Apprentissage statistique : TP noté

Contexte Les fermes de haute précisions sont de grandes fermes agricoles dans lesquelles certaines tâches, autrefois réalisées manuellement par les éleveurs, sont automatisées. Dans un contexte de bien-être animal, une ferme de haute précision cherche à détecter automatiquement le comportement anormale de ses vaches laitières. Ce comportement anormal peut correspondre à différentes situations : maladies, chaleurs, perturbation du parc comme par exemple une alarme,... et demande le plus souvent une action de la part de l'éleveur.

L'étude a été réalisée en hiver, lorsque les vaches sont à l'étable. L'étable se caractérise par trois grandes salles : la première salle contient des logettes où les vaches peuvent se reposer, la second salle des auges dans lesquelles les vaches peuvent manger et enfin il existe des allées où elles peuvent se promener.

Pour obtenir des données, l'éleveur a positionné des capteurs sur les vaches et a récupéré pour chaque heure, le nombre de secondes dans chaque salle.

Format Le jeu de données est composé de trois fichiers : *xTrain.csv*, *yTrain.csv* et *xEval.csv*. Les fichiers *xTrain.csv* et *xEval.csv* correspondent aux caractéristiques qui doivent permettre la prédiction d'un comportement anormal ou non. Les fichiers contiennent les attributs suivants :

- *date_hour* : date et heure du premier enregistrement
- *idCow* : l'identifiant de la vache
- *all* + numéro i entre 0 et 23 : nombre de secondes dans les allées pour l'heure *data_hour + i*
- *rest* + numéro i entre 0 et 23 : nombre de secondes dans les logettes pour l'heure *data_hour + i*
- *eat* + numéro i entre 0 et 23 : nombre de secondes dans la salles des auges pour l'heure *data_hour + i*

Le fichier *yTrain.csv* comprend l'état de la vache après les 24 heures. Les valeurs sont : 0 si son état est considéré comme normal, 1 sinon.

Remarques

- Les capteurs positionnés sur les vaches n'émettent pas tout le temps. Il est donc courant que la somme des attributs *all*, *rest* et *eat* pour une heure donnée fasse moins de 3600 secondes.
- Le fichier *xTrain.csv* contient des valeurs manquantes (sous forme de NaN) alors que *xEval.csv* a déjà été nettoyé.
- Les classes sont très déséquilibrées.

Consignes Vous devez trouver le modèle qui vous permettra la meilleure classification. Pour cela, vous pouvez combiner plusieurs algorithmes vus en cours et/ou modifier des algorithmes vus en cours. Attention, l'utilisation de réseaux de neurones n'est pas permise.

Vous avez la possibilité d'envoyer une fois votre résultat de classification sous forme d'un fichier csv de la même forme que *yTrain.csv*.

Attention, il faut envoyer ce résultat minimum 5 jours avant la date de rendu ! L'évaluation correspondra à la fonction sklearn.metrics.f1_score et la matrice de confusion.

Rendu Vous fournirez sur moodle **une prédiction sous format csv** et un **rapport** sous le format pdf (et pas un notebook).

Rapport L'explicabilité de vos résultats est important. Votre rapport sera noté en fonction de :

- La fluidité dans l'explication de votre pipeline pour la prédiction (un schéma peut aider)
- Les détails permettant la reproduction (sans regarder le code) de votre pipeline
- Les justifications sur le choix de la pipeline (qui doit être bien sûr choisi en fonction des données et appuyées par de la bibliographie [1])
- L'explication du choix des paramètres et le protocole expérimentale mis en place
- L'interprétation des résultats et une perspective de travail
- L'originalité de la méthode proposée (par rapport aux autres groupes)
- Le résultat de la classification (notez cependant que ce point est moindre par rapport aux autres)

Pour vous aider dans la rédaction du rapport, vous pouvez suivre le plan suivant : introduction (problématique), analyse des données brutes, pipeline, résultats, conclusion et perspective. Notez bien également dans votre rapport le résultat intermédiaire que je vous aurais fournis et les conclusions que vous en tirez.

1. Le support de cours peut faire référence de bibliographie.