Mario Becerra 124362

Tarea 3: 3 Feb 2015

① Factorización

Demostrar que $X \not\perp Y \mid Z, W$.

La conjunta se puede factorizar como

$$p(x,y,z,w,a,c) = p(z)\,p(a|z)\,p(y|z)\,p(x|a)\,p(w|a,y)\,p(c|w)$$

La marginal de x,y,z,w es:

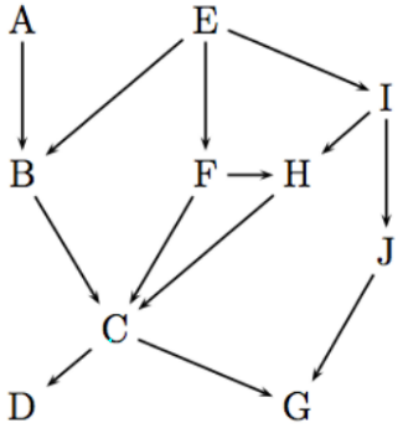
$$p(x,y,z,w) = \sum_a \sum_c p(x,y,z,w,a,c) = p(z)\,p(y|z) \sum_a p(a|z)\,p(x|a)\,p(w|a,y)$$

Para que haya independencia condicional, se debe cumplir que

$$p(x,y,z,w) \propto g(x,z,w)\,h(y,z,w) \quad \ldots (1)$$

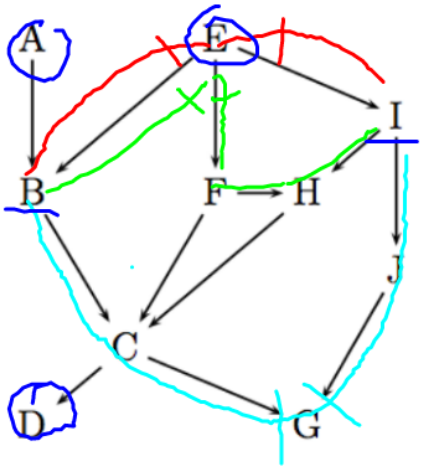
pero en la factorización anterior, se tiene que 'x' y 'y' interactúan a través de 'a', por lo que no existen funciones 'g' y 'h' tales que se cumpla (1).

② D-separación



a) $B \perp H \mid E$ Verdadero

b) $B \perp I \mid A, D, E$ Falso

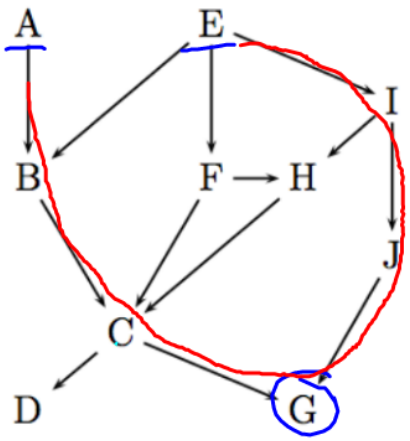


El camino verde tiene un colisionador (H) que no tiene descendientes en $\{A, D, E\}$ ∴ no está activo

El camino azul tiene un colisionador (C) con descendientes en $\{A, D, E\}$ ∴ sí está activo

∴ B e I no están d-separados

c) $A \perp E \mid G$ Falso



El camino rojo tiene un colisionador (G) que está en $\{G\}$.

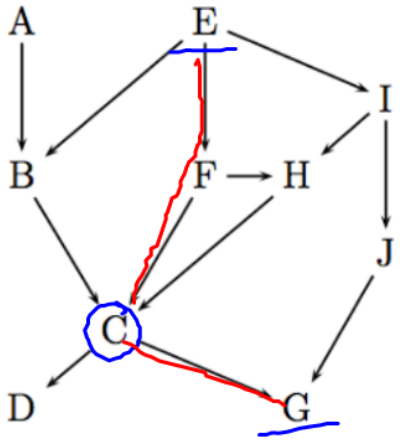
$\therefore$ El camino está activo.

d) $C \perp J$ Falso

El camino $C \to H \to I \to J$ está activo

e) $A \perp I$ Verdadero

f) $E \perp G \mid C$ Falso



El vértice C está en el camino rojo. Por lo tanto, está activo.

