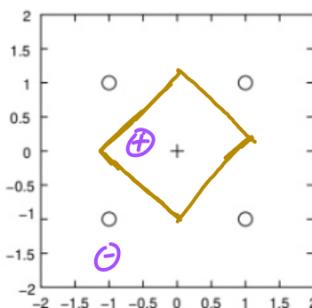
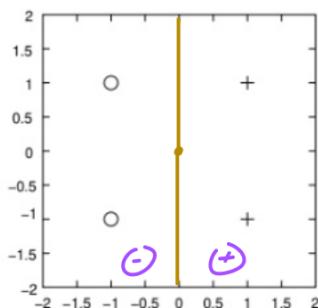


21.

(Algoritmul 1-NN: diagrame Voronoi)

• * CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 3.1

Desenați suprafețele de decizie pentru clasificatorul 1-NN pentru fiecare dintre seturile de date din figurile de mai jos. Folosiți distanță euclidiană. Hașurați fin zonele corespunzătoare clasei +.



464

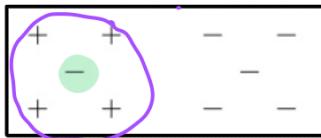
24.

(Algoritmul k -NN:CVLOO: comparație pentru diferite valori ale lui k ; eroarea la antrenare: comparație cu alți clasificatori)

• * CMU, 2010 fall, Aarti Singh, midterm exam, pr. 2

- a. Care dintre clasificatorii de mai jos realizează o eroare de tip CVLOO (Leave-One-Out Cross-validation) mai mare pe setul de date alăturat?

1-NN $\frac{1}{2} > \square$ 3-NN $\frac{1}{10}$



- b. Considerăm setul de date din figura alăturată. Care dintre clasificatorii de mai jos obține / obțin eroare nulă la antrenare pe acest set de date?

eroare nula

- arborii de decizie ID3 de adâncime 2 regresia logistică
 clasificatorul 3-NN SVM (cu nucleu pătratic)

 $\hookrightarrow \text{eroare} = 1$ a) 1-NN *Vom avea 5 eroare.*

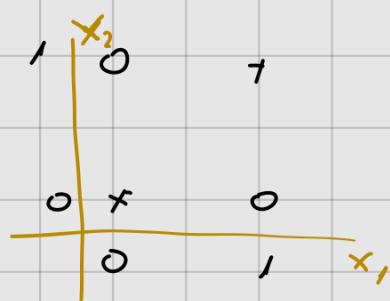
$$\text{CVLOO}, \text{eroare} = \frac{5}{10} = \frac{1}{2}$$

3-NN

$$\text{CVLOO}, \text{eroare} = \frac{1}{10}$$

b) ID3 \rightarrow decarece obține datele ca un XORObservația: setul de date \Rightarrow consistent \Rightarrow \Rightarrow eroare la antrenare este 0.

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	1

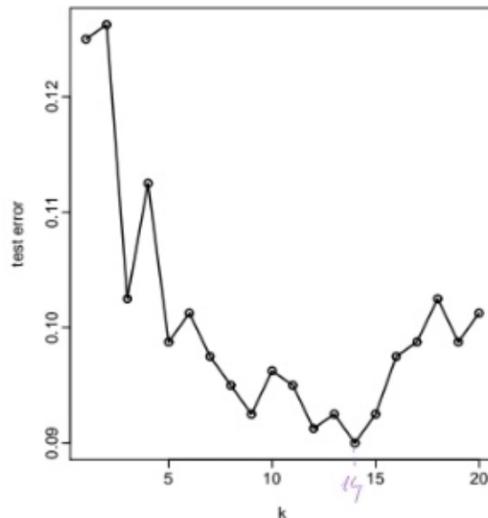


25.

(Algoritmul k -NN: alegerea valorii convenabile pentru k)• prelucrare de Liviu Ciortuz, după
CMU, (?) spring, ML course 10-701, HW1, pr. 5

Pe un anumit set de date format din date de antrenament, date de validare și date de test, după ce a fost antrenat algoritmul k -NN pentru diferite valori ale lui k , rezultatele obținute la validare au fost reprezentate în graficul care urmează.

Care este — în conformitate cu aceste rezultate — valoarea optimă care trebuie aleasă pentru k , în vederea folosirii ulterioare pe datele de test? Justificați alegerea făcută.



