Licence d'Informatique 2
Analyse de Données Utilisateur (C5–160412)
TD 6 – Évaluation(s)

Carl FRÉLICOT – Dpt Info / Lab MIA

1. Évaluation d'une Règle (Prédiction)

On dispose d'un tableau de données X_L (*learn* or *train data*) et d'une indicatrice de groupe Y_L qui a permis d'apprendre une règle de classement. Pour évaluer son *pouvoir prédictif*, on peut utiliser un autre tableau X_T (*test data*) d'individus dont on connait l'appartenance aux groupes (*actual*) ; évaluer la performance consistera à confronter les prédictions Y_T (*predicted*) à cette vérité-terrain. Si $X_T = X_L$, on parle de *reclassement* (*resubstitution*).

Ci-contre, le tableau (ou matrice) de *confusion* de reclassement du tableau des notes d'élèves de terminale par la règle des 2-PPV avec la distance euclidienne usuelle.

Actual \ Predicted	D	F	P	Σ
D	2	0		2
F	0	4	0	4
P	0	1	2	3
Σ	2	5	2	9

(a) Pourquoi avoir choisi 2-PPV et pas 1-PPV ?
(b) À votre avis, qu'appelle-t-on *Classification Accuracy* ? *CA*
(c) Ainsi utilisée, est-ce une bonne mesure (globale) d'évaluation, et pourquoi ?

De nombreuses mesures d'évaluation (ou de *performance*) sont issues de la classification binaire et s'étendent au cas multi-groupes.

Pour chaque groupe j, on considère les nombres de : *True Positives*, *True Negatives*, *False Positives* et *False Negatives*, à partir desquels on définit (entre autres) les mesures suivantes :

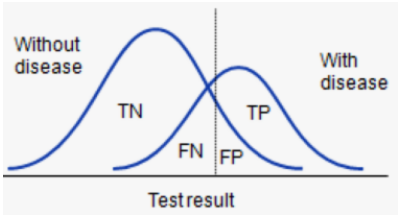
- $R_j = \frac{TP}{TP+FN}$ *Recall*
- $P_j = \frac{TP}{TP+FP}$ *Precision* (TPR)
- $FA_j = \frac{FP}{TP+FP} = 1 - P_j$ *False Alarm* (FPR)
- $F1_j = 2\,\frac{R_j \times P_j}{R_j + P_j}$ *F1 score*

(d) A_j ? *Accuracy*

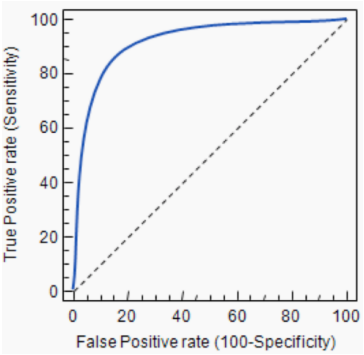
↓ actual / redicted →	*positifs*	*négatifs*
positifs (∈)	TP	FN
négatifs (∉)	FP	TN

Exercice 1

1-1) Calculez les mesures de Rappel, Précision, Fausse Alarme et Justesse des trois groupes d'étudiants (D,P,F).
(HW) Vous calculerez les *F1 scores*.
1-2) Comment faire de ces mesures conditionnelles des mesures globales ?
1-3) Calculez les mesures globales de Rappel, Précision, Fausse Alarme et Justesse.
(HW) Vous calculerez le *F1 score* global.



À l'origine, ces nombres TP, FP, TN et FN proviennent de la séparation signal/bruit et ont été utilisées en diagnostic médical : $1_{x>seuil}$ (figure ci-dessus). Pour différentes valeurs du *seuil* de détection, on peut obtenir autant de valeurs des mesures. La courbe (TPR,FPR) qui en découle s'appelle ROC (*Receiver Operating Characteristics*).



(e) À quoi peut bien servir une telle courbe ?
(f) Faut-il mieux que l'AUC (*Area Under Curve*) soit grande ou petite ?
(g) Orange ne propose pas d'afficher simultanément pour tous les groupes les courbes ROC. En revanche, on peut les afficher pour plusieurs règles.

2. Stratégie(s) d'Évaluation de Règles

Le reclassement introduit bien évidement un biais d'apprentissage (évaluation optimiste). Or, on ne dispose en général que d'un jeu de données et pas deux.

(a) Comment s'en accomoder ?

La solution consiste à échantillonner aléatoirement : $X = X_L \cup X_T$. On fixe en général des % (par ex. : 70-30) ou des effectifs. *hold-out method.*

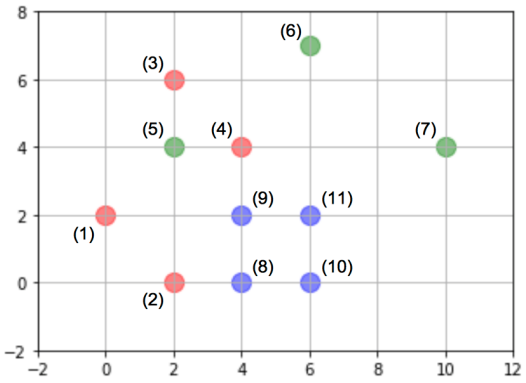
(b) Comment résoudre un éventuel problème de représentativité ? *stratified*
(c) Si on prend 50%-50%, on n'utilise que la moitié des individus. Comment améliorer cet état de fait ?

