A1: Preprocessament de les dades

Enunciat

6 de març, 2019

Introducció

Les dades que s'analitzen en aquesta activitat corresponen a una anàlisi de satisfacció laboral. L'objectiu és investigar si la satisfacció laboral dels treballadors està relacionada amb la qualificació del treball i el nivell d'estudis, entre d'altres variables.

Les dades recollides per a cada treballador són:

1. La ciutat del lloc de treball
2. El sexe del treballador
3. El nivell d'estudis del treballador (1: sense estudis, 2: estudis primaris, 3: educació secundària o educació professional, 4: universitaris)
4. El tipus de treball (C: qualificat, PC: poc qualificat)
5. La satisfacció laboral del treballador (compresa entre 0 i 10)
6. L'edat (anys)
7. L'antiguitat (en anys)
8. Els dies de baixa en el darrer any laboral
9. Si el treballador va estar de baixa o no en el darrer any (variable binària)
10. Les hores que treballa a la setmana en promig
11. El nivell de colesterol en la penúltima revisió mèdica (mmol/L)
12. El nivell de colesterol en la darrera revisió mèdica (mmol/L).

L'arxiu es denomina *rawData.csv*, i conté 1920 registres i 12 variables. Aquestes variables són: city, sex, educ_level, job_type, happiness, age, seniority, sick_leave, sick_leave_b, work_hours, Cho_initial, Cho_final

Preprocessament de les dades

L'objectiu concret d'aquesta activitat és preparar el fitxer per a la seva posterior anàlisi. Per a guiar l'activitat us suggerim quins tipus de preprocessament cal fer per a resoldre satisfactòriament aquesta activitat. A continuació, us expliquem quins són els passos necessaris:

1. Carregueu l'arxiu de dades en R i feu una breu descripció del mateix, on s'indiqui el nombre de registres llegits, el nombre de variables i els noms de les variables. Recordeu que abans de carregar l'arxiu, cal inspeccionar quin tipus de format csv es tracta per tal que la seva lectura sigui apropiada.
2. Indiqueu el tipus de variable estadística de cada variable.
3. En el cas que R no assigni el tipus apropiat a alguna variable, realitzeu la conversió necessària per tal que el tipus final de cada variable sigui l'apropiat. Per exemple, possibles errors de variables quantitatives amb confusió en el separador decimal.
4. Normalitzeu / Estandaritzeu les variables quantitatives.

5. Normalitzeu / Estandaritzeu les variables qualitatives.
6. Reviseu possibles inconsistènciess entre variables: age versus seniority i número d'hores setmanals promig amb valor negatiu.
7. Busqueu valors atípics en les variables quantitatives.
    i. Presenteu un boxplot per cada variable quantitativa.
    ii. Realizeu un quadre amb les estimacions robustes i no robustes de tendència central i dispersió per a cada variable quantitativa.
8. Valors perduts
    i. Busqueu quines variables i quins registres tenen valors perduts.
    ii. Imputeu els valors a partir dels k-veïns més propers usant la distància de Gower amb la informació de totes les variables.
9. Realitzeu un breu estudi descriptiu de les dades un cop depurades.
10. Finalment, emmagatzemeu les dades en un arxiu de dades corregit.

Criteris de preprocessament

Quan es realitza un preprocessament, l'analista decideix com estandaritzar/normalitzar les variables. Per això, estableix uns criteris, que idealment s'escriuen per a homogeneïtzar versions o per a posteriors preprocessaments que apareguin sobre noves dades. A continuació, us indiquem els **criteris a seguir**:

- El punt (.) és el separador decimal de qualsevol variable numèrica.
- Els valors de la variable **work_hours** s'estandaritzen en una sola xifra decimal.
- Els noms de les ciutats s'estandaritzen amb la primera lletra de cada paraula en majúscules (excepte les preposicions) i la resta de lletres en minúscula. Per exemple: Barcelona, Las Palmas de Gran Canaria, ...
- Els valors de gènere s'estandaritzen com M i F.
- El nivell d'estudis s'estandaritza passant el número a un caràcter: 1 (N), 2 (P), 3 (S), 4 (U).
- El tipus de treball s'estandaritza amb la següent abreviatura: 1 (C), 2 (PC), on C correspon a qualificat i PC a poc qualificat.
- La inconsistència entre les variables **age** i **seniority** es produeix quan un treballador té una diferència d'edat (age) i antiguitat (seniority) menor de 18 anys. Si és així, el criteri de correcció és assignar a la variable antiguitat l'edat del treballador menys 18 anys.
- Una altra inconsistència pot ser la provocada per valors negatius en el nombre d'hores setmanals de promig (**work_hours**). El criteri de correcció és passar el valor a positiu.

Comentaris importants

1. **No** es pot inspeccionar ni corregir de manera manual el fitxer de dades. Per exemple, **no** es pot fer una assignació del tipus:

```
`mydata[1,5] <- 32.5`
```

Aquest tipus de transformacions s'han de fer amb funcionalitats de búsqueda (cercar els registres que tenen errors o inconsistències) i després fer les correccions oportunes amb funcionalitats d'R. Així el procediment de neteja és útil, independentment del fitxer de dades i de la posició i valors concrets del fitxer.

2. **No** es poden fer llistats complets de les dades del fitxer a pantalla, perquè genera fitxers de sortida excessivament grans. Si voleu testejar el resultat d'una instrucció sobre les dades, podeu usar la funció **head** que mostra les primeres files de la taula de dades.

Puntuacions dels apartats

- Apartats 1 i 2 (10%)
- Apartats de 3 a 5 (20%)
- Apartat 6 (10%)
- Apartat 7 (20%)
- Apartat 8 (10%)
- Apartats 9 i 10 (20%)
- Qualitat de l'informe dinàmic (10%)