Redes Bayesianas (1)

Carlos Hurtado L.

Depto. de Ciencias de la Computación, Universidad de Chile

Referencia

Bayesian networks without tears: making Bayesian networks more accessible to the probabilistically unsophisticated. Eugene Charniak

AI Magazine Volume 12 , Issue 4 (Winter 1991) Pages: 50 - 63
Year of Publication: 1991

Clasificación

- Dado un conjunto de registros en una base de datos, donde cada registro pertenece a una clase.
- Construir un modelo que permita predecir la clase de un nuevo registro.

Clasificación: Tipos de Modelos

- Enfoque Discriminante
  - Arboles de Decisión
  - Reglas de Decisión
  - Discriminantes lineales
- Enfoque Generativo
  - Clasificador Bayesiano Naive
  - Redes Bayesianas
- Enfoque de Regresión
  - Redes Neuronales
  - Regresión Logística

Notación de Probabilidades

Los datos de entrenamiento son eventos instancias de las variables aleatorias $(\vec{X}, C)$, donde $\vec{X} = (X_1, \ldots, X_n)$.

Ejemplo: weather.nominal

Ejemplo: weather.nominal

Outlook	Temp.	Humidity	Windy
Sunny	Hot	High	FALSE
Sunny	Hot	High	TRUE
Overcast	Hot	High	FALSE
Rainy	Mild	High	FALSE
Rainy	Cool	Normal	FALSE
Rainy	Cool	Normal	TRUE
Overcast	Cool	Normal	TRUE
Sunny	Mild	High	FALSE
Sunny	Cool	Normal	FALSE
Rainy	Mild	Normal	FALSE
Sunny	Mild	Normal	TRUE
Overcast	Mild	High	TRUE
Overcast	Hot	Normal	FALSE
Rainy	Mild	High	TRUE

Probabilidad Conjunta

Denotamos $P(X_1 = x_1, \ldots, X_n = x_n, C = c)$ al número de datos donde $X_1 = x_1, \ldots, X_n = x_n, C = c$ dividido por el número total de datos.

Cuando las variables se entiendan del contexto denotamos $P(X_1 = x_1, \ldots, X_n = x_n, C = c)$ simplemente como $P(x_1, \ldots, x_n, c)$.

Probabilidad Condicional

En general, para dos variables W, Y, la *probabilidad condicional* $P(W = w \mid Y = y)$ se define como

$$\frac{P(W=w, Y=y)}{P(Y=y)}.$$

Independencia/Dependecia Probabilística

Dos variables W, Y son **independientes** si $P(W = w, Y = y) = P(W = w)P(Y = y)$, para toda instance w, y de las variables.

Es decir tenemos que $P(W = w \mid Y = y) = P(W = w)$ y lo mismo para Y.

Uso de la Probabilidad Conjunta

- Si tenemos la probabilidad conjunta de un conjunto de variables, podemos calcular cualquier probabilidad sobre ellas.
- Esto se hace con dos reglas:
  - Regla de probabilidad condicional (ya la vimos)
  - Marginación

Marginación

Sea $P(X_1 = x_1, \ldots, X_n = x_n, C = c)$ la probabilidad conjunta.

Sea $\vec{S}$ un vector subconjunto de las variables y $\vec{s}$ un vector subconjunto de sus valores. Sea $\vec{Y}$ el vector de variables que no está en $\vec{S}$. Tenemos:

$$P(\vec{S} = \vec{s}) = \Sigma_{\vec{y}} P(\vec{y}, \vec{s})$$

Ejercicio: probabilidades para weather.nominal

- El vector de variables aleatorias es (Outlook,Temp,Humidity,Windy,Play)
- La variable de la clase es Play.
- Exprese las siguientes probabilidades en términos de la probabilidad conjunta:
  - P(Play=yes) P(Play=no)
  - P(Play=yes | Windy=false)

Definición del Problema de Clasificación en Términos de Probabilidades

Dado un dato $\vec{x}$, que clase le asignamos ?

Respuesta: la clase c tal que

$$Pr(c \mid x)$$

es máximo.

Clasificador Basado en Probabilidades

- Aquí el modelo se codifica como probabilidades
- Por ejemplo: basta almacenar la probabilidad conjunta
- Problemas:
  - Requiere mucho espacio almacenar todas las probabilidades.
    - Para n variables binarias podríamos tener una tabla de tamaño 2^n.
    - En la práctica del orden del tamaño del conjunto de entrenamiento.
  - ¿Cómo estimamos estas probabilidades?
    - Contando (para el caso categórico), pero necesitamos datos repetidos para tener soporte estadístico.

Ejercicio

- Cuál es el tamaño de la tabla de probabilidades conjunta para weather.nominal
- Comparar con el tamaño de la tabla de datos.

Enfoque Bayesiano

Teorema de Bayes:

likelihood

prior

posterior

$$P(c \mid x) = \frac{P(x \mid c)Pr(c)}{P(x)}$$

evidencia

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{\sum_{c_i \in C} P(x \mid c_i)P(c_i)}$$

Enfoque Bayesiano (cont.)

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{\sum_{c_i \in C} P(x \mid c_i)P(c_i)}$$

Necesitamos estimar

- Para toda clase $c_i \in C$, $P(c_i)$.
- Para toda clase c, $P(x \mid c)$.

Teorema de Bayes : Ejemplo

	Bowl #1	**Bowl #2**	**Totals**
Chocolate Chip	10	20	30
Plain	30	20	50
Total	40	40	80

Cuál es la probabilidad de que hayamos sacado una galleta
Bowl 1 (A) dado que esta es “plain” (B)?

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} = \frac{0.75 \times 0.5}{0.625} = 0.6$$

Clasificador Bayesiano

$$P(a_1, a_2, \ldots, a_n \mid c_i) \cdot P(c_i)$$ ¿¿¿

Enfoque Bayesiano (cont.)

- Tenemos:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{\Sigma_{c_i \in C} P(x \mid c_i)P(c_i)}$$

- Luego necesitamos almacenar
  - Para toda clase $c_i \in C$, $P(c_i)$.
  - Para toda clase c, $P(x \mid c)$.

Clasificador Bayesiano Naive

- $P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
- Como $P(x)$ es la misma para toda las clases, sólo necesitamos estimar $P(x \mid c)P(c)$ para toda clase c.
- Suponemos independencia condicional.

$$P(x \mid c) = P(x_1 \mid c) \times \ldots \times P(x_n \mid c)$$

Clasificador Bayesiano Naive

- Luego, podemos codificar el clasificador almacenando:
  - Para toda clase c, $P(c)$
  - Para toda clase c y valor x_1 de un atributo X_1, $P(x_1 \mid c)$
- El tamaño del clasificador es lineal en el número de variables y no exponencial.

Variables Categóricas

Estimar probabilidades contando frecuencias sobre datos de entrenamiento.

