Datos

CLASE 3: Preprocesamiento de Datos

Dr. Edgar Acuna
Departmento de Matematicas

Universidad de Puerto Rico-Mayaguez

website: academic.uprm.edu/eacuna

Por qué preprocesar los datos?

- Los datos en el mundo real son “sucios”:
  - incompletos: falta valores en los atributos, carecen de algunos atributos de interés, o contienen sólo totalizaciones.
  - ruidosos: contienen errores o valores anómalos.
  - inconsistentes: contienen discrepancias en códigos o nombres (Notas: A,AB,B,C,D,F,W).
  - datos duplicados.
- Si no hay calidad en los datos, no hay calidad en los resultados!
  - Las decisiones de calidad deben estar basadas en datos de calidad.
  - Para hacer Data Warehouse se necesita integrar consistentemente datos de calidad.

Ruido y “Outliers”

- El ruido se refiere a la modificación de valores originales .Ejemplos: distorsión de la voz de una persona , “nieve” en la pantalla de televisión
- Los “outliers” son casos con características que difieren bastante de la mayoría de los casos en el conjunto de datos.

Principales Tareas de Preprocesamiento de Datos

- Limpieza de datos
  - Completar valores faltantes, suavizar datos ruidosos, identificar o eliminar “outliers”, y resolver inconsistencias.
- Integración de datos
  - Integración de múltiples bases de datos.
- Transformación de datos
  - Normalización y totalización.
- Reducción de datos
  - Se obtiene una representación más reducida en volumen pero que produce los mismos o similares resultados analíticos.
- Discretización de datos
  - Parte de la reducción de datos pero que es particularmenete mportante para datos numéricos.

Limpieza de Datos

- Tareas de la limpieza de datos:
  - Completar valores faltantes.
  - Identificar “outliers” y suavizar datos ruidosos.
  - Corregir datos inconsistentes.

Preprocesamiento - Datos Faltantes

- Los datos no siempre están disponibles.
  - E.g., muchas filas no tienen registrados valores para muchos atributos, tales como los ingresos del cliente en datos de ventas.
- La falta de valores se puede deber a:
  - mal funcionamiento de equipos.
  - inconsistencia con otros datos registrados y por lo tanto han sido eliminados y considerados faltantes.
  - datos no ingresados debido a equivocaciones o malos entendidos.
  - algunos datos pudieron no considerarse importantes al momento de ingresar datos y fueron omitidos.
- Puede ser necesario estimar los valores faltantes.

Datos faltantes (cont)

- Los valores faltantes son un problema común en análisis estadístico.
- Se ha propuesto muchos métodos para el tratamiento de valores faltantes. Muchos de estos métodos fueron desarrollados para el tratamiento de valores faltantes en encuestas por muestreo.
- Bello (1995), tratamiento de valores faltantes in regression
- Troyanskaya et al (2001), tratamiento de datos faltantes en clasificacion no supervisada.

Datos faltantes (cont)

- Impacto de valores faltantes:
  - 1% datos faltantes – trivial
  - 1-5% - manejable
  - 5-20% - requiere métodos sofisticados
  - 20% o mas- efecto perjudica las interpretaciónes

Conjunto de datos Census

Tambien conocido como Adult.

48842 instancias, contiene variables continuas , ordinales y nominales (entrenamiento=32561, prueba=16281).

Cuando se eliminan las instancias con valores faltantes quedan 45222 (entrenamiento=30162, prueba=15060).

Tamano=3.8MB (entrenamiento), 1.9MB(prueba)

Disponible en: http://archive.ics.uci.edu/ml/

Donantes: Ronny Kohavi y Barry Becker (1996).

Variables en Census

- 1- age: continua.
- 2- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. Nominal
- 3- fnlwgt (final weight) : Continua.
- 4- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. Ordinal.
- 5- education-num: continua.
- 6- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. Nominal
- 7- occupation: Nominal

Variables en Census

- 8-relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. Nominal
- 9-race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. Nominal
- 10-sex: Female[0], Male[1]. Nominal-Binaria
- 11-capital-gain: continua.
- 12-capital-loss: continua.
- 13-hours-per-week: continua.
- 14-native-country: nominal
- 15 Salary: >50K [2], <=50K [1].

