Université Côte d'Azur
Enrico FORMENTI

Année 2021/22
M1 Info
TP NNL

Pre-traitement (Pre-processing)

1 Le problème des données manquantes (suite et fin)

Dans la feuille précédente nous avons vu quelques méthodes pour palier au problème des données manquantes. Nous nous sommes aussi rendu compte d'à quel point ça puisse être compliqué dans certains cas.

Dans cette feuille, nous proposons une nouvelle méthode de traitement des valeurs manquants et chercherons à en estimer la qualité de ses résultats.

Exercice 1

1. Reprenons la base contenue dans le fichier `winemag-data-130k-v2.csv`, supposons d'avoir fait toutes les opérations demandées dans la feuille précédente jusqu'à la fin de l'exercice 4.
2. Reprenez les valeurs d'origine pour la colonne `prix`, y compris les 'NA'.
3. Re-écrivez une fonction `distance(L1,L2,nom)` qui calcule la distance entre les lignes `L1` et `L2` mais sans prendre en compte la valeur de la colonne `nom`.
4. Ecrivez une fonction `getPrix(k,L, nom)` où `k` est un entier non-nul, `L` est un numéro de ligne de la base et `nom` est une chaîne de caractères indiquant le nom d'une colonne. Le but de cette fonction est de sélectionner les `k` lignes les plus *proches* de `L` disons $\mathtt{L}_1$, $\mathtt{L}_2$, ..., $\mathtt{L}_k$ grâce à la fonction `distance` sans prendre en compte les valeurs de la colonne `nom`. Ensuite, la fonction renvoie la valeur moyenne de $\mathtt{L}_1\mathtt{[nom]}, \mathtt{L}_2\mathtt{[nom]}, \ldots, \mathtt{L}_k\mathtt{[nom]}$ approchée à l'entier supérieur *ie.*
$$\mathtt{val} = \left\lceil \frac{\sum_{i=1}^{k} \mathtt{L}_i\mathtt{[nom]}}{k} \right\rceil$$
5. Ajoutons à présent à notre base une nouvelle colonne que nous allons appeler `cp` (Catégorie de Prix). Soit M le max de la colonne `prix`. Supposons de vouloir diviser les vins en 5 catégories de prix allant de 1 pour les moins chers à 5 pour les plus chers. La première plage contiendra tous les vins ayant un prix entre 0 et $\lfloor \frac{M}{5} \rfloor$. La deuxième tous ceux ayant un prix entre $\lfloor \frac{M}{5} \rfloor + 1$ et $\lfloor \frac{2 \cdot M}{5} \rfloor$. Et ainsi de suite pour les autres plages. Une fois la colonne remplie `cp`, créez une colonne `pp` (prix prévu) et utilisez la fonction `getPrix(k, L, 'prix')` pour calculer un nouveau prix attendu pour chaque vin.

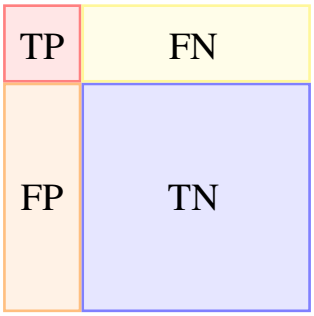
Dans le prochain exercice on cherchera d'estimer la qualité du travail de notre fonction `getPrix()`, c'est-à-dire on cherchera à juger jusqu'à quel point elle est capable d'attribuer une catégorie de prix correcte à un vin.

Exercice 2

1. Définissez les compteurs suivants `TP`, `TN`, `FP` et `FN` et mettez-le à zéro.
2. Construisez une matrice `M` où chaque élément $\mathtt{m}_{ij}$ représente le nombre de lignes de la base pour lesquelles la valeur de la colonne `pp` tombe dans la plage j alors que la valeur de la colonne `prix` est dans la catégorie i. Bien sûr, on fera cela seulement pour toutes les lignes dont le prix initial est différent de 'NA'.
3. Définissons à présent, `TP` comme $\sum_{i=1}^{5} m_{ii}$; `FN`=$\sum_{\substack{j=1 \\ j \neq i}}^{5} m_{i,j}$; `FP`=$\sum_{\substack{k=1 \\ k \neq i}}^{n} m_{k,i}$ et `TN`=`Total-FP-FN+TP` où `Total` est le nombre total d'éléments dans `M` *ie.* `Total`=$\sum_{i=1}^{n} \sum_{j=1}^{n} m_{i,j}$.

Distribué le 30 septembre 2021.

La figure suivante illustre graphiquement la structure de la matrice `M`.



Notre nouvelle méthode sera d'autant plus fiable qu'elle concentre les valeurs dans `TP` et `TN`.

Exercice 3

Il est clair que la méthode est sensible à la valeur du paramètre `k`. Proposez une méthode pour estimer la meilleure valeur de `k` à utiliser.