Variables Numéricas

- Se asume una distribución de probabiliad para variables numéricas $\Rightarrow$ densidad $f(x)$
  - normal

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 2\sigma^2}.$$

  - Estimación de parámetros

- Similar como antes, se clasifica:

$$P(c) \cdot \prod_i f(a_i|c) ¿$$

Ejercicio

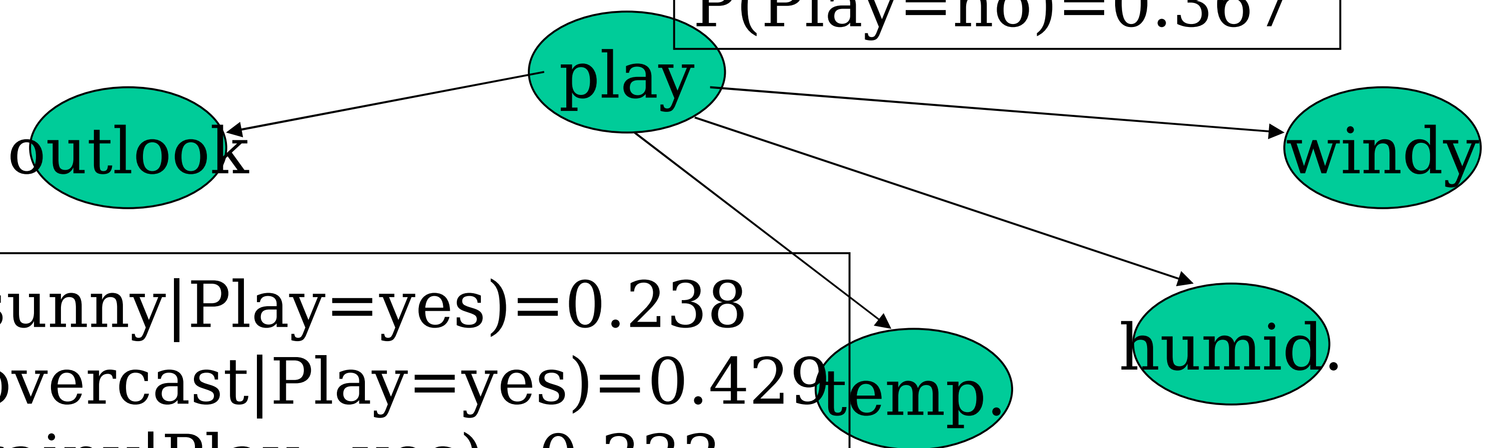
- Construya un clasificador Bayesiano Naive para weather.nominal
  - Diga qué probabilidades debe almacenar el clasificador
  - Calcule las probabilidades para la clase Yes y la variables Outlook
    - Si llega un nuevo registro diga los pasos que se realizan para encontrar su clase: (Sunny,Hot,High,FALSE)

Ejemplo: weather.nominal

Outlook	Temp.	Humidity	Windy
Sunny	Hot	High	FALSE
Sunny	Hot	High	TRUE
Overcast	Hot	High	FALSE
Rainy	Mild	High	FALSE
Rainy	Cool	Normal	FALSE
Rainy	Cool	Normal	TRUE
Overcast	Cool	Normal	TRUE
Sunny	Mild	High	FALSE
Sunny	Cool	Normal	FALSE
Rainy	Mild	Normal	FALSE
Sunny	Mild	Normal	TRUE
Overcast	Mild	High	TRUE
Overcast	Hot	Normal	FALSE
Rainy	Mild	High	TRUE

Clasificador Bayesiano Naive para weather.nominal

P(Play=yes)=0.633
P(Play=no)=0.367



(O=sunny|Play=yes)=0.238
(O=overcast|Play=yes)=0.429
(O=rainy|Play=yes)=0.333
(O=sunny|Play=no)=0.538
(O=overcast|Play=no)=0.077
(O=rainy|Play=no)=0.385

Uso del Clasificador Bayesiano Naive

- Dado un nuevo dato:

(Outlook=sunny,Temp=hot,Humid.=high,Windy=false)

- Para la clase Play=yes, calculamos

A= P(Outlook=sunny,Temp=hot,Humid.=high,Windy=false | Play=yes) x P(Play=yes)

- Por suposición de independencia (Naive) tenemos:

A= P(Outlook=sunny | Play=yes) x P(Temp=hot | Play=yes) x

P(Humid.=high | Play=yes) x P(Windy=false | Play=yes) x P(Play=yes)

Uso del Clasificador Bayesiano Naive (cont.)

- Por suposición de independencia (Naive) tenemos:

A= P(Outlook=sunny | Play=yes) x P(Temp=hot | Play=yes) x

P(Humid.=high | Play=yes) x P(Windy=false | Play=yes) x P(Play=yes)

- Obtenemos estos valores de las tablas del clasificador

A= 0.238 x 0.238 x 0.35 x 0.65 x 0.633 = 0.00815716

Uso del Clasificador Bayesiano Naive (cont.)

- Análogo para Play=no, calculamos:

B= P(Outlook=sunny | Play=no) x P(Temp=hot | Play=no) x

P(Humid.=high | Play=no) x P(Windy=false | Play=no) x P(Play=no)

- Obtenemos estos valores de las tablas del clasificador

B= 0.538 x 0.385 x 0.75 x 0.471 x 0.367 = 0.02685 > A

- Luego para el nuevo dato nuestro pronóstico es Play=no.

Ejemplo

- Clasificador Bayesiano Naive para detectar spam

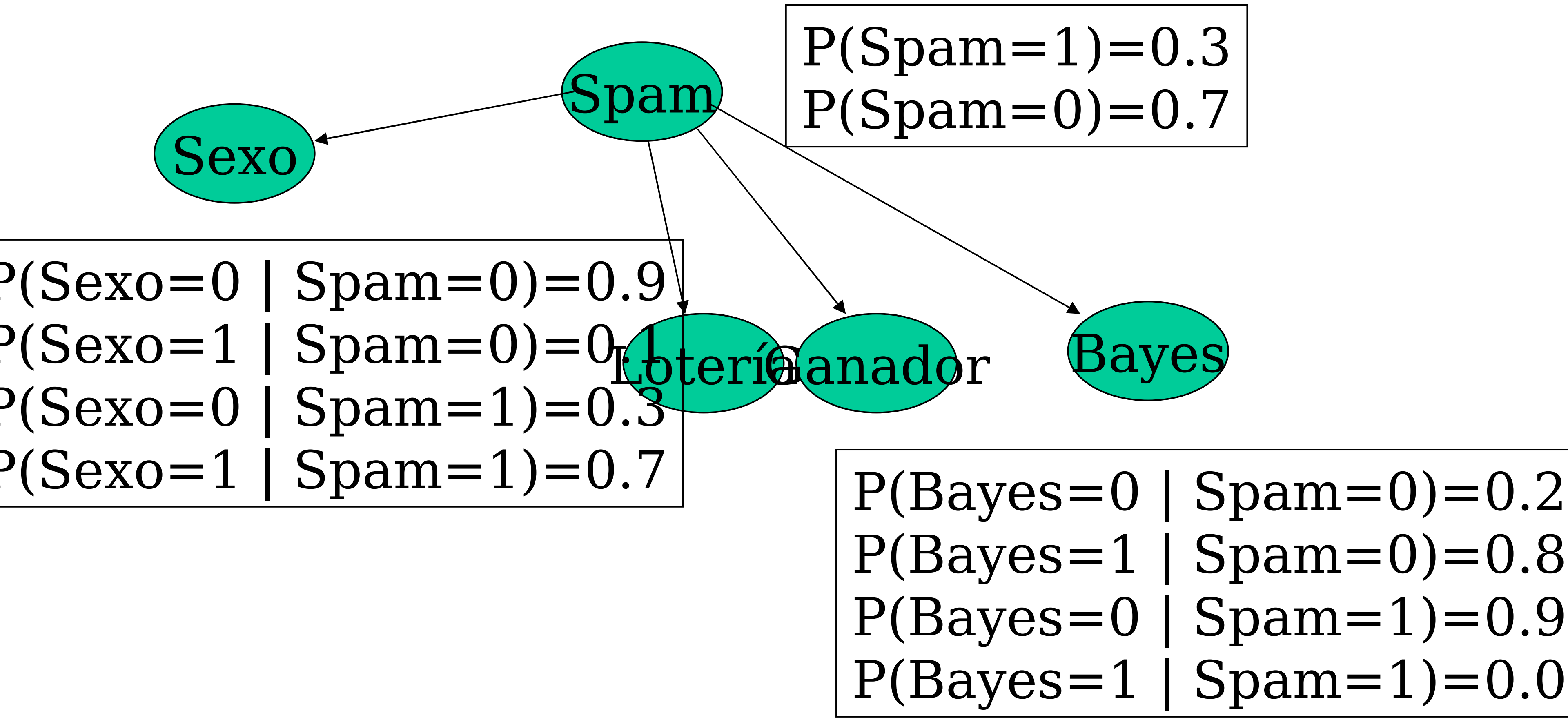
Clasificador Bayesiano Naive para detectar SPAM

