ÉCOLE POLYTECHNIQUE DE MONTRÉAL

TRAVAIL DE SESSION : PREMIÈRE PARTIE

PAR
SAMUEL FERRON 1843659
JEAN-FRÉDÉRIC FONTAINE 1856632
PATRICK ST-PIERRE 1840634

DANS LE CADRE DU COURS MTH2302D
(PROBABILITÉS ET STATISTIQUES POUR INGÉNIEURS)
VENDREDI 6 OCTOBRE 2017

Contexte général des données:

Tout d'abord, il est pertinent de mentionner que chaque membre de notre équipe est inscrit au programme de génie logiciel. Ceci nous a permis de rapidement trouver un point d'intérêt commun qui se prêterait bien au travail de session. Effectivement, nous sommes tous les trois très intéressé par l'apprentissage machine (AM). Ce domaine de l'informatique connait un essor faramineux depuis la dernière décennie. Notre objectif est donc d'utilisé les notions de régression linéaire et multiple (joint aux tests d'hypothèses) afin de s'initier à l'AM. Ainsi, nous avons exploré plusieurs banques de données utilisées afin de développer et tester des algorithmes d'AM. Notre choix s'est arrêté sur le jeu de données « *Forest Fires* » disponible sur le site de l'université Irvine (Californie) [1]. En plus de respecter les exigences du travail de session, il semble que ces données se prêtent bien aux analyses de régression [2].

Plus précisément, ces données décrivent plusieurs instances de feux de forêts sur le territoire du parc naturel de Montesinho situé au nord-est du Portugal. La faune et la flore de cette région présentent une grande diversité ce qui en fait un sujet d'étude intéressant. Ces données ont étés recueillies entre janvier 2000 et décembre 2003. D'ailleurs, il sera intéressant d'analyser ceux-ci afin d'en apprendre d'avantage sur le phénomène des feux de forêts, leurs causes et les éléments catalyseurs.

Provenances des données

Comme mentionné dans la section précédente, les données proviennent du répertoire de jeux de données publié par l'université d'Irvine [1]. La première partie des données a été récoltée par l'inspecteur responsable de l'observation des feux de forêt du parc Montesinho. Ainsi, pour chaque feu, l'inspecteur a pris soin de noter l'heure, la date et le type de végétation du secteur en question. De plus, la position de chaque feu a été enregistré selon un système de coordonné *x,y* (Figure1). Les autres variables mesurées par l'inspecteur concernent les métriques d'indice forêt météo[1] (IFM). Ces métriques proviennent du système Canadien servant à l'évaluation du niveau de dangerosité des feux de forêts [3]. Ce système consiste en 6 variables (FFMC, DMC, DC, ISI, BUI et FWI) qui sont explicitées dans la prochaine section. Ensuite, la deuxième partie des variables a été récoltée par l'institut Polytechnique de Brangança. Celles-ci correspondent à quelques observations météorologiques (e.g vitesse du vent) enregistrées par l'institut météo du parc.

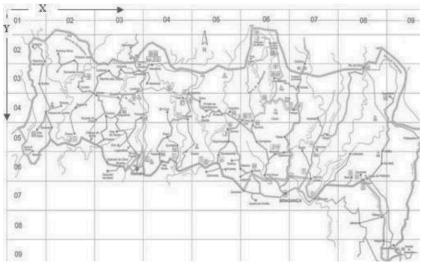

Figure 1 Carte du parc Montesinho

[1] De l'anglais "Forest Fire Weather Index" .

Description des données

Chaque instance de feu (518) a été examinée selon 13 attributs. Nous avons les informations sur la température lors des feux, leur position, leur superficie et les composantes de l'Indice Forêt-Météo (IFM). Les descriptions concernant l'IFM ont étés tirés du site de Ressources Naturelles du Canada [3]. Ici-bas, nous avons énumérés les attributs du jeu de données de la même façon qu'ils sont organisés dans le fichier Excel.

