8INF867 – Fondamentaux de l'apprentissage automatique

Travail No. 2 – Dimensionalité et anomalies

1. Travail à effectuer

Dans le cadre de ce travail pratique, vous aurez à utiliser un ensemble de données débalancées appelé « Covertype » qui vise à faire la classification de morceaux de forêts en fonction du type d'arbres dominant. Le dataset est disponible sur UCI :

https://archive.ics.uci.edu/dataset/31/covertype

Voici ce que vous devez faire avec les données :

1. Vous devez visualiser les classes avec T-SNE et noter vos observations.
2. Installez UMAP et comparer la visualisation avec T-SNE. Tentez d'utiliser la version paramétrique (exploitant un réseau de neurone). Vous trouverez les instructions dans la documentation de UMAP.
3. Testez un algorithme de classification de votre choix avec minimalement l'Accuracy et le F1-Score. Vous pouvez choisir de prétraiter les données ou non. Attention! Une mauvaise méthodologie pourrait affecter vos résultats.
4. Vous devez ensuite installer Imbalanced-Learn. En premier lieu, vous devez exploiter une méthode d'*under-sampling* pour équilibrer vos données. Répétez les étapes 1, 2, 3. Qu'observez-vous?
5. Enfin, vous devez tester avec SMOTE pour équilibrer vos classes par *over-sampling*. Répétez de nouveau les étapes 1, 2, 3. Qu'observez-vous par rapport à précédemment?

2. Remise

Le travail devra être déposé sur Moodle dans la section « Travaux et dépôts » sous l'intitulé Travail pratique #2. Attention! tout travail en retard recevra une pénalité minimale de 20% jusqu'à possibilité d'obtenir zéro.

Vous devez me remettre une copie de votre code et un document contenant minimalement une page titre (vos noms, codes permanents et titre du cours), les contraintes de fonctionnement (librairies nécessaires), vos observations/résultats ainsi qu'une brève explication de votre travail.

Équipe : Équipe de 2 à 3 personnes

Date limite de remise : Vendredi le 22 novembre à 23h59 (EST).

3. Évaluation

Le travail vaut 7.5% de votre note finale. Voici la grille d'évaluation que j'utiliserai pour vous noter. Il s'agit d'une évaluation *négative*, c'est-à-dire que vous démarrez avec 100% et que chaque élément peut vous faire perdre un certain nombre de points :

Dimension	Élément spécifique	Pénalité max
Code	Code difficile à lire	2
	Dépendances non justifiées	1
	Bogues	5
Fonctionnalités	Ensemble de données	3
	UMAP	3
	Under-sampling	3
	SMOTE	3
	Méthodologie	3
Qualité de la remise	Présence des éléments demandés	2
	Qualité du français (ou de l'anglais)	2