Obțin că în $k=14$ găsim rezolvarea minimă pentru care alg K-NN a înregistrat cele mai puține erori de test.

Prădător, vom alege $k=14$, deoarece ne descurcă cel mai bine pe setul de teste pentru datele noi.

22.

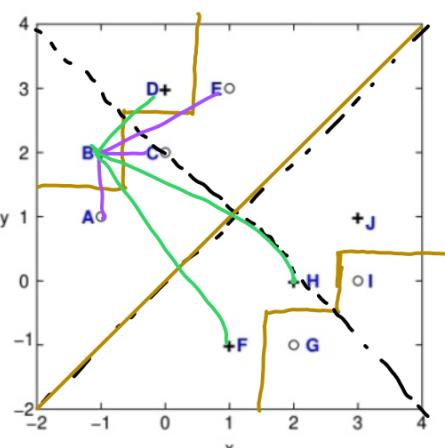
(Algoritmul k -NN: diagrama Voronoi, eroarea la CVLOO; comparație pentru diferite valori ale lui k)• prelucrare de L. Ciortuz, după
CMU, 2012 spring, Ziv Bar-Joseph, midterm exam, pr. 2

La acest exercițiu veți aplica algoritmul k -NN folosind distanța euclidiană pe setul de date din figura de mai jos. Fiecare punct aparține la una din două clase, desemnate cu + și respectiv o.

a. Pentru $k = 1$, trasați diagrama Voronoi și hașurați zona / zonele de decizie corespunzătoare etichetei +.

b. Care este eroarea la cross-validation cu metoda "Leave-One-Out" (CVLOO) dacă se folosește algoritmul 1-NN?

c. Care dintre următoarele valori ale lui k va conduce la o valoare minimă a erorii de tip CVLOO: 3, 5, 7 sau 9? Comentați succint rezultatul.



Care mi se pare multă eroare?

dacă: 0

L,



rezultatul

b)

$$\text{+) } \frac{1}{d^2(B,A)} + \frac{1}{d^2(B,C)} + \frac{1}{d^2(B,E)}$$

$$\text{o: } \frac{1}{d^2(B,D)} + \frac{1}{d^2(B,F)} + \frac{1}{d^2(B,H)}$$

Data	Et.	Vecinătate					Clasif. la CVLOO					Eroare? (da/nu)				
		1-	3-	5-	7-	9-NN	1-	3-	5-	7-	9-NN	1-	3-	5-	7-	9-NN
A	o	B	C D	E F	G H	J I	-	+	-	-	-	1	1	1	1	1
B	+	A C	D	D, F, H	J	G I	0	0	0	0	O	1	1	1	0	1
C	o	B D	A E	H	F J	G I	+	+	+	+	+	1	1	1	1	1
...																

$$d(B, G) = \left\| \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\| = \left\| \begin{bmatrix} -3 \\ 3 \end{bmatrix} \right\| = \sqrt{9+9} = \sqrt{18}$$

$$d(B, J) = \left\| \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix} \right\| = \left\| \begin{bmatrix} -4 \\ 1 \end{bmatrix} \right\| = \sqrt{16+1} = \sqrt{17}$$

1 1 1 $\frac{6}{10}$ 1 (deșteapta
 par + at tabelul,
 mă folosesc de
 simetria)

dacă date multe, cel mai apropiat vecin ar fi eticheta cunoscută, de astăzi avem eroare maximă.

De ce $k=2$ dă mai bine? (nu stiu, nu important)

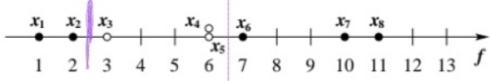
23.

(Diagramme Voronoi pentru dataset-uri din \mathbb{R} , folosind 1-NN și 5-NN;
 calcularea erorii la CVLOO pentru algoritmul 1-NN;
 comparație cu algoritmul ID3 cu atribute numerice continue)

• * MIT, ML 6036 course, Review Material Problems, ex. 37

Comentariu: În acest exercițiu, spre deosebire de toate celelalte cazuri când am construit diagrame Voronoi (întotdeauna pentru algoritmul 1-NN și seturi de date din \mathbb{R}^2), vom construi astfel de diagrame pentru seturi de date din \mathbb{R} , mai întâi pentru algoritmul 1-NN și apoi pentru algoritmul 5-NN. (Absolut similar se poate proceda și pentru alte valori ale lui $k \neq 1$, atunci când datele de antrenament sunt din \mathbb{R} .)