- Considere las variables:

(Sexo,Lotería,Ganador,Bayes,Spam)

- La variable de la clase es Spam
- Todas son binarias ({0,1}).

Clasificador Bayesiano Naive para Detectar Spam



Ejercicio

- Suponga que tenemos 2 variables (Sexo,Bayes)
- Diga cómo se clasifica los siguientes datos:
  - (Sexo=0,Bayes=1)
  - (Sexo=1,Bayes=1)
  - (Sexo=0,Bayes=0)

¿Se cumple siempre la suposición de Independencia Condicional de un Clasificador Bayesiano Naive?

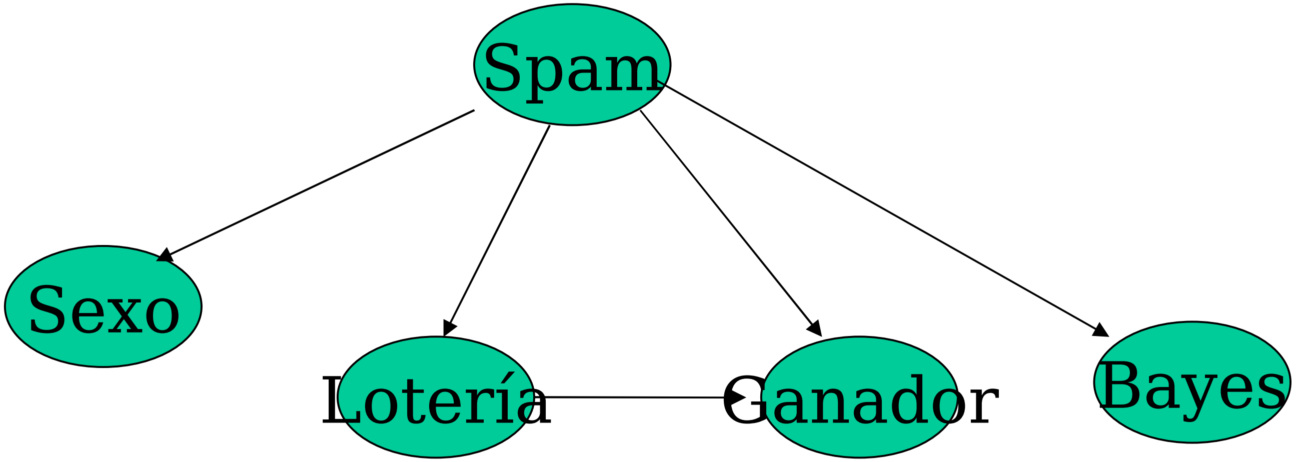
- En muchos caso la suposición Naive no es verdadera:

$$P(x_1,\ldots,x_n \mid c) \neq P(x_1 \mid c) \times \ldots \times P(x_n \mid c)$$

- Ya que pueden existir dependencias dentro de la clase. Por ejemplo:

$$P(x_1,x_2 \mid c) \neq P(x_1 \mid c) \times P(x_2 \mid c)$$

Problema de Clasificador Bayesiano Naive para Detectar Spam



En realidad existe una dependencia entre Lotería y Ganador para la clase Spam=1

Dependencia de Lotería y Ganador en clase Spam=1

- En realidad se cumple:

$$P(\text{Lotería}=1,\text{Ganador}=1 \mid \text{Spam}=1) \;>$$
$$P(\text{Lotería}=1 \mid \text{Spam}=1) \times P(\text{Ganador}=1 \mid \text{Spam}=1)$$

- Por ejemplo, supongamos que en todos los spams “lotería” y “ganador” aparecen juntos:
  - P(Lotería=1 | Spam=1) = .3
  - P(Ganador=1 | Spam=1) = .3
  - P(Lotería=1,Ganador=1 | Spam=1) = .3
- En este ejemplo: 0.3 > 0.3 x 0.3 = 0.09

Dependencia de Lotería y Ganador en clase Spam=1 (cont.)

- La mala supocisión Naive nos puede llevar a un error en la clasificación. Consideremos el dato (Lotería=1,Ganador=1)
- Para clasificarlo en la clase Spam=1, calculamos:

A= P(Lotería=1 | Spam=1) x P(Ganador=1 | Spam=1) x P(Spam=1)

A= 0.3 x 0.3 x 0.3 = 0.027

- Pero en realidad debemos calcular

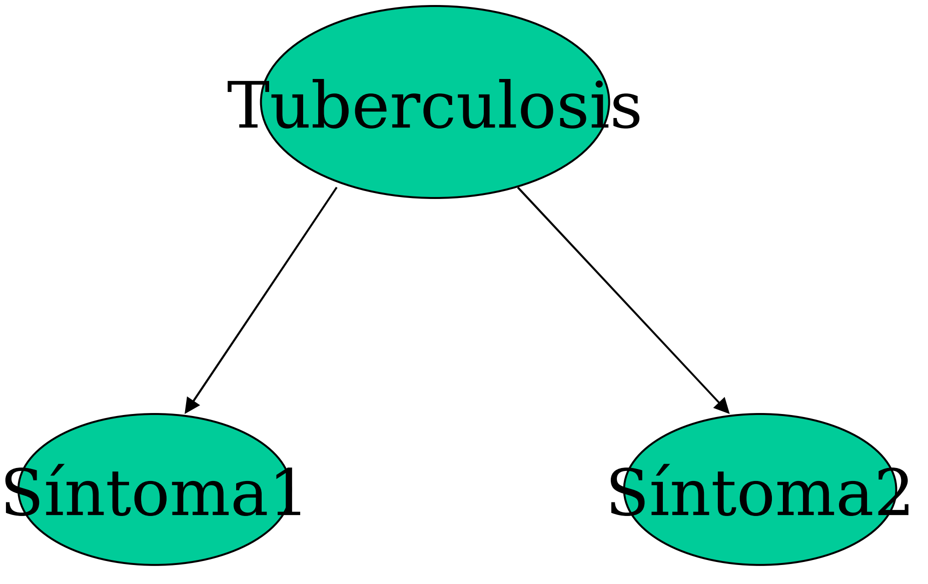
Adep= P(Lotería=1,Ganador=1 | Spam=1) x P(Spam=1)

Adep= 0.3 x 0.3 = 0.09

Dependencia de Lotería y Ganador en clase Spam=1 (cont.)

