Traitement de données

mercredi 16 mars 2022 10:42

Dans un dataset, on a souvent des valeurs absentes, des trous…

Si on a une donnée absente, on ne peut pas faire comme si on l'avait observé.
Quelles sont les stratégies à employer mm en présence de valeurs manquantes ?

=> stratégies pertinentes quand on a quelques 10zaine de % de valeurs manquantes…
=> d'autres pb apparaissent avec les données réelles => outlayers (valeurs extrêmes, éloignées de celle que l'on attend en fo nction de notre point de vue, anomalies…) => il y a des techniques de détection d'outlayers

A cause de ces valeurs, certains algo peuvent refuser de fonctionner (ACP par exemple)…

Il est donc nécessaire de s'occuper de ce type de valeurs (sur Kaggle par exemple, cela a déjà été traité pour que le dataset soit parfaitement exploitable).

DEFINITION :
Quelle est une valeur manquante dans un jeu de donnée ?
Information que j'aurais aimé avoir, mais que je n'ai pas eu… information qui m'aurait été utile mais que je n'ai pas eu.
=> observations, éléments du modèle dont je ne dispose pas (ex : pour une classification des individus, la variable qui assig ne un individu à une classe est manquante et je veux trouver la valeur de cette variable)
=> la classification est donc une résolution de valeurs manquantes
=> variables latentes : on veut évaluer la satisfaction d'un client d'une marque de téléphone portable => le score de satisfa ction est une variable latente puisque ce n'est pas mesurable directement

=> on regarde l'aspect des valeurs manquantes dans le sens où elle est manquante dans notre jeu de donnée

Comment elles apparaissent ? Quelles sont les mécanismes d'apparition des valeurs manquantes ?
3 mécanismes qui induisent des outils différentes.

- 2 types de valeurs manquantes aléatoires : on imagine que les valeurs manquantes arrivent aléatoirement dans le jeu de donnée s => on peut remplacer les trous par des valeurs, supprimer les individus concernées => **MCAR**
- **Déterministes** => attention car peuvent fausser l'analyse ! => il faut connaitre le mécanisme avec lequel les données ont été collectées => problème d'outil de mesure ?

===> Valeurs manquantes
- %tage de valeurs manquantes
- Type : MCAR (+ simple => le probabilité que la valeur soit manquante est = pour toutes les observations), MAR (la probabilité d'être manquant dépend des autres variables), MNAR (pas aléatoire, mesure impossible par exemple…)

Existe-t-il des outils pour déterminer le type de valeurs manquantes présent dans le dataset ?
Oui, il y a des outils de visualisation pour tenter de répondre à cette question.
On peut calculer la proportion de valeurs manquantes par variable.
Comment faire des hypothèses réalistes ?

J'ai des valeurs manquantes, je veux savoir comment elles apparaissent, mais pourquoi ?
=> Stabiliser l'analyse
=> Amélioration continue du processus
=> Eviter les trous dans les données car les analyses ne les supportent pas, on veut donc les boucher => on a vraiment besoin de ça ?
=> quel algo utiliser en fonction de ce type ?

=====> visualisation, détermination de la nature des valeurs manquantes
=====> imputation, remplacer ces valeurs manquantes par des valeurs plausibles

---

https://fbertran.github.io/ESIEA_MD/
MDP : lob-OLDER-deltoid-consent

=> Supports de cours

ID	Y1	Y2	Y3	Y4	Y5
1	0.51	0.85	0.31	0.60	0.89
2	0.42	0.35	0.76	0.27	0.45
3	0.34	0.07	0.08	0.49	0.24
4	0.35	0.15	0.90	0.73	0.21
5	0.41	0.37	0.53	0.36	0.63
6	0.50	0.79	0.38	0.04	0.06
7	0.69	0.53	0.17	0.75	0.90
8	0.74	0.33	0.30	0.55	0.02
9	0.36	0.11	0.43	0.91	0.04
10	0.58	0.63	0.15	0.26	0.33
11	0.15	0.61		0.14	0.45
12	0.18	0.36		0.48	0.01
13	0.30	0.89		0.32	0.06
14	0.68	0.18		0.19	0.36
15	0.26	0.79		0.23	0.27
16	0.17	0.52		0.95	0.35
17	0.48	0.60		0.47	0.98
18	0.80	0.09		0.43	0.34
19	0.96	0.23		0.61	0.43
20	0.06	0.55		0.89	0.71
21	0.85	0.55		0.44	0.91
22	0.84	0.84		0.73	0.58
23	0.01	0.90		0.79	0.61
24	0.32	0.88		0.86	0.69
25	0.50	0.37		0.36	0.13
26	0.22	0.27		0.71	0.46
27	1.00	0.23		0.22	0.24
28	0.42	0.24		0.85	0.90
29	0.13	0.06		0.21	0.33
30	0.34	0.84		0.94	0.52