Ejemplo: Conjunto de datos census

age	employme	education	edun	marital	…	job	relation	race	gender	hour	country	wealth
					…							
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	…	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_c	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	…	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	…	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	…	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	…	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	…	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	…	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	…	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	…	Adm_cleric	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acd	12	Never_mar	…	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	…	Craft_repai	Husband	Asian	Male	40	*MissingVa	rich
34	Private	7th_8th	4	Married	…	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	…	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	…	Machine_c	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	…	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	…	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	…	Prof_speci	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

Leyendo files de datos en R

El comando basico es read.table(“filename”)

Si los datos estan en el formato csv se usa
> a=read.csv("c://datos1.csv")
>a
Si los datos estan en Excel se usa la librerias xlsx, que tiene el comando read.xlsx(“filename”), La libreria gdata que tiene el comando read.xlx o La libreria XLConnect

Tambien se puede leer datos en otros formatos usando la libreria foreign.

Datos de hasta un millon de registros pueden ser leidos usando la funcion fread de la libreria data.table de R

Para el ejemplo read.table(https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data,header=F,sep=", ",na.strings="? ")

Leyendo los datos en Python

```
import pandas as pd
df = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/adult/adult.data', header=None, sep=',‘,na_values=“ ?”)
df.columns=['v1', 'v2', 'v3', 'v4',v5','v6','v7','v8','v9','v10','v11','v12','v13','v14','class']
df.dropna(how="all", inplace=True) # elimina la vacia vacia al final del archivo

Otra  form:
 import pandas
names=['v1','v2','v3','v4','v5','v6','v7','v8','v9','v10','v11','v12','v13','v14','clase']
data=pandas.read_csv('c://espol/census.csv',names=names)
print(data.shape)
(32562, 15)
data.describe()
```

Datos Census en la libreria dprep

Asumiendo que se ha instalado la libreria dprep en el espacio de trabajo

```
>library(dprep)
>data(census)
>#Viendo la primera fila de los datos
>head(census)
```

Funciones en R para Valores Faltantes

Para detectar las columnas con missing values

colmiss=which(colSums(is.na(census))!=0)

- Para detectar las filas con missing values
  rmiss=which(rowSums(is.na(census))!=0,arr.ind=T)
- Para hallar el porcentaje de filas con missing values

  length(rmiss)*100/dim(census)[1]
- Para hallar el porcentaje de missing values por columna

  per.miss.col=100*colSums(is.na(census) [,colmiss]))/dim(census)[1]
- Para eliminar los missing values

  census.omit=na.omit(census)

  dim(census.omit)

  [1] 30162 15

Explorando el conjunto de datos usando

imagmiss() de la libreria dprep

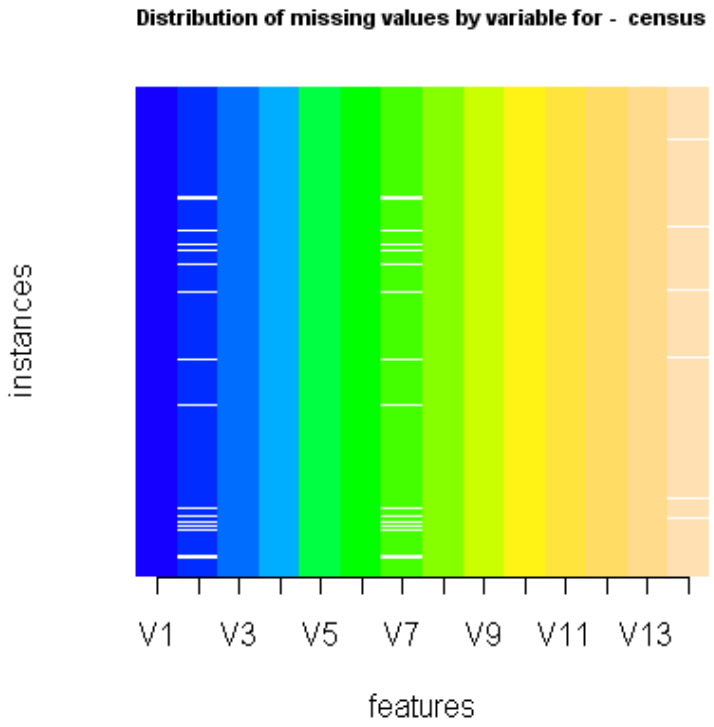
Instalar primero la libreria dprep de R

> imagmiss(data, name="dataname")

