Alicia Vázquez
Antonio García
David Herrera

APriori - Ejercicio Adult Data Set

Tratamiento de datos

Antes de empezar es conveniente conocer los datos con los que trabajaremos, adecuarlos a nuestras necesidades, agrupar los ítems y/o eliminar parámetros que no aporten valor o no nos interesen. Las agrupaciones las hacemos de tal manera que valores semejantes y poco frecuentes estén juntos, consiguiendo así que el conjunto sea más frecuente.
En el ejercicio, ya se proponen las siguientes transformaciones. Agrupaciones de ítems para los siguientes parámetros:

- **age**, agrupación según la distribución de la variable.
  Young - 16 a 28, Middle-aged - 29 a 37, Senior - 38 a 48,Old - +49
- **hours-per-week,** agrupación por franjas horarias conocidos
  Part-time - 0 a 25, Full-time - 26 a 40, Over-time - 41 a 60, Workaholic - +60
- **capital-gain** y **capital-loss**, conversión de categóricas a tres niveles:
  Low, Medium, Higth

Por nuestra cuenta, hemos agrupado valores que creemos que tienen sentido que estén juntos, ya que si no dispersan la información sin aportar mucha información individualmente.

- **workclass**, agrupación:
  State-gov, *Federal-gov* y *Local-gov* bajo el tipo *Public-sevants*
  Self-emp-not-inc y *Self-emp-inc* bajo el tipo *Self-inc*
- **relationship**, agrupación:
  Husband, Other-relative y *Wife* bajo *Married, Not-in-family* y *Unmarried* bajo *Unmarried*
- **native-country**, hemos decidido agrupar los países por zonas geográficas: Norteamérica, Latinoamérica, Caribe, Europa y Asia.
- **education**, hemos agrupado todas las categóricas que indican una educación previa a la mayoría de edad, es decir, hasta doceavo grado.
- **occupation**, hemos agrupado en high-qualify, medium-qualify y low-qualify

Y hemos eliminado los valores que tan solo aportan duplicación de información, como:

- **marital-status** lo hemos eliminado en favor de *relantionship*

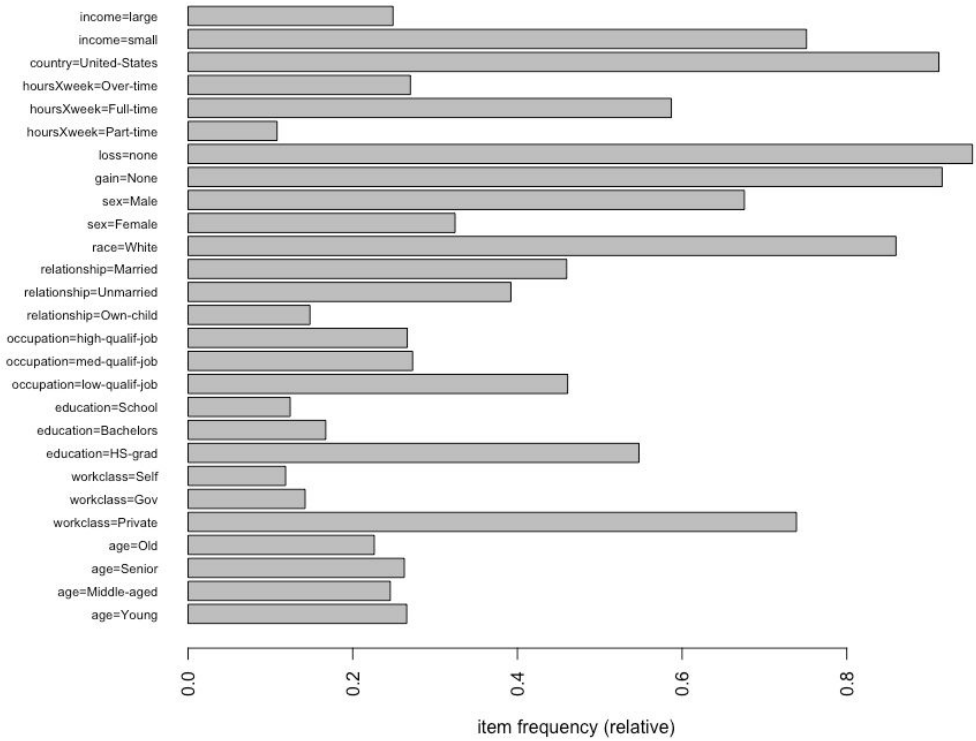
Tratamiento de los NAs

Hay varias técnicas posibles para tratar los NAs. Una dice que se le pueden dar valores como la media pero dejar la fila, para no perder el resto de información. Pero, dado que en este caso los campos que presentan NAs son buenos predictores, como la **workclass**, **native-country** o **income** hemos decidido eliminarlos. Para tratar solamente los casos en los cuales tenemos toda la información.