(A)

ID	Y1	Y2	Y3	Y4	Y5
1	0.51	0.85	0.31	0.60	0.89
2	0.42	0.35	0.76	0.27	0.45
3	0.34	0.07	0.08	0.49	0.24
4	0.35	0.15	0.90	0.73	0.21
5	0.41	0.37	0.53	0.36	0.63
6	0.50	0.79	0.38	0.04	0.06
7	0.69	0.53	0.17	0.75	0.90
8	0.74	0.33	0.30	0.55	0.02
9	0.36	0.11	0.43	0.91	0.04
10	0.58	0.63	0.15	0.26	0.33
11	0.15	0.61	0.49	0.14	
12	0.18	0.36	0.76	0.48	
13	0.30	0.89	0.16	0.32	
14	0.68	0.18	0.82	0.19	
15	0.26	0.79	0.60	0.23	
16	0.17	0.52	0.68		
17	0.48	0.60	0.45		
18	0.80	0.09	0.97		
19	0.96	0.23	0.64		
20	0.06	0.55	0.24		
21	0.85	0.55			
22	0.84	0.84			
23	0.01	0.90			
24	0.32	0.88			
25	0.50	0.37			
26	0.22				
27	1.00				
28	0.42				
29	0.13				
30	0.34				

(B)

ID	Y1	Y2	Y3	Y4	Y5
1	0.51	0.85	0.31	0.60	0.89
2		0.35		0.27	
3	0.34	0.07		0.49	0.24
4	0.35	0.15	0.90	0.73	0.21
5	0.41	0.37	0.53	0.36	0.63
6	0.50	0.79		0.04	
7	0.69	0.53	0.17	0.75	0.90
8	0.74		0.30	0.55	0.02
9	0.36	0.11		0.91	0.04
10		0.63	0.15	0.26	0.33
11	0.15	0.61	0.49	0.14	
12	0.18	0.36	0.76	0.48	0.01
13	0.30	0.89	0.16		0.06
14	0.68	0.18	0.82	0.19	0.36
15	0.26		0.60		0.27
16	0.17	0.52	0.68	0.95	0.35
17		0.60		0.47	0.98
18	0.80	0.09	0.97	0.43	0.34
19		0.23	0.64	0.61	0.43
20	0.06		0.24	0.89	0.71
21	0.85	0.55	0.77	0.44	
22	0.84	0.84		0.73	0.58
23	0.01	0.90	0.55	0.79	0.61
24	0.32			0.86	0.69
25	0.50	0.37	0.81		0.13
26		0.27	0.45	0.71	0.46
27	1.00	0.23		0.22	
28	0.42			0.85	0.90
29	0.13	0.06		0.21	0.33
30	0.34	0.84	0.24	0.94	0.52

(C)

A => **univarié** => donc pas complètement aléatoire => **MNAR** (
lorsque pour une variable donnée, si une observation est absente, alors toutes les observations suivantes pour cette variable sont absentes)

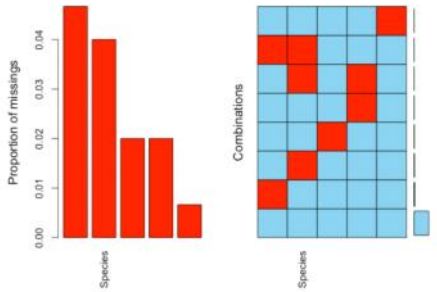
B => **manquant monotone** => pas aléatoire => dès qu'on a une première valeur manquante, les suivantes sont manquantes aussi, pour plusieurs variables (ex, capteur défaillant)