- Tenemos Adep=0.9 > A=0.027
- Ahora supongamos que para la clase Spam=0 tenemos B=0.04
- En este caso clasificaríamos el mail (Lotería=1,Ganador=1) incorrectamente como Spam=0.

Otro ejemplo: Tuberculosis



¿Hay dependencia entre S1 y S2 dado T?

Otro ejemplo: Tuberculosis

- Depende del comportamiento de los síntomas S1 y S2 cuando hay o no hay Tuberculosis. Debemos preguntarnos:

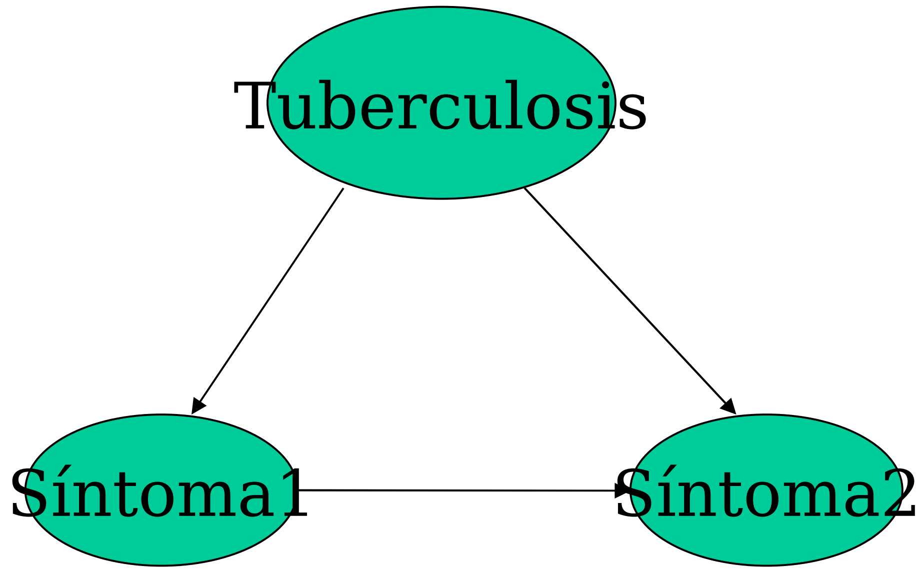
  Para T=no, serían independientes ?

  Para T=si, serían independientes ?

- Recordar que basta que sean dependientes para un caso para que haya dependencia entre las variables

Otro ejemplo: Tuberculosis (cont.)

Si para las personas que tienen tuberculosis,
los síntomas 1 y 2 son excluyentes hay dependencia.



Otro caso, para las personas que tienen tuberculosis,
l síntoma 1 causa el síntoma 2, también tenemos dependenc

Otro ejemplo: Tuberculosis (cont.)

Conclusión:

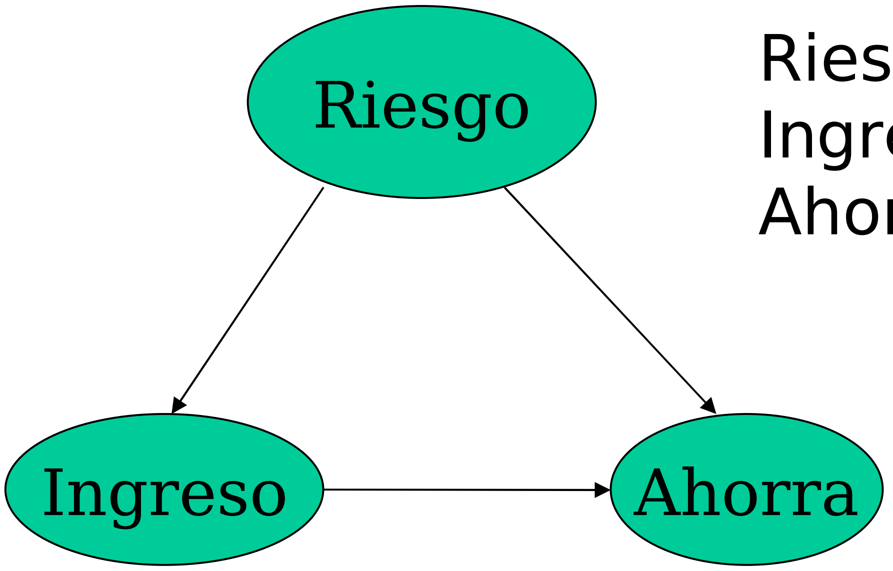
En algunos casos puede ser realista:

$$Pr(\texttt{Sintoma1}, \texttt{Sintoma2} \mid \texttt{Tuberculosis}) \approx$$

$$Pr(\texttt{Sintoma1} \mid \texttt{Tuberculosis}) \times Pr(\texttt{Sintoma2} \mid \texttt{Tuberculosis})$$

En otros casos no.

Otro ejemplo: detección de riesgo financiero



Riesgo {alto,bajo}
Ingreso {alto,bajo}
Ahorra {si,no}

Existe dependencia entre I y A dado R?

Dado que Riesgo es bajo, una persona con ingreso alto tiende a ahorrar

Otro ejemplo: detección de riesgo financiero (cont.)

P(Ahorra=si,Ingreso=alto | Riesgo = bajo)

>

P(Ahorra=si | Riesgo=alto) x P(Ingreso=alto | Riesgo = bajo)

Pero podríamos tener:

P(Ahorra=si,Ingreso=alto | Riesgo = alto)

=

P(Ahorra=1 | Riesgo=alto) x P(Ingreso=alto | Riesgo = bajo)

Redes Bayesianas

- Recordemos que, dada una tupla x a clasificar, necesitamos estimar $P(x \mid c)P(c)$ para toda clase c.
- El problema se reduce a computar

$$P(x \mid c) = \frac{P(x,c)}{P(c)} = \frac{P(x,c)}{\sum_x P(x,c)}$$

Redes Bayesianas (cont.)

- Idea: codificar eficientemente las probabilidades conjuntas $P(x, c_i)$.
  - Identificar qué variables para simplificar probabilidades conjuntas.

Ejemplo: Detección de Fraude

- Variables:
  - `Fraude`(F): si, no
  - `Gasolina`(G): si (compró), no
  - `Joyas`(J): si (compró), no
  - `Edad`(E): < 30, $30 - 50$
  - `Sexo`(S): h, m

Ejemplo: Detección de Fraude (cont.)