Tarea 3

Mario Becerra 000124362

Problema 3. Modelos locales

En estos ejemplos construiremos modelos locales usando tablas de frecuencias. Se utiliza los datos en el archivo admisiones.csv, en donde cada renglón representa a un alumno que solicitó entrar a una universidad, las variables son el departamento al que aplicó, el género del aplicante y su resultado (aceptación o rechazo).

Se pide:

Considera la relación Genero -> Admisión (solo dos variables). Construye un modelo local con la tabla de frecuencias y explica qué relación observas entre estas dos variables.

Considera ahora el modelo Género -> Departamento -> Admisión. Construye los dos modelos locales con las tablas de frecuencias correspondientes. Bajo este modelo, ¿cómo se explica la relación de Género con Admisión que observaste en el ejemplo anterior?

Explica qué significa que no haya una flecha directa de Género a Admisión. Verifica tu respuesta calculando el porcentaje de admitidos para cada sexo dentro de cada departamento, y explica qué tienen qué ver estas tablas con la ausencia de una flecha directa de Género a Admisión.

```
options(digits=2)
setwd('/home/mbc/Dropbox/ITAM_Dropbox/Estadística Multivariada/Tareas_Git/Tarea_03')

library(bnlearn)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)

datos <- read.csv('admisiones.csv')
```

Modelo con la relación Género -> Admisión.

```
white <- data.frame(from = c('Gender'), to = c('Admit'))
net_gen_adm <- hc(data.frame(Gender=datos$Gender, Admit=datos$Admit), whitelist=white)
fit_net_gen_adm <- bn.fit(net_gen_adm,
                          data = data.frame(
                              Gender=datos$Gender,
                              Admit=datos$Admit),
                          method = 'mle')
fit_net_gen_adm
```

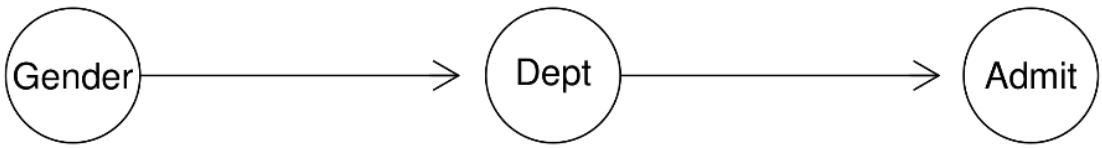
```
##
##   Bayesian network parameters
##
##   Parameters of node Gender (multinomial distribution)
##
## Conditional probability table:
##
## Female   Male
##   0.41   0.59
##
##   Parameters of node Admit (multinomial distribution)
##
## Conditional probability table:
##
##           Gender
## Admit      Female Male
##   Admitted   0.30 0.45
##   Rejected   0.70 0.55
```

La primera tabla indica que hay una proporción menor de mujeres que de hombres en los datos, mientras que en la segunda tabla se ve que la probabilidad estimada de ser admitido dado que es una mujer es menor que la de ser admitido dado que es un hombre. Esto podría llevar a pensar que existe cierta preferencia hacia el género masculino a la hora de hacer la selección de alumnos.