În desenul de mai jos este reprezentat un set de antrenament care conține 8 instanțe, fiecare dintre ele având doar o trăsătură (engl., feature), notată cu f .



Remarcăți faptul că sunt două instanțe pentru care valoarea trăsăturii f este aceeași, și anume 6. Aceste două instanțe sunt reprezentate prin două simboluri \circ , situate unul deasupra celuilalt, dar de fapt ele ar fi trebuit să fie reprezentate ca două simboluri \circ suprapuse (unul peste celălalt), însăciute instanțe au exact aceeași valoare pentru trăsătură f .

Vă readucem aminte *convenția* noastră de notare: simbolul \bullet desemnează instanțe pozitive (+), iar simbolul \circ instanțe negative (-).

a. La acest punct al problemei veți folosi algoritmul 1-NN.

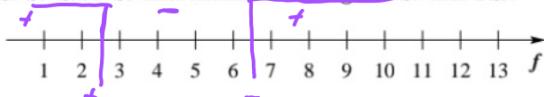
Convenții / reguli:

- În cazul în care există două sau mai multe instanțe situate exact pe „marginile” [adică, pe conturul circular al] k -NN-vecinătății asociate instanței de clasificat, se va considera că toate aceste instanțe aparțin respectivei vecinătăți, iar fiecare dintre ele dispune de un vot întreg.
- În caz de *paritate de voturi*, veți alege eticheta instanței de antrenament care este situată la *distanță minimă* către stânga față de instanța de test respectivă. (Distanța minimă este 0 atunci când instanța de test coincide cu o instanță de antrenament!)
- Acste reguli vor fi aplicate și la punctele următoare.

—> Convenție. alegem cel din stânga

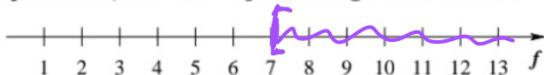
a)

Vă cerem ca pe linia de mai jos să hașurați — și, bineînțeles, să delimitați prin separatori decizionali — zonele în care algoritmul 1-NN va prezice semnul +, dat fiind setul de date de antrenament din figura de mai sus.



În prealabil, vă cerem să stabiliți cu ce etichetă vor fi clasificate instanțele de test $f = 2.5$ și $f = 6.5$.⁴⁷² Justificați.

- b. Dacă faceți cross-validation folosind metoda "leave-one-out" pe acest set de date, în conjuncție cu algoritmul 1-NN, cât va fi eroarea produsă? Veți arăta în mod clar — alcătuind un tabel care să conțină 1-NN-vecinătățile respective — cum anume ați ajuns la rezultatul pe care l-ați indicat.
- c. Similar cu punctul a, însă aici veți folosi algoritmul 5-NN.⁴⁷³



Indicație: Justificați în mod riguros rezultatul, referindu-vă la diverse intervale de valori (sau valori particulare) ale lui f :

Cazul 1: $f \in (-\infty, 4)$: ...

Cazul 2: ...

...

În prealabil, vă cerem să stabiliți cu ce etichetă vor fi clasificate instanțele de test $f = 4$, $f = 6$ și $f = 7$.⁴⁷⁴ Justificați.

Observație: Se poate demonstra faptul că în cazul seturilor de date de antrenament $x_1, x_2, \dots, x_n \subset \mathbb{R}$ (adică, atunci când se lucrează pe axa reală) k -NN-vecinătățile nu sunt discontinue, adică: dacă fiind un punct de test oarecare $x_q \in \mathbb{R}$, dacă instanțele de antrenament x_i și x_j , cu $x_i \leq x_j$, se află în k -NN-vecinătatea lui x_q , atunci orice instanță de antrenament x_l care satisface proprietatea $x_i \leq x_l \leq x_j$ aparține și ea respectivăi k -NN-vecinătăți a lui x_q .⁴⁷⁵

- d. Construiți — în mod riguros — arborele de decizie corespunzător setului de date din enunț, considerând f ca fiind atribut numeric continuu. Comparați rezultatul obținut de data aceasta cu rezultatele care au fost produse de clasificatorii 1-NN și 5-NN.

ea = $\frac{2}{8}$

app · 103.