- El problema de clasificar se reduce a estimar probabilidades conjuntas:

$$P(F = si | G = si, J = si, E < 30, S = m)$$

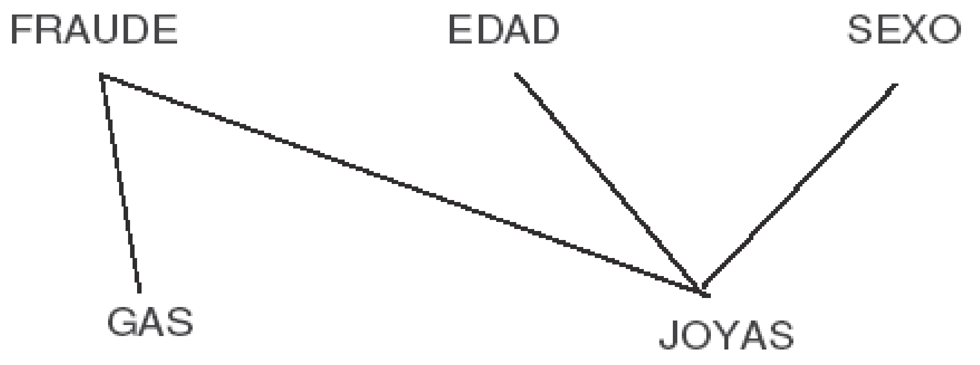
$$= \frac{P(F{=}si,G{=}si,J{=}si,E{<}30,S{=}m)}{P(G{=}si,J{=}si,E{<}30,S{=}m)}$$

$$= \frac{P(F{=}si,G{=}si,J{=}si,E{<}30,S{=}m)}{P(F{=}si,G{=}si,J{=}si,E{<}30,S{=}m)+P(F{=}no,G{=}si,J{=}si,E{<}30,S{=}m)}$$

Red Bayesiana

- Modelo gráfico que codifica probabilides conjuntas. Consiste en:
- Grafo dirigido acíclico (DAG) que contiene un nodo por variable
- Tablas de distribuciones de probabilidades locales

Ejemplo: Detección de Fraude



Fraude: $P(f = si) = 0..00001$

Edad: $P(e < 30) = 0.25$, $P(e = 30 - 50) = 0.4$

Sexo: $P(s = h) = 0.5$

Ejemplo: Detección de Fraude (cont.)

Gas: $P(g = si \mid f = si) = 0.2$, $P(g = si \mid f = no) = 0.01$

Joyas:

$$P(j = si \mid f = si, e = *, s = *) = 0.05$$

$$P(j = si \mid f = no, e =< 30, s = h) = 0..0001$$

$$P(j = si \mid f = no, e = 30 - 50, s = h) = 0.0004$$

$$P(j = si \mid f = no, e => 50, s = h) = 0.0002$$

$$P(j = si \mid f = no, e =< 30, s = m) = 0..0005$$

$$P(j = si \mid f = no, e = 30 - 50, s = m) = 0.002$$

$$P(j = si \mid f = no, e => 50, s = m) = 0.001$$

Regla de la Cadena

Sea $x = \{x_1, \ldots, x_n\}$

$$P(x) = \Pi_{i=1..n} P(x_i | x_1, \ldots, x_{i-1})$$

Ejemplo: $n = 4$

$$P(x_1, x_2, x_3, x_4) =$$

$$P(x_1) \times P(x_2|x_1) \times P(x_3|x_1, x_2) \times P(x_4|x_1, x_2, x_3)$$

Codificación de la Prob. Conjunta usando Regla de la Cadena

Por cada orden de las variables podemos reformular $P(x)$ usando la regla de la cadena. Tenemos $n!$ reformulaciones distintas.

Ejemplo: usando el orden: F, E, S, G, J se tiene:

$$P(f, e, s, g, j) =$$

$$P(f) \times P(e|f) \times P(s|f, e) \times P(g|f, e, s) \times P(j|f, e, s, g)$$

Simplificación de Probabilidades Condicionales

$$P(f, e, s, g, j) =$$

$$P(f) \times P(e|f) \times P(s|f, e) \times P(g|f, e, s) \times P(j|f, e, s, g)$$

La independencia de algunas variables nos lleva a:

$$P(e|f) = P(e)$$

$$P(s|f, e) = P(s)$$

$$P(g|f, e, s) = P(g|f)$$

$$P(j|f, e, s, g) = P(j|f, e, s)$$

Estas igualdades se codifican en la red Bayesiana.

Independencia Condicional en Redes Bayesianas

- Ejemplo: $P(g|f,e,s) = P(g|f)$, esto se puede leer como:
  - G es independiente de E, S dado F.

O mejor: G es independiente de E y S dada la evidencia F.

Dada una probabilidad conjunta: ¿cómo se codifica en una Red Bayesiana?

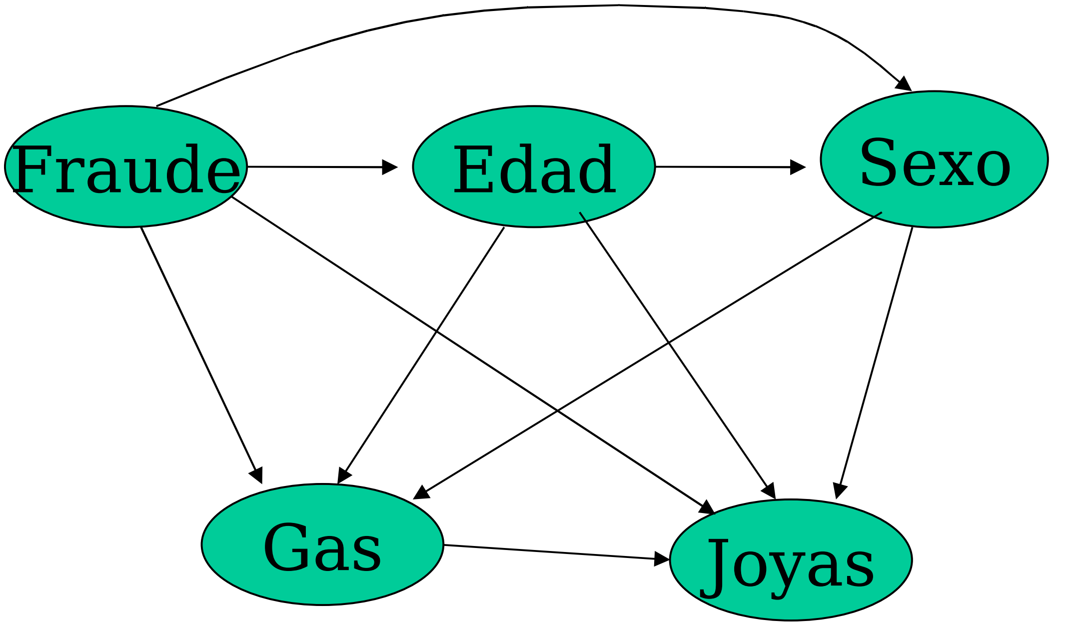
- Ordenamos las variables, ej., F, E, S, G, J.
- Construimos un grafo con un nodopor variable, los padres de X_i son las vaiables $X_1, \ldots, X_{i-1}$.
  - Ej., los padres de S son F, E.
- Eliminamos los arcos (X_j, X_i) tales que

$$P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|(X_1, \ldots, X_{i-1}) - X_j)$$

Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

Red sin suposiciones de independencia

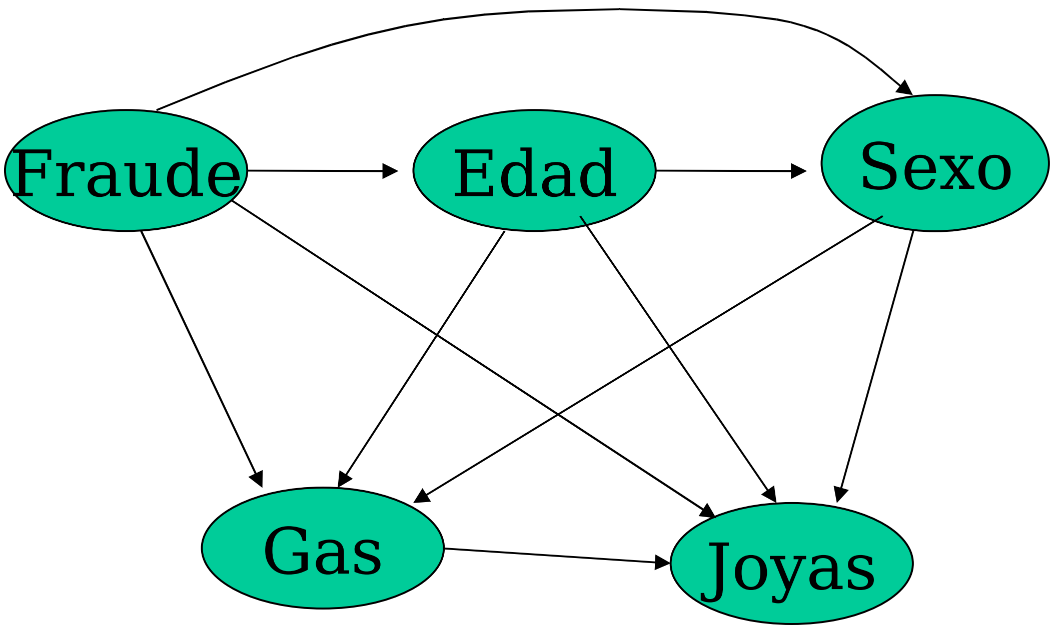


P(f,e,s,g,j) = P(f) P(e|f) P(s|f,e) P(g|f,e,s) P(j|f,e,s,g)

Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

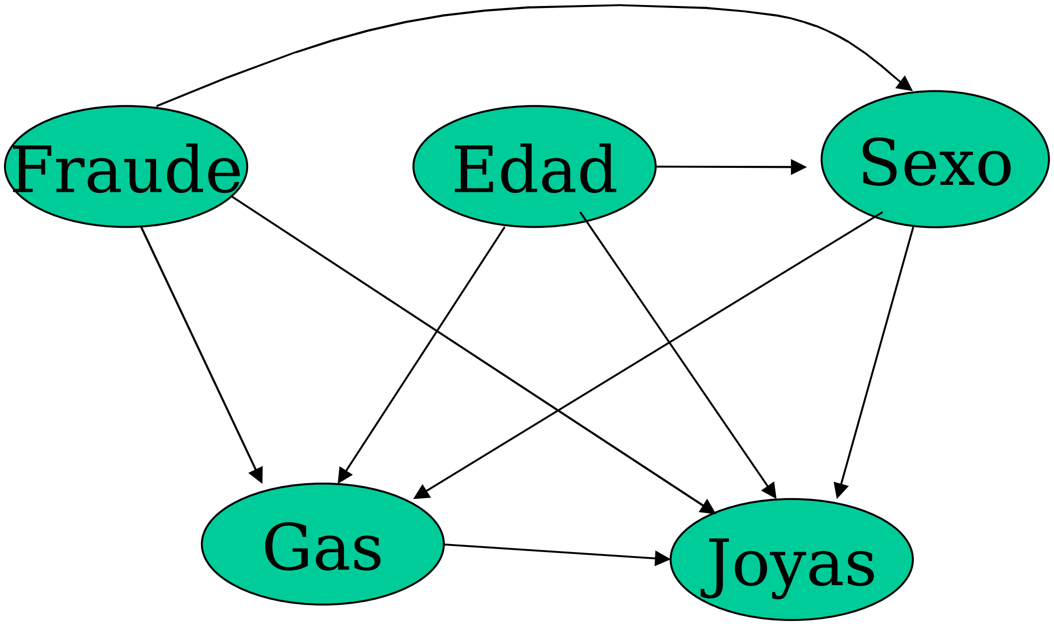
Eliminamos arco debido a P(E | F) = P(E)



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

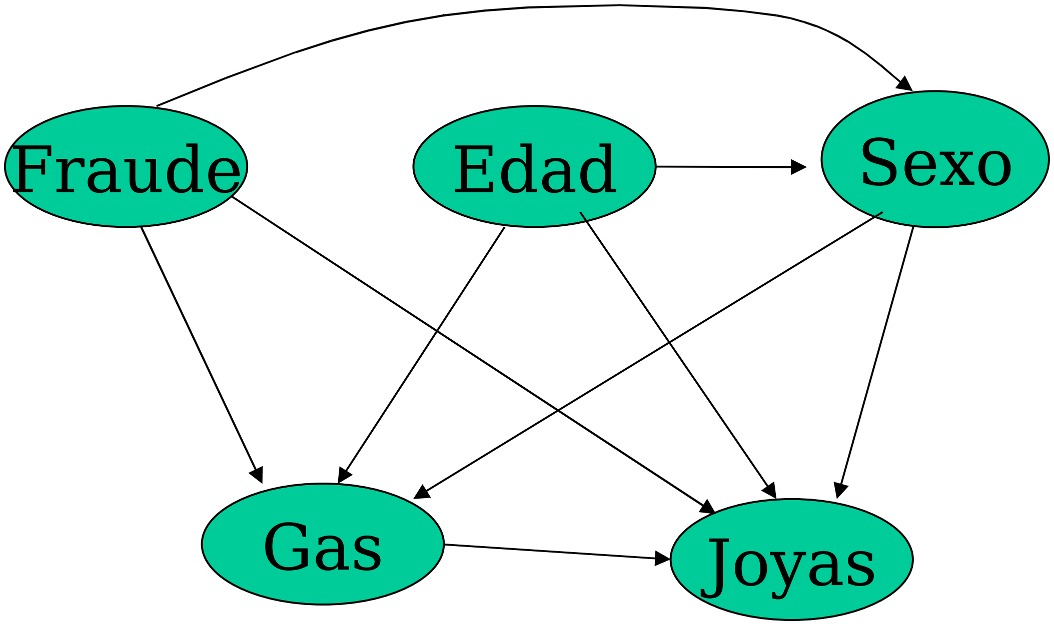
Eliminamos arco debido a P(E | F) = P(E)



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

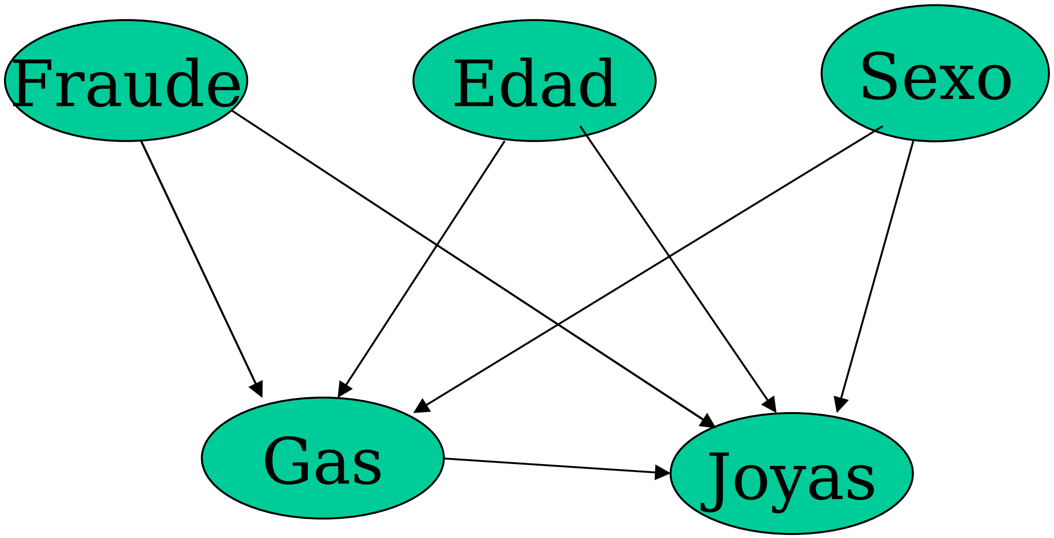
Eliminamos arco debido a P(S | E,F) = P(S)