Report on missing values for Census :

Number of missing values overall: 4262
Percent of missing values overall: 0.9349485
Features with missing values (percent):
V2 V7 V14
5.638647 5.660146 1.790486

Percent of features with missing values:
21.42857
Number of instances with missing values: 2399
Percent of instances with missing values:
7.36771



La función clean

- Esta función elimina columnas y filas que tienen un gran número de valores faltantes.

```
census.cl=clean(census,tol.col=.5,tol.row=.3,name="cl.census")
```

	Variables	Percent.of.missing
1	V2	5.6386474616873
2	V6	5.66014557292466
3	V13	1.79048555019809

Maximum number of values to be imputed: 4262

Tratamiento de valores faltantes en clasificacion

Eliminacion de casos (CD) – Este método consiste en descartar todas las instancias (casos) con valores perdidos en por lo menos un atributo. Una variante de este método consiste en determinar el grado de valores faltantes en cada instancia y atributo, y eliminar las instancias y/o atributos con altos niveles de valores faltantes. Antes de eliminar cualquier atributo es necesario evaluar su relevancia en el análisis.

Tratamiento de valores faltantes

- Imputacion usando la media (MI) – Reemplazar los valores faltantes de un atributo dado por la media de todos los valores conocidos de ese atributo en la clase a la que la instancia con el valor faltante pertenece. Si la variable es nominal se usa la moda en lugar de la mediana
- Imputacion usando la mediana (MDI). Como la media se ve afectada por la presencia de outliers, parece natural usar la mediana en su lugar para asegurar robustez. En este caso los valores faltantes para un atributo dado es reemplazado por la mediana de todos los valores conocidos de ese atributo en la clase a la que la instancia con el valor faltante pertenece.

  census.mimp=ce.mimp(census ,"mean",1:14, c(2,7,14))

  census.mdimp=ce.mimp(census,"median",atr=1:14,nomatr=c(2,7,14))

Imputación con los k-vecinos mas cercanos (KNNI)

- Dividir el conjunto de datos D en dos partes. Sea *D*m el conjunto que contiene las instancias en las cuales falta por lo menos uno de los valores. Las demás instancias con información completa forman un conjunto llamado *D*c.

  - Para cada vector x en *D*m:

    A) Dividir el vector en dos partes: una la de información observada y otra la de información faltante, *x* = [*x*o; *x*m].

    B) Calcular la distancia entre *x*o y todos los vectores del conjunto *D*c. Usar solo aquellos atributos en los vectores de Dc que están observados en el vector x.

    C) Usar los K vectores más cercanos (K-nearest neighbors) y considerar la moda como un estimado de los valores faltantes para los atributos nominales. Para atributos continuos, reemplazar el valor faltante por la media del atributo en la vecindad de los k vecinos mas cercanos (k-nearest neighborhood).

Imputación con los k-vecinos mas cercanos (KNNI)[2]

A1	A2	A3	A4	Clase
4	5	NA	6	1
5	1	4	1	1
7	9	5	2	1
8	2	5	8	1
6	4	6	2	1

D(r1,r2)= 6.48 D(r1,r3)=6.40, D(r1,r4)=5.38, D(r1,r5)= 4.58
Luego, si se usa k=1 vecinos (r5), NA deberia ser reemplazado por 6,si se usa k=3 vecinos (r3,r4 y r5), NA deberia ser reemplazado por el promedio de 5,5 y 6 que es 5.33

Imputación con los k vecinos mas cercanos KNNI[3]

- El metodo no se podria aplicar si todas las filas de la matriz contienen al menos un valor faltante.
- Usualmente, se toma k igual a 10. Pero el numero de vecinos a usar es a lo sumo igual al numero de filas completas de la matriz.
- En Python, los missing values, son estimados usando predicción con una regresión basada en vecinos mas cercanos.

.