C => aléatoire => **manquant arbitraire => MCAR**

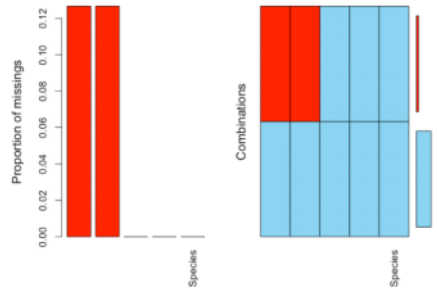
On peut faire de la visualisation pour représenter ces valeurs manquantes.
Rouge = % de variables manquantes par variable

```
##
##  Variables sorted by number of missings:
##       Variable       Count
##   Sepal.Width 0.046666667
##       Species 0.040000000
##  Petal.Length 0.020000000
##   Petal.Width 0.020000000
##  Sepal.Length 0.006666667
```

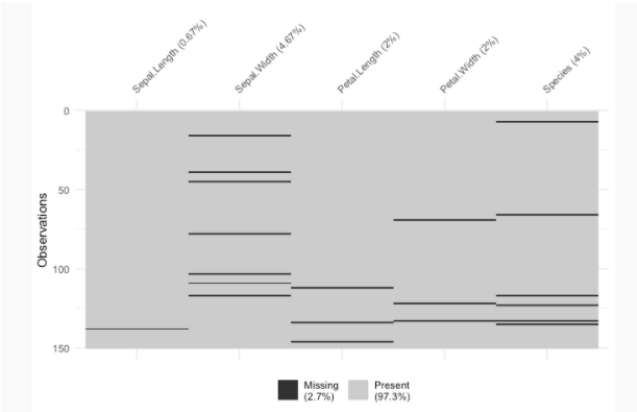
```
summary(aggr(iris_NA, sortVar = TRUE))
```



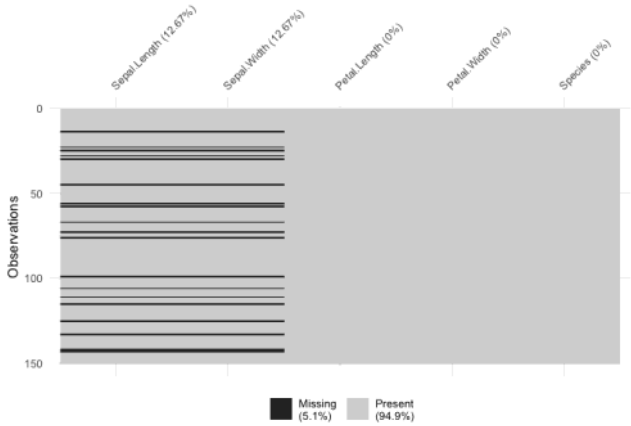
=> MCAR



=> MNAR // monotone

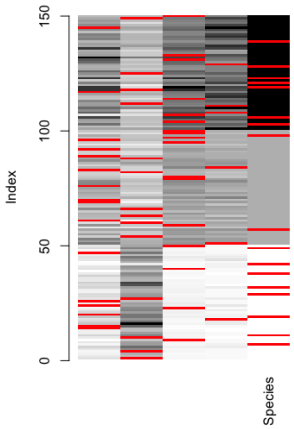


=> MCAR

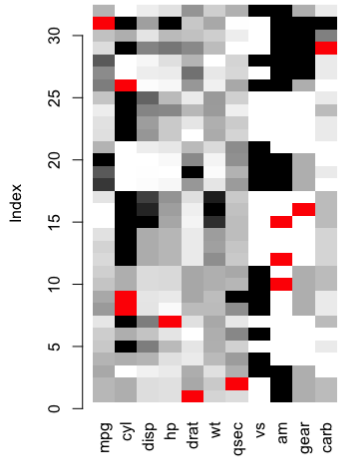


=> MAR

Il est possible de visualiser les données manquantes en rouge, tandis que les données présentes prennent un niveau de gris se lon leur valeur.



=> **MCAR : manquantes dans toutes les colonnes**
Lorsque la probabilité qu'une valeur soit manquante ne dépend ni des valeurs observées, ni de celles manquantes, les données sont dites manquantes complètement au hasard (* MCAR, missing completely at random*). La probabilité d'absence est donc la même pour toutes les observations et elle ne dépend que de paramètres extérieurs indépendants de cette variable.