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

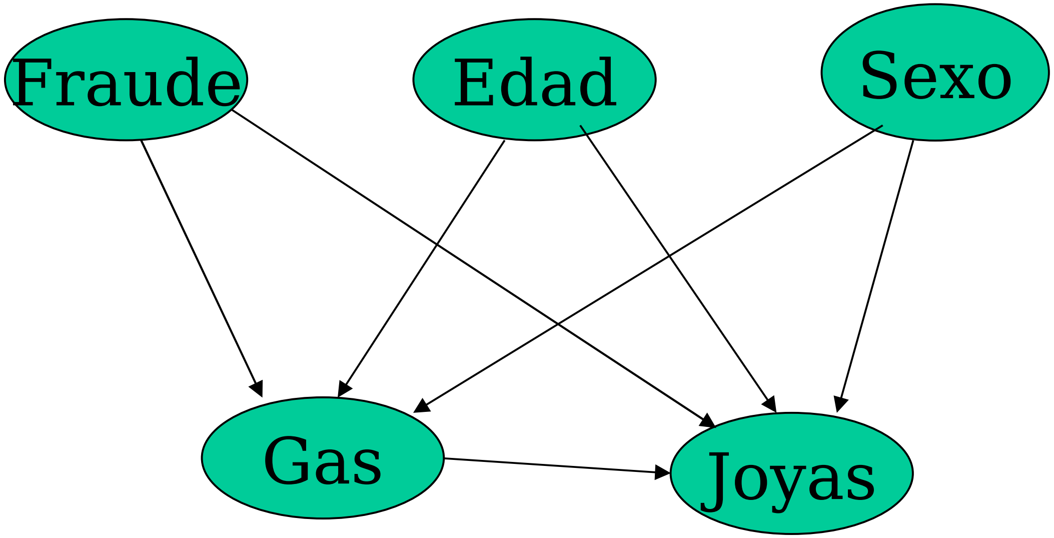
Eliminamos arco debido a P(S | E,F) = P(S)



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

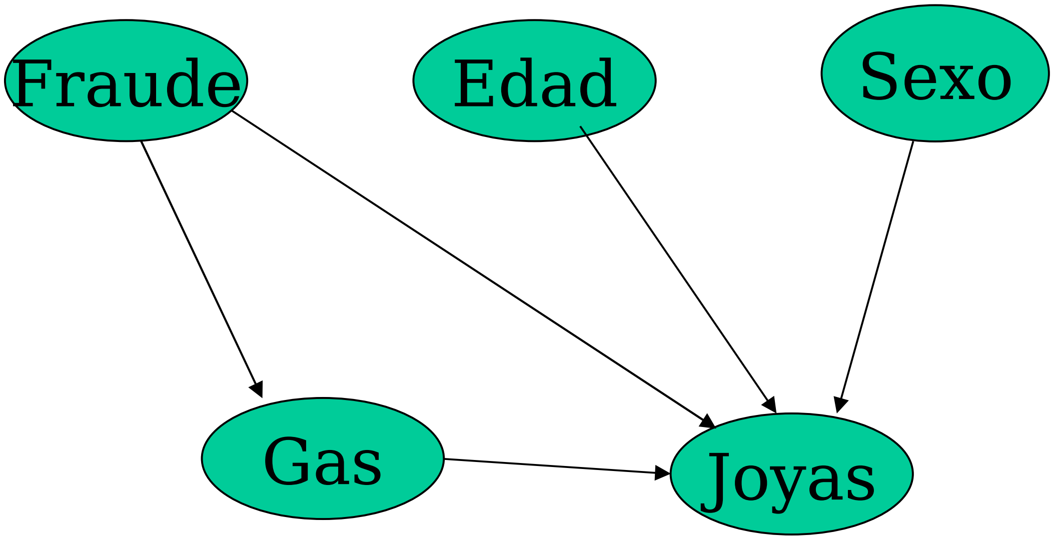
Eliminamos arco debido a P(G | S,E,F) = P(G | F)



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

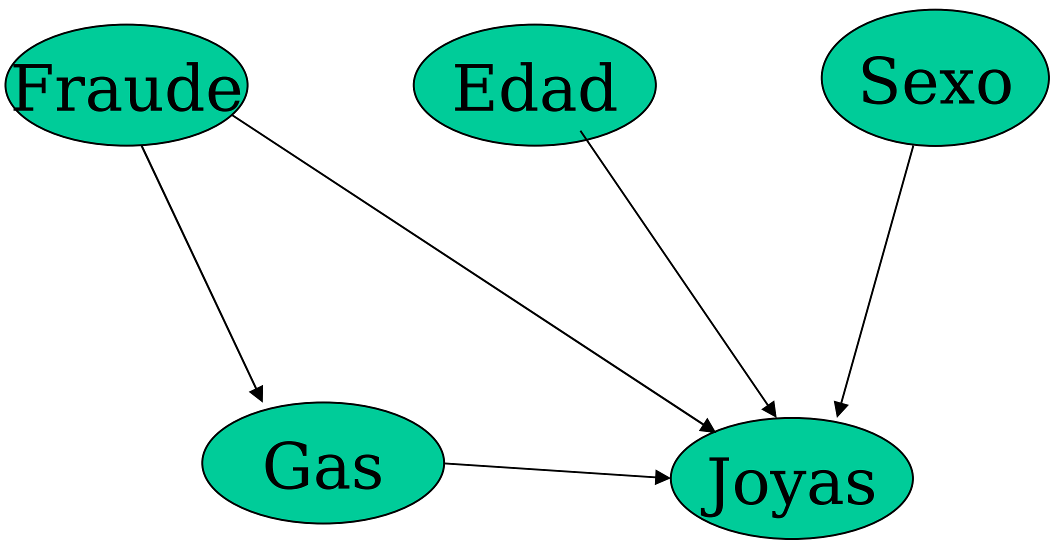
Eliminamos arco debido a $P(G \mid S,E,F) = P(G \mid F)$



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

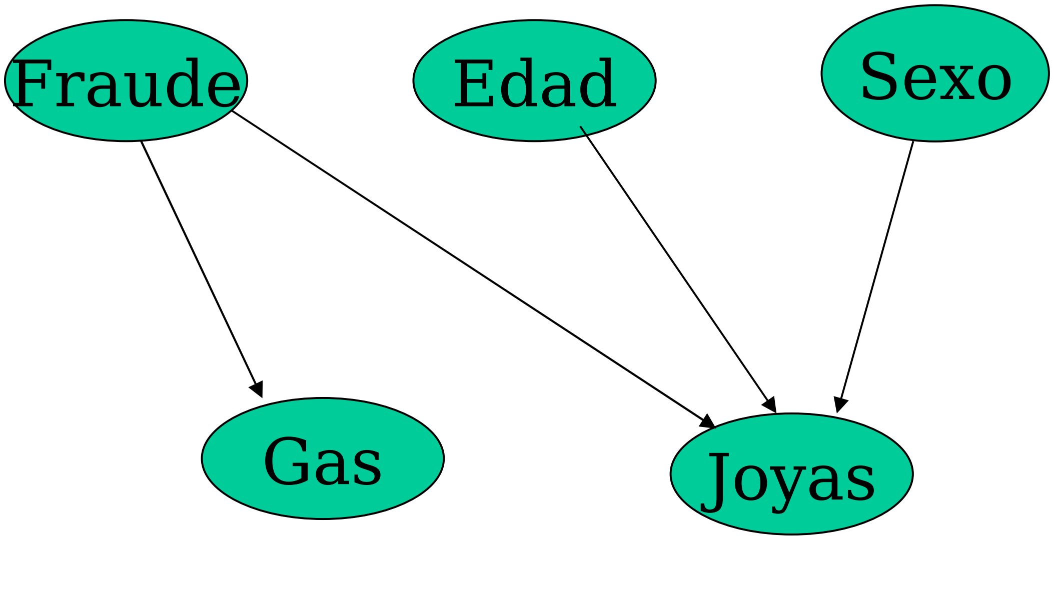
Eliminamos arco debido a P(J | G,S,E,F) = P(J | S,E,