La *k-validation croisée* (*k-fold cross-validation*) consiste à choisir aléatoirement k échantillons de n/k individus pour X_T, le reste formant X_L, et moyenner les k mesures de performance induites. Une valeur usuelle pour k est : 10.

(d) Que pensez-vous du cas où $k = 2$ en termes de biais d'apprentissage et temps de calcul ?
(e) À quel cas peut se rapporter la méthode dite du *leave-one-out* ?
(f) Finalement, quelle stratégie recommanderiez-vous si on a peu ou beaucoup de données ?

Exercice 2

Soit le jeu de données 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
$X =$	0	2	2	4	2	6	10	4	4	6	6	4.182	6.876
	2	0	6	4	4	7	4	0	2	0	2	2.818	5.234
$Y =$	1	1	1	1	2	2	2	3	3	3	3	$m = 3$	

2-1) Composez, selon les règles de l'art, un ensemble d'apprentissage (X_L, Y_L) et un ensemble de test (X_T, Y_T) en *holdout* à 30% (à peu près).
2-2) Évaluez la règle du 1-NN au sens de la distance euclidienne, puis celle des 3-NN, en calculant la *Classification Accuracy* à partir de la matrice de confusion.
(HW) Recommencez en *holdout* à 50% (à peu près).
(HW) Lorsque vous aurez fait ce (HW), aurez-vous le résultat de 2-validation croisée ?
2-1) Quelle stratégie est la plus recommandée pour ces données ?
2-2) Réalisez visuellement les prédictions par la règle du 3-NN avec la distance euclidienne selon cette stratégie. Vous donnerez la matrice de confusion, puis les scores de *Recall*, *Precision*, et *Classification Accuracy* par classe, puis globaux.

3. Sélection *vs* Extraction d'Attributs

Les méthodes factorielles en classification non supervisée (ACP) ou supervisée (AFD) consistent à construire q nouvelles variables en combinant linéairement[1] les p variables (quantitatives) initiales ; on parle d'*extraction*. Une autre approche consiste à *sélectionner* un sous-ensemble des p variables initiales de cardinalité q. Dans les deux cas, il s'agit aussi de réduire la dimensionnalité des données.

(a) Combien de sous-ensembles de q variables peut-on obtenir à partir de p variables ?

Les méthodes de sélection nécessitent donc :

- un algorithme de recherche — (sous-)optimal, (non) déterministe, (non) séquentiels
- un critère d'arrêt
- une mesure de performance (*scoring*)

(b) Avez-vous des idées pour chaque point ?
(c) Soit une mesure de performance S à maximiser, utilisée avec $q = 1$. Notons $X_{(1)}$, $X_{(2)}$, ..., $X_{(p)}$ les variables rangées dans l'ordre décroissant de S, et considérons deux ensembles de variables $E_1 = \{X_{(1)}, X_{(2)}, X_{(3)}\}$ et $E_2 = \{X_{(1)}, X_{(2)}, X_{(4)}\}$. Pensez-vous qu'on a obligatoirement $S(E_1) > S(E_2)$?
(d) Pourquoi ne pas utiliser pour S une des mesures de performances vues précédemment (Rappel, Précision, etc) ?
(e) Orange ne propose que le cas univarié ($q = 1$), mais donne bon nombre de mesures de performance[2] comme le montre le tableau ci-contre (*widget* `Rank`).

Exercice 3 On souhaite passer de $p = 2$ variables à $q = 1$. Il faut donc garder X_1 ou X_2. Pour décider, on propose d'utiliser la statistique F[3] d'ANOVA qui repose sur les variances inter s_B^2 – et intra s_W^2 – groupes.

3-1) Partagez-vous le calcul des cellules manquantes dans le tableau ci-contre.
(HW) Calculez ce que vous n'avez pas choisi de calculer.
3-2) Afin de compléter le tableau, calculez, la matrice de covariance (intra-classes) du 3ème groupe par produit matriciel.
3-3) Calculez s_B^2 et s_W^2 pour X_1 ou pour X_2.
3-4) Calculez la statistique d'ANOVA $F = \frac{s_B^2/(m-1)}{s_W^2/(n-m)}$ pour la variable choisie.
(HW) Reprenez les calculs pour la variable que vous n'avez pas choisie.
3-5) Connaissant F pour les deux variables, laquelle retenir ?

groupe	stat. par groupe	variables X_1	variables X_2
1	$\overline{x}_j$	2	3
	s_j^2		5
2	$\overline{x}_j$	6	5
	s_j^2	10.67	
3	$\overline{x}_j$	5	1
	s_j^2	1	1

[1] en supervisé, on peut discriminer linéairement (LDA) ou quadratiquement (QDA)
[2] la plupart fondées sur la théorie de l'information
[3] $F = \frac{s_B^2/(m-1)}{s_W^2/(n-m)}$