=> **MAR : (manquantes dans certaines colonnes/variables)**

Lorsque la probabilité qu'une valeur soit manquante dépend uniquement de la composante observée "O" (une ou plusieurs variables observées) mais pas des valeurs manquantes elles-mêmes, les données sont dites manquantes au hasard (* MAR: missing at random*).

1 case = 1 observation
1 variable = 1 feature

Pour traiter les données manquantes :

=> CLEAN LE DATASET (enlever les NA)

Analyse des cas complets :
On enlève les individus/observations (lignes) pour lesquels on a des manquants.
On peut le faire sous deux conditions :
=> les données à supprimer doivent être en faible quantité
=> on ne peut faire ça que sur des données manquantes complètement aléatoires (MCAR)

Analyse des cas disponibles :
On enlève les colonnes/variables pour lesquels on a des manquants.

Il faut jauger si on doit enlever une variable ou les cas.

Imputation :
On a plusieurs stratégie :

- **Imputation simple/univariée** : remplacer le manquant par une seule valeur (par plus proche voisin, par la moyenne de la variable, par la médiane, par forêt aléatoire) => remplace les données manquantes par des valeurs provenant d'individus similaires pour lesquels toute l'information a été observée. L'imputation peut aussi se faire par régression en remplaçant les valeurs manquantes par des valeurs prédites selon un modèle de régression.
  On estime que si la fraction des observations avec des données manquantes est inférieure à par exemple 5% (10-15%), **et le mécanisme est ignorable (MCAR ou MAR)**, on reste en imputation simple même si l'imputation multiple peut-être privilégiée (car en imputation simple, l'estimateur peut-être biaisé si une valeur est systématiquement manquante)

- **Imputation multiple** : remplacer le manquant par plusieurs valeurs. On le fait avec le package mice de R => gérer l'incertitude dû à l'imputation => faire des imputations multiples => **on le fait que si l'analyse en cas complet est possible**
  - method = rf
  - m (nombre de jeux de données) est à choisir en fonction du FMI (fraction missing information) de l'échantillon (=% de NA dans l'échantillon). On choisit m = max(FMI,5) -> même si FMI < 5, toujours au moins 5 jeux de données imputés.
    Règles de Rubin => 1ère règle de Rubin => on fait la moyenne des estimations sur les différents jeux de données
    => on calcule la moyenne des variances de la moyenne sur les jeux de données imputées
    => on regarde l'écart-type associé
    => évaluer l'incertitude liée à l'imputation => si elle est faible => c'est rassurant car l'effet des valeurs manquantes sur la moyenne est limitée

On prend notre dataset et on divise en 5 jeux de données minimum => on impute les NA avec la moyenne (imputation simple) => on calcule la moyenne des variances de la moyenne sur les 5 jeux de données => on regarde l'écart-type associé => on évalue l'incertitude liée à l'imputation => si faible => TROP BIEN.

6 datasets	Moyenne
1	5,854
2	5,843
3	5,851
4	5,845
5	5,845
6	5,847

<u>Pour traiter les données aberrantes/extrêmes :</u>

Définition :

En analyse univariée/complet, une valeur aberrante est une “donnée observée” pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l’échantillon = raisonnement en ligne

En analyse multivariée/dispo, l’échantillon aberrant résulte d’une erreur importante se trouvant dans un des composants du ve cteur de réponse, ou de petites erreurs systématiques dans chacun de ses composants, et qui de ce fait, ne partage pas les relations entre les variab les de la population = raisonnement en colonne

L’examen des valeurs aberrantes dans une base de données a pour objectif de les identifier pour soit les supprimer, soit les conserver, ou les corriger avant d’ajuster des modèles non robustes.

La valeur extrême peut être liée à un événement atypique, mais néanmoins connu et intéressant à étudier. Dans ce cas elle est importante à conserver.

La correction (ou accommodation) évite le rejet des observations aberrantes et consiste à estimer les valeurs des paramètres de la distribution de base de façon relativement libre sans déformation des résultats liés à leur présence.

Pour les observer graphiquement => boxplot ou boîte à moustache => regarder la distance interquartile // cote Z