Modelo Género -> Departamento -> Admisión.

```
white <- data.frame(from = c('Gender', 'Dept'), to = c('Dept', 'Admit'))
net_gen_dep_adm <- hc(
                      data.frame(
                          Gender=datos$Gender,
                          Admit=datos$Admit,
                          Dept=datos$Dept),
                      whitelist=white)
graphviz.plot(net_gen_dep_adm, layout='circo')
```

```
## Loading required namespace: Rgraphviz
```




```
fit_net_gen__dep_adm <- bn.fit(net_gen_dep_adm,
                               data = data.frame(
                                  Gender=datos$Gender,
                                  Admit=datos$Admit, Dept=datos$Dept),
                               method = 'mle')
fit_net_gen__dep_adm
```

```
##
##   Bayesian network parameters
##
```

```
##   Parameters of node Gender (multinomial distribution)
##
## Conditional probability table:
##
## Female   Male
##   0.41   0.59
##
##   Parameters of node Admit (multinomial distribution)
##
## Conditional probability table:
##
##           Dept
## Admit          A     B     C     D     E     F
##   Admitted 0.644 0.632 0.351 0.340 0.252 0.064
##   Rejected 0.356 0.368 0.649 0.660 0.748 0.936
##
##   Parameters of node Dept (multinomial distribution)
##
## Conditional probability table:
##
##     Gender
## Dept Female  Male
##    A  0.059 0.307
##    B  0.014 0.208
##    C  0.323 0.121
##    D  0.204 0.155
##    E  0.214 0.071
##    F  0.186 0.139
```

Bajo este modelo, se puede decir que el departamento al que se van las personas depende del género, por ejemplo, el departamento C es el más popular entre las mujeres, mientras que el más popular entre los hombres es el A; y el menos popular entre las mujeres es el B y entre los hombres es el E.

Asimismo, este modelo dice que la proporción de estudiantes admitidos es distinta para cada departamento. El departamento A es el que más admite alumnos y el F es el que menos alumnos admite.

```r
counts <- dplyr::summarise(group_by(datos, Gender, Dept, Admit), count = n())

k=rep(0,nrow(counts))
for(i in seq(to=nrow(counts)/2, 1)){
  k[2*i]=counts$count[2*i] + counts$count[2*i-1]
  k[2*i-1] = k[2*i]
}

counts$prop <- counts$count/k

df <- cbind(counts[counts$Gender=='Male',], counts[counts$Gender=='Female',])

kable(df, format = "markdown")
```

Gender	Dept	Admit	count	prop	Gender	Dept	Admit	count	prop
Male	A	Admitted	512	0.62	Female	A	Admitted	89	0.82
Male	A	Rejected	313	0.38	Female	A	Rejected	19	0.18

Gender	Dept	Admit	count	prop	Gender	Dept	Admit	count	prop
Male	B	Admitted	353	0.63	Female	B	Admitted	17	0.68
Male	B	Rejected	207	0.37	Female	B	Rejected	8	0.32
Male	C	Admitted	120	0.37	Female	C	Admitted	202	0.34
Male	C	Rejected	205	0.63	Female	C	Rejected	391	0.66
Male	D	Admitted	138	0.33	Female	D	Admitted	131	0.35
Male	D	Rejected	279	0.67	Female	D	Rejected	244	0.65
Male	E	Admitted	53	0.28	Female	E	Admitted	94	0.24
Male	E	Rejected	138	0.72	Female	E	Rejected	299	0.76
Male	F	Admitted	22	0.06	Female	F	Admitted	24	0.07
Male	F	Rejected	351	0.94	Female	F	Rejected	317	0.93

En la tabla anterior se puede ver el porcentaje de admitidos para cada sexo dentro de cada departamento, si se ve con detenimiento se puede observar que las proporciones son muy parecidas para ambos géneros; excepto en el departamento 1. Esto tal vez se asimile mejor en la siguiente tabla.

```
df2 <- data.frame(Dept = counts$Dept[counts$Gender=='Male'], Dif=abs(counts$prop[counts$Gender=='Male']
df2<-df2[seq(to=nrow(df2),by=2),]
kable(df2, format = "markdown")
```

	Dept	Dif
1	A	0.20
3	B	0.05
5	C	0.03
7	D	0.02
9	E	0.04
11	F	0.01

Esta tabla muestra la diferencia en valor absoluto de proporciones entre hombres y mujeres por cada departamento. Se puede ver que la diferencia es muy poca (excepto en el departamento A), por lo que se podría decir que el género no afecta directamente la probabilidad de ser admitido o no, esto explica la ausencia de una flecha directa de Género a Admisión.