Alicia Vázquez
Antonio García
David Herrera

Frecuencia de los valores

Tras las transformaciones previamente explicadas, obtenemos el siguiente gráfico de barras para los valores más frecuentes:



Variaciones de *Confidence* y *Support*

Hemos ido haciendo diferentes ejecuciones variando los valores de support y confidence, para valorar los diferentes resultados. A continuación adjuntamos una tabla con los análisis obtenidos para las 10 primeras reglas ordenadas de más Lift a menos.

Support	Confidence	NumReglas	LiftMax	Observacions de las reglas obtenidas con mayor Lift
0.05	0.6	39791	3.08	Población blanca originaria de EEUU, con estudios HS-Grad, sin inversiones son mayoritariamente Young.
0.05	0.8	27240	3.08	Más información sobre el mismo grupo poblacional.
0.05	0.9	17871	2.01	La población con estudios HS, blanca, hombre sin inversiones, proveniente de EEUU, con un sueldo alto, se sabe que están casados.
0.1	0.7	10153	3.01	Seguimos obteniendo info sobre el grupo Young y que les caracteriza: relation=Own-child, workclass=Private, country=EEUU y income=small y sin inversiones
0.15	0.8	3849	1.90	Obtenemos resultados parecidos a la prueba número 3, que habla sobre la población relationship=Married

Alicia Vázquez
Antonio García
David Herrera

0.30	0.7	899	1.30	Conjunto de reglas que da información sobre la población masculina: realtionship=Married, race=White, gain=None y country=EEUU son sex=Male
0.50	0.8	144	1.10	Habla de forma dispar sobre grupos con income=small, race=White and gain=None. Adjuntadas a continuación.

```
lhs                                              rhs    support   confidence lift
[1]{workclass=Private,gain=None,loss=none} =>{income=small}  0.54  0.83  1.11
[2]{workclass=Private,gain=None}           => {income=small} 0.55  0.81  1.08
[3]{gain=None,loss=none}                   => {income=small} 0.70  0.80  1.07
[4]{sex=M,country=US}           => {race=White}  0.56 0.90  1.05
[5]{sex=M,loss=none,country=US} => {race=White}  0.53 0.90  1.05
[6]{sex=M,gain=None,country=US} => {race=White}  0.50 0.90  1.05
[7]{workclass=Priv.,income=small}           => {gain=None} 0.58 0.96  1.04
[8]{workclass=Priv.,loss=none,income=small} => {gain=None} 0.53 0.95  1.04
[9]{income=small}                           => {gain=None} 0.71 0.95  1.04
[10] {country=US,income=small}              => {gain=None} 0.65 0.95  1.04
```

Conclusiones

Aumentando el **Support**, reducimos significativamente el número de reglas obtenidas, sin embargo, el valor aportado por dichas reglas baja (el **Lift** también va disminuyendo). Si aumentamos la confianza, el lift se mantiene pero el número de reglas disminuya.

Después de probar hacer “zoom” en algunos sets de reglas para algunos valores concretos de variables, observamos que las reglas suelen hacer referencia a las transacciones que tienen los ítems más frecuentes, es muy difícil obtener reglas para minorías. A pesar de haber agrupado ítems minoritarios, (pej: Nacionalidad por continente) siguen siendo valores poco frecuentes y no aparecen en las reglas o con lifts muy bajos. Pero si encuentras una reglas que tenga por resultado una minoría tendrá un lift enorme, aunque no tiene porque serle útil al analista.

```
> rules <- apriori(Adult, parameter = list(support = 0.01, confidence = 0.6))
> rulesAsia <- subset(rules, subset = rhs %in% "country=Asia" & lift > 1.2)
> inspect(head(sort(rulesAsia, by = "confidence"), n = 3))
    lhs                            rhs               support confidence     lift
[1] {workclass=Private,race=Asian-Pac-Isl, hoursXweek=Full-time}
                                               => {country=Asia} 0.01 0.65 30.96
[2] {race=Asian-Pac-Islander,sex=Male}         => {country=Asia} 0.01 0.62 29.92
[3] {workclass=Private, race=Asian-Pac-Isl} => {country=Asia} 0.01 0.62 29.77
```

Entendemos pues que el objetivo de APriori es encontrar las reglas más comunes y generales aunque muchas de ellas pueden parecernos obvias, al ser las más recurrentes.