1 Fiche de TD

Exercice 1

Pourquoi et comment traiter les valeurs manquantes. Expliquer en utilisant des exemples les méthodes traitées au cours et TP.
Pour diverses raisons (oublis, erreurs, non disponibilité, non communiqué,...) des valeurs sont manquantes (absentes), Il parait judicieux de les remplacer par des valeurs estimées pour permettre aux algorithmes d'apprentissage automatique de mieux fonctionner et d'améliorer leurs performances. Sans oublier que certains algorithmes ne fonctionnent pas si le dataset contient des valeurs manquantes.

Réponses :

1. Supprimer tout simplement l'instance contenant des valeurs manquantes
2. Considérer la valeur manquante comme une valeur à part
3. Remplacer par la valeur (moyenne, médiane, mode) des valeurs des attributs de la même classe
4. Remplacer par la valeur la plus fréquente.
5. Faire une prédiction de la valeur manquante en construisant un modèle à l'aide d'un algorithme d'apprentissage KNN ou autres.

Exercice 2

Supposez que l'on veuille développer un modèle de prédiction du succès d'un smartphone. Expliquer en détail comment on procédera dans l'approche data mining. Quel est l'intérêt d'une telle approche ? Proposer un bon résumé de fichier arff pour, selon vous, résoudre ce problème.

Réponses : Créer une base de données en faisant une recherche sur le net ou ailleurs. Puisque la classe est succès, la base contiendra un max d'exemples positifs et négatifs. Chaque smartphone sur le marché devient un exemple. Chaque caractéristique deviendra une valeur d'attribut possible. L'ensemble des caractéristique sera maximum défini et fixé dès le départ.
Intérêt : Seulement à partir d'exemples (positifs et négatifs), des algorithmes dédiés sont capable la connaissance descriptive du concept « smartphone à succès ». Aucun autre apport n'est nécessaire. D'où l'intérêt de cette science

Exercice 3

Donner le pourquoi et le principe des méthodes ensemblistes. Comparer les algorithmes bagging et boosting (vus au cours) étape par étape (apprentissage et classification), donner les similarités et les différences. Expliquez les paramètres de chacun. Quel est le meilleur et pourquoi ?
Réponses :

Pourquoi et principe : Faire Travailler ensemble plusieurs modèles de performances moyennes d'une même application développés sur un même dataset modifiés à chaque fois, ou avec différentes algorithmes d'apprentissage automatique. Dans l'espoir de voir l'ensemble prédire, classifier, clusteriser avec plus d'efficacité que le modèle unique.

BAGGING	BOOSTING
1. Un seul dataset initial 2. Générer n dataset exactement dès le départ (parallèlisable) 3. Appliquer le même algorithme sur tout les datasets (parallelisable) 4. Prédit la classe en utilisant tout les modeles crées. 5. La classe a plus fréquente en cas de classification. La moyenne des réponses en cas de régression	1. Un seul dataset initial 2. N ou moins datasets seront générés les uns après les autres par repondération des instances mal classées par les modèles precedents. 3. Appliquer le meme algorithme de manière sequentielle. 4. Calculer l'erreur sur le dataset pondéré. 5. S'arreter si l'erreur est nulle ou sup »rieur à 0.5 6. Reponderer les instances en utilisant cette erreur avant d'appliquer de nouveau l'algorithme. 7. Nomaliser les poids. 8. Prédit la classe en utilisant tout les modeles crées.

Les paramètres sont l'algorithme et le nombre de modèles à combiner. Le meilleur et pourquoi : Chaque méta algorithme fonctionnera de manière optimale, si des conditions précises sont vérifiées. Pour le bagging, par exemple, on aura besoin d'un algorithme instable, sans oublier que des instances seront repetés ans le dataset construit, alors qu'ils ne le sont pas dans le dataset original. Pour le boosting, par exemple, les instances seront pondérés de manière particulière, d'où la nécessité pour l'algorithme de les gérer.

Exercice 4 :

Selon la kappa statistic, lequel des deux modèles de classification est le plus performant :

		Classe Prédite		
		a	b	c
Classe réelle	a	45	11	9
	b	13	15	6
	c	10	3	32

Table 1 – Matrice de confusion Modèle 1

		Classe Prédite		
		a	b	c
Classe réelle	a	47	8	10
	b	5	16	13
	c	5	5	35

Table 2 – Matrice de confusion Modèle 2

réponse : $KappaM1 = 0.43$
$KappaM2 = 0.50$
Le meilleur modèle est M2

Exercice 5 :

Selon la MAE (mean absolut error), lequel des deux modèles de régression est le plus performant :

Valeurs réelles A_i	12.3	6.5	10.7	7.9	11.5	13.7	12.6	9.4
M1 Valeurs prédites P_i	11.7	7.8	10.9	9	10.8	13.2	11.6	9.3
M2 Valeurs prédites P_i	13	6.2	11.2	8.5	12.1	12.9	12.5	9.2

TABLE 3 – données récoltées

réponse : $MAE1 = 5.5/8 = 0.6875$
$MAE2 = 3.8/8 = 0.475$
Le meilleur modèle est M2