Stim că la aplicarea algoritmului k -NN, clasificarea unei instanțe date se face pe baza votului majoritar obținut în „vecinătatea“ instanței respective. Preșupunem că se dă două clase de instanțe, fiecare clasă având $n/2$ puncte, întrepătrunse într-o anumită măsură, într-un spațiu bidimensional.

- a. Descrieți ce se întâmplă cu eroarea la antrenare (folosind toate datele disponibile) când numărul k al vecinilor considerați variază de la n la 1.

468

Probleme propuse

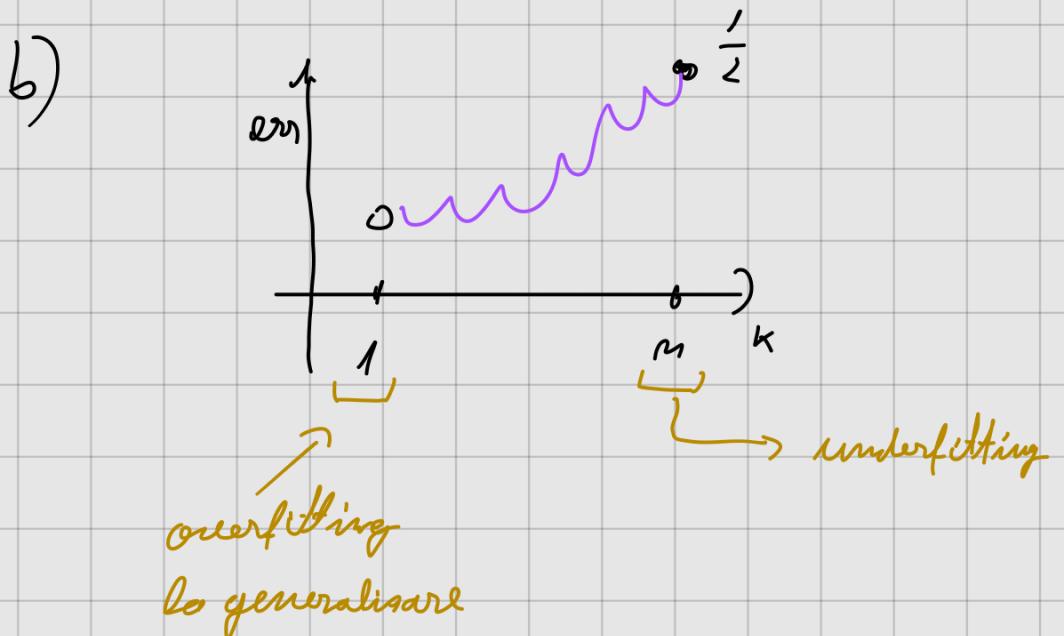
ÎNVĂȚARE bazată pe MEMORARE

- b. Schițați grafic cum anume ar evoluă eroarea la generalizare (de exemplu, reținând o parte din date pentru testare) atunci când k variază. Explicați modul în care ați raționat.
- c. Propuneți o metodă de determinare a unei valori adecvate pentru k .
- d. La folosirea algoritmului k -NN, odată ce s-a stabilit valoarea lui k , toți cei mai apropiati k vecini ai punctului de clasificat au ponderi egale (adică, aceeași importanță) la stabilirea etichetei respectivului punct. Sugerați o modificare a algoritmului k -NN care elimină această limită.
- e. Dați două motive pentru care este de preferat să nu folosim algoritmul k -NN atunci când dimensiunea spațiului datelor de intrare este mare.



1 nu testare :

$$\left\{ \begin{array}{l} + : \frac{m}{2} \\ - : \frac{n}{2} \end{array} \right.$$



c) facem un grafic cu fătă k , și-l lăudăm măcarul
Eș lo ex. anterior

d) \rightarrow ponderi la distanțe.

e)
1)
2) alg.

mai avem 27, 28

21.

(Algoritmul 1-NN: diagrame Voronoi)

• * CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 3.1

Desenați suprafețele de decizie pentru clasificatorul 1-NN pentru fiecare dintre seturile de date din figurile de mai jos. Folosiți distanță euclidiană. Hașurați fin zonele corespunzătoare clasei +.