1. X - La coordonnée spatiale sur l'axe x (figure 1) dans la carte du Montesinho park : 1 à 9 (**discret**).

2. Y - La coordonnée spatiale sur l'axe y (figure 1) dans la carte du Montesinho park : 2 à 9 (**discret**).

3. Month (mois) - Le mois de l'année : Janvier à Décembre (**discret**).

4. Day (jour) - Jour de la semaine : Lundi à Vendredi (**discret**).

5. FFMC (Indice combustible léger) - Évaluation numérique qui donne une indication de l'inflammabilité du combustible léger : 18.7 à 96.20 (**continue**).

6. DMC (Indice d'humidité de l'humus) - Évaluation numérique qui donne une indication du facteur de combustion d'un combustible dans les couches organiques de moyenne épaisseur et les matières ligneuses de taille moyenne : 1.1 à 291.3 (**continue**).

7. DC (Indice de sècheresse) - Évaluation numérique sur la sècheresse des combustibles forestiers et du degré de latence du feu dans les épaisses couches organiques et les grosses billes : 7.9 à 860.6 (**continue**).

8. ISI (Indice de propagation initiale) - Évaluation numérique du taux prévu de propagation du feu : 0.0 à 56.10 (continue)

9. temp - Température en degré Celsius : 2.2 à 33.30 (**continue**).

10. RH - L'humidité relative en % : 15.0 à 100 (**continue**).

11. wind (vent) - Vitesse du vent en km/h : 0.40 à 9.40 (**continue**).

12. Rain (pluie) - Pluie en mm/m2 : 0.00 à 6.4 (**continue**).

13. Area (superficie) - La superficie de forêt brulé en ha : 0.0 à 6.4 (**continue**).

Malgré le fait que nous n'avons pas toutes les informations correspondant à l'IFM, chaque instance de feu nous donne une bonne quantité d'information pour faire des analyses pertinentes et intéressantes.

Questions ouvertes

Comme mentionné dans la section d'ouverture, notre objectif est d'utiliser les notions que nous allons voir prochainement dans le cours afin de s'initier aux bases de l'apprentissage machine. Ainsi, après avoir discuté avec le chargé de cours, nous en sommes venu à la conclusion qu'il serait intéressant de faire des analyses de régression linéaire entre les variables afin d'extirper de l'information quant au propriétés des feux de forêts. Ces analyses seraient jointes à des tests d'hypothèses afin de vérifier si le taux de variation entre les variables est significatif. De plus, il serait pertinent de mesurer l'intervalle de confiance des différentes pentes à l'étude afin de préciser la pertinence des relations. Ainsi, une fois ces analyses faites, nous pourrions essayer de faire des prédictions sur des variables connus qui ont été isolées de nos modèles. Cette notion d'isoler des observations du modèle afin d'en prédire ses caractéristiques ultérieurement est le fondement même de l'apprentissage machine supervisé. Plusieurs questions pourraient être ainsi exploré. Par exemple, existe-t-il une relation entre le niveau d'humidité du humus et l'indice de propagation initiale d'un feu ? Quel est le lien entre la superficie d'un feu de forêt et son indice de combustible léger ? Nous serait-il possible de prédire les endroits à risque grâce à ces relations ? Finalement, il serait aussi intéressant de faire des analyses de régressions multiples afin de voir les relations entre plusieurs variables.

Références :

[1] University Irvine of California, Machine Learning Repository (2017). [En ligne]. Disponible : http://archive.ics.uci.edu/ml/index.php

[2] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data", dans "New trends in artificial intelligence". 2007. [En ligne]. Disponible: https://repositorium.sdum.uminho.pt/handle/1822/8039

[3] Système canadien d'information sur les feux de végétation, Ressources naturelles Canada (2017). [En ligne]. Disponible: http://cwfis.cfs.nrcan.gc.ca/renseignements/sommaire/fwi