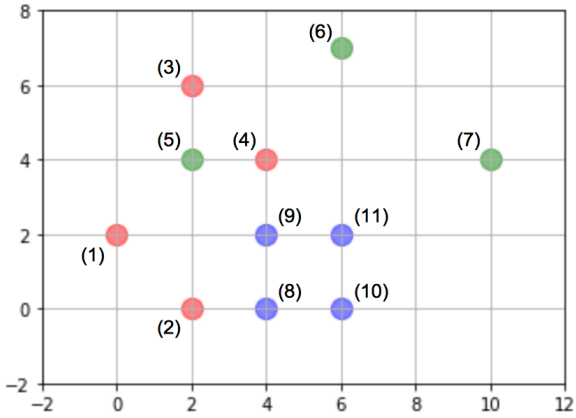
Licence d'Informatique 3
AD Programmeur (C5–160512-INFO)
TD 4 – Classification Supervisée

Carl FRÉLICOT – Dpt Info / Lab MIA

Soient $n = 11$ points en mode supervisé (X, Y), dont le nuage en dimension $p = 2$ est représenté ci-contre :



n^o :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	$\overline{x}$	s^2
${}^tX =$	0	2	2	4	2	6	10	4	4	6	6	4.18	6.88
	2	0	6	4	4	7	4	0	2	0	2	2.82	5.23
$Y =$	1	1	1	1	2	2	2	3	3	3	3	$m = 3$	

1. *Nearest-Prototype Rule*

1-1) On souhaite prédire le groupe auquel associer un/des nouveau/x point/s à l'aide de la règle du *Plus Proche Prototype* au sens de la distance euclidienne. De quoi a-t-on besoin ?

1-2) Faites les calculs nécessaires afin de prédire le groupe pour ${}^tx = (3, 2)$.

1-3) Visuellement, pensez-vous qu'on obtiendrait les mêmes résultats avec la distance de Manahattan ?

(HW) Refaites les calculs nécessaires au classement des points ${}^ty = (10, 2)$ et ${}^tz = (5, 3)$ pour ces deux distances.

(HW) De même avec la distance de Chebychev (à la main et/ou avec vos fonctions python).

2. *Nearest-Neighbors Rule*

Ci-contre, les distances des points x,y et z aux données de X.

2-1) Donnez les prédictions par la règle du *Plus Proche Voisin* pour au moins une distance, vous le ferez chez vous pour celle/s qui reste/nt.

2-2) Recommencez avec la règle des 5-*PPV*. Vous donnerez les vecteurs d'étiquettes (ce qu'on appelle abusivement des "probabilités", ex. Orange).

2-3) On lit dans certains ouvrages spécialisés (ou dans des cours) qu'il vaut mieux prendre K impair. Qu'en pensez-vous ?

d_1 :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
x	3	3	5	3	3	8	9	3	1	5	3
y	10	10	12	8	10	9	2	8	6	6	4
z	6	6	6	2	4	5	6	4	2	4	2

d_2^2 :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
x	9	5	17	5	5	34	53	5	1	13	9
y	100	68	80	40	68	41	4	40	36	20	16
z	26	18	18	2	10	17	26	10	2	10	2

d_∞ :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
x	3	2	4	2	2	5	7	2	1	3	3
y	10	8	8	6	8	5	2	6	6	4	4
z	5	3	3	1	3	4	5	3	1	3	1

3. Évaluation de Règle de Classification

3-1) Composez, selon les règles de l'art, un ensemble de test (X_{test}, Y_{test}) et un ensemble d'apprentissage (X_{learn}, Y_{learn}) en *holdout* à 30% (à peu près).

(HW) Recommencez en *holdout* à 50% (à peu près). Serez-vous en mesure d'avoir le résultat de la 2-validation croisée ?

3-2) Quelle autre stratégie est la plus recommandée pour ces données ?

3-3) Réalisez visuellement les prédictions par la règle du 3-NN avec la distance de Manhattan selon cette stratégie. Vous donnerez la matrice de confusion (table de contingence $C = [c_{ij}]_{i,j=1,m}$ croisant Y_{test} et Y_{pred}), en déduirez la *Classification Accuracy* , puis les scores de *Recall*, *Precision*, et *Classification Accuracy* par classe$^{(a)}$, puis globaux.
$R_j = \frac{TP}{TP+FN}$, $P_j = \frac{TP}{TP+FP}$, $FA_j = \frac{FP}{TP+FP} = 1 - P_j$, $F1_j = 2\,\frac{R_j \times P_j}{R_j + P_j}$, $A_j = \frac{TP+TN}{TP+FN+FP+TN}$

(HW) Calculez les autres scores.

(HW) Pourquoi est-ce plus difficile de réaliser visuellement les prédictions par la règle du NP selon cette stratégie ?

3-4) Évaluez la règle du NP au sens de la distance euclidienne par reclassement de l'ensemble d'apprentissage.

3-5) Pourquoi n'est-ce pas une bonne stratégie d'évaluation ?

(HW) Dans un espace réduit

Soit le tableau de données du ressenti de 10 apprenants sur l'utilisation de logiciels (`Blue` et `Yellow`) de *Machine Learning* et leur association `Green` du TD précédent. On donne ci-contre les vecteurs directeurs des 2 premiers (seuls ?) axes discriminants.

-0.0656	0.5379	0.8404
0.8693	-0.4441	0.2172

${}^tu_1$
${}^tu_2$

3-1) Projetez les 10 apprenants dans l'espace discriminant.

3-2) Vérifiez que les centres des groupes sont bien ceux donnés ci-contre.

- ${}^t\overline{x}_{Blue} = (11.23, -1.58)$
- ${}^t\overline{x}_{Green} = (21.87, 6.83)$
- ${}^t\overline{x}_{Yellow} = (7.73, 11.20)$

3-3) Á l'aide de la règle *NP* ou *1-NN*, prédisez le mode d'utilisation d'un nouvel apprenant qui aurait donné comme évaluation ${}^tx = (10, 10, 10)$.