Imputación con los k vecinos mas cercanos (KNNI)[4]

- Cuando la base de datos tiene atributos de distintos tipos, ya no se puede usar las distancias usuales como Euclideana y Manhattan que solo sirven para atributos numericos.
- Hay muchas alternativas para medir la distancia entre los registros (ver Wilson y Martinez).
- Sin embargo la distancia mas usada es la distancia Gower.
- data(crx)
- crx.knn=ec.knnimp(crx,nomatr=c(1,4:7,9:10,12:13),k=5)

  imputacion por KNN no esta implementada en Rattle.

Distancia Gower(disponible en la librería StatMatch)

- Supongamos que tenemos p atributos algunos de los cuales son cuantitativos y otros nominales no ordinales.

La distancia Gower entre los registros Xi y Xj

Xi=(xi1,xi2,……,xip) y Xj=(xj1,xj2,……..,xjp)

Estaa data por $D_G(Xi,XJ)= \sum_{k=1}^{p} d(xik,xjk)/p$

Donde d(xik,xjk)=|xik-xjk|/rango(Xk) si la variable es continua y d(xik,xjk)=1 si la variable Xk es nominal y sus valores son distintos y d(xik,xjk)=0 si sus valores son iguales.

El valor de la distancia Gower esta estandarizado entre 0 y 1.

Insertando aleatoriamente valores faltantes[1]

```
> mat1=cbind(c("a","b","ba","d","c","ab"),c("ac","ad","bf","ba","ac","ba"))
> mat1
     [,1] [,2]
[1,] "a"  "ac"
[2,] "b"  "ad"
[3,] "ba" "bf"
[4,] "d"  "ba"
[5,] "c"  "ac"
[6,] "ab" "ba"
> dim(mat1)
[1] 6 2
```

Insertando aleatoriamente datos faltantes[2]

```
>  mat2=as.vector(mat1)
> mat2
 [1] "a"  "b"  "ba" "d"  "c"  "ab" "ac" "ad" "bf" "ba" "ac" "ba"
>#insertando al azar 6 valores faltantes
>mat2[sample(1:12,6)]=NA
> m2
 [1] "a"  NA   "ba" "d"  NA   NA   "ac" NA   "bf" NA   "ac" NA
> m3=matrix(m2,6,2)
```

Insertando aleatoriamente datos faltantes[3]

```
>mat3
     [,1] [,2]
[1,]  NA "ac"
[2,] NA   NA
[3,] "ba" NA
[4,] "d"  NA
[5,] “c”   "ac"
[6,] NA   “ba”
>#anadiendole una columna ficticia de clases
>mat3=cbind(mat3,rep(1,6))
>#convirtiendo la matriz en dataframe
>mat3=data.frame(mat3)
```

Imputando los datos faltantes

```
>ce.mimp(mat3,"mean",1:2,nomatr=c(1,2))
Summary of imputations using substitution of  mean (mode for nominal features):
     Row Column Class Imput.value
[1,] "1" "1"    "1"   "d"
[2,] "2" "1"    "1"   "d"
[3,] "2" "2"    "1"   "ac"
[4,] "3" "2"    "1"   "ac"
[5,] "4" "2"    "1"   "ac"
[6,] "6" "1"    "1"   "d"
Total number of imputations per class:
1
6
```

Otros métodos de imputación.

- **Hot deck and Cold deck. [nces.ed.gov/statprog].** En Cold deck se usan valores de estudios similares para reemplazar valores perdidos en el estudio actual. En Hot deck se usa valores de atributos correlacionados con el atributo que contiene el valor faltante para sustituirlos.
- **Modelo predictivo:** Regresión Lineal (atributos continuos), Regresión Logística (atributos binarios), logística Polychotomous (atributos nominales). El atributo con valor faltante es usado como la variable de respuesta y los demás atributos son considerados predictoras. En general, se puede usar modelos de Classification and Regression Trees.
- Desventajas: Puede crear sesgo, requiere correlación alta entre predictoras. Cómputo lento.

Otros métodos de imputación.

- Imputación Múltiple. Se imputan varias veces los valores faltantes con valores simulados de una distribucion que se asume para cada variable.
- Metodo de la SVD. Iterativamente, considerando inicialmente a los missings como las medias de cada columna. Y hsta que la norma de la matriz converja.
- Algoritmo EM. Asumir un disribucion para las predictoras y estimando los parametros iterativamente hasta alcanzar convergencia.
- Los árboles de decisión tienen su propio enfoque para tratar valores faltantes.