Ejemplo: Detección de Fraude

Orden: F, E, S, G, J

Eliminamos arco debido a P(J | G,S,E,F) = P(J | S,E,



Ejemplo Spam

- Probabilidad conjunta:

P(Spam,Sexo,Loteria,Ganador,Bayes)

- Regla de la Cadena:

P(Spam,Sexo,Loteria,Ganador,Bayes)

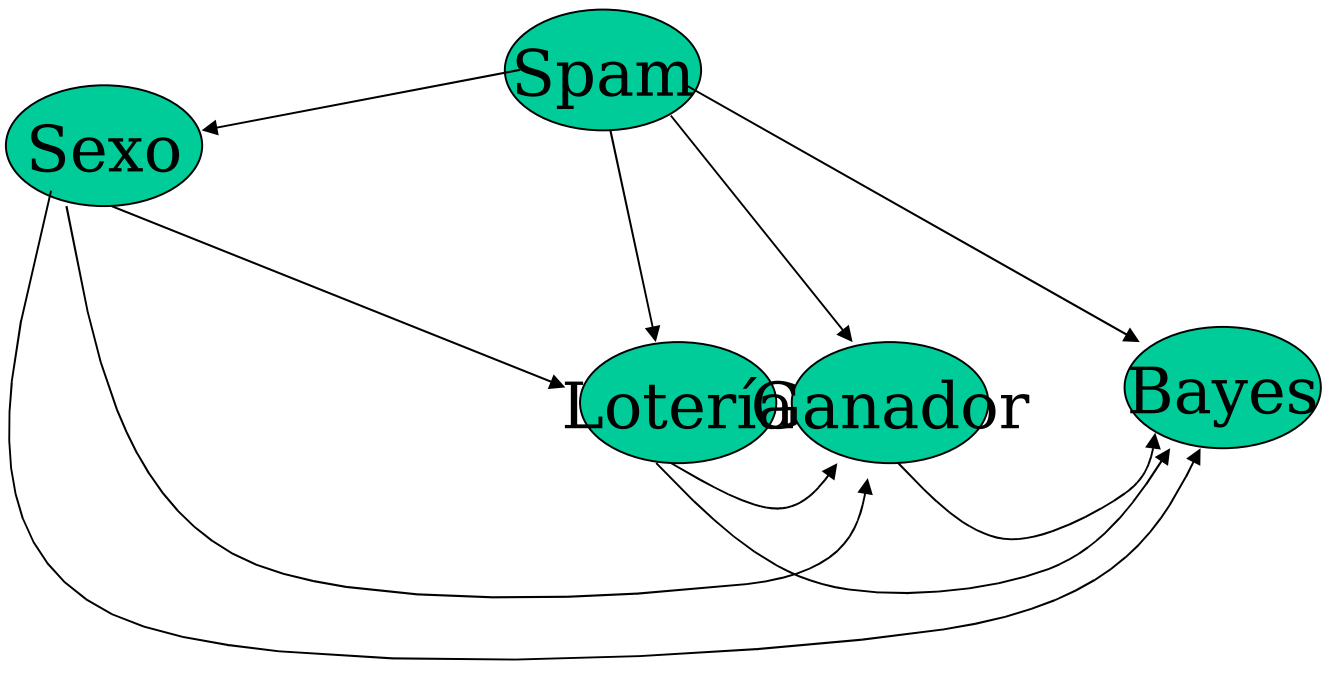
=

P(Spam) P(Sexo|Spam) P(Loteria|Spam,Sexo) P(Ganador|Spam,Sexo,Loteria) P(Bayes|Spam,Sexo,Loteria,Ganador)

Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

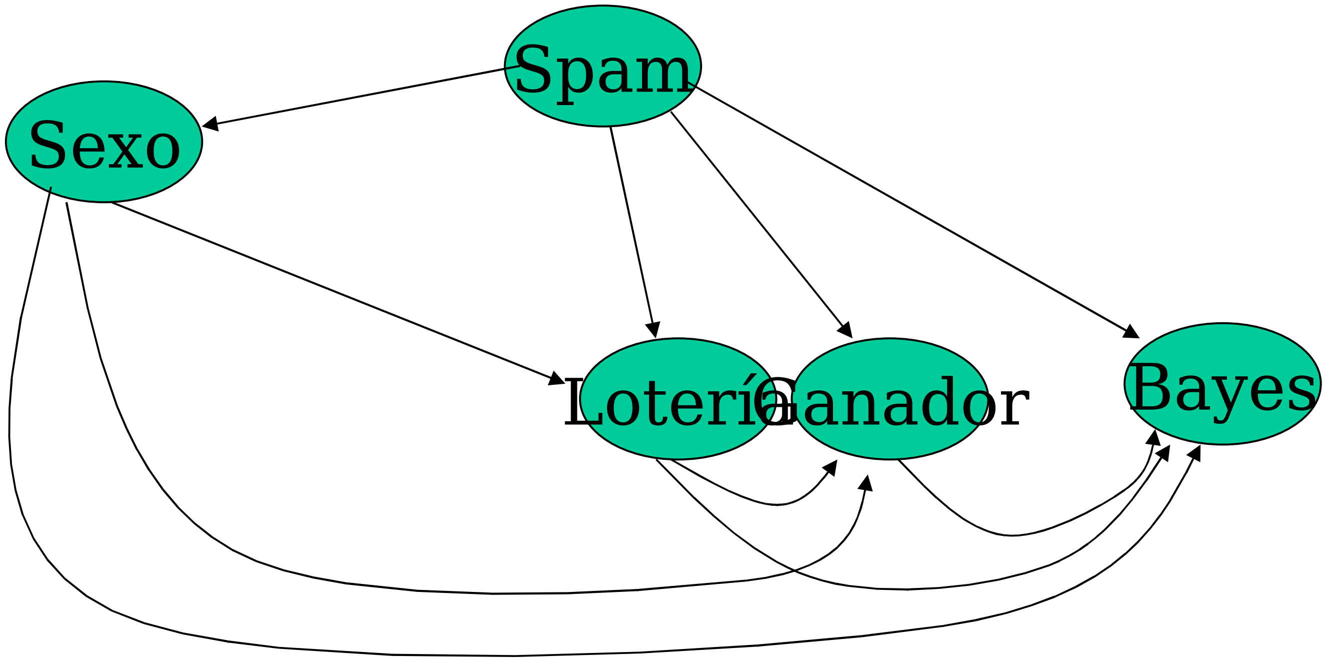
Tenemos una red inicial basada en regla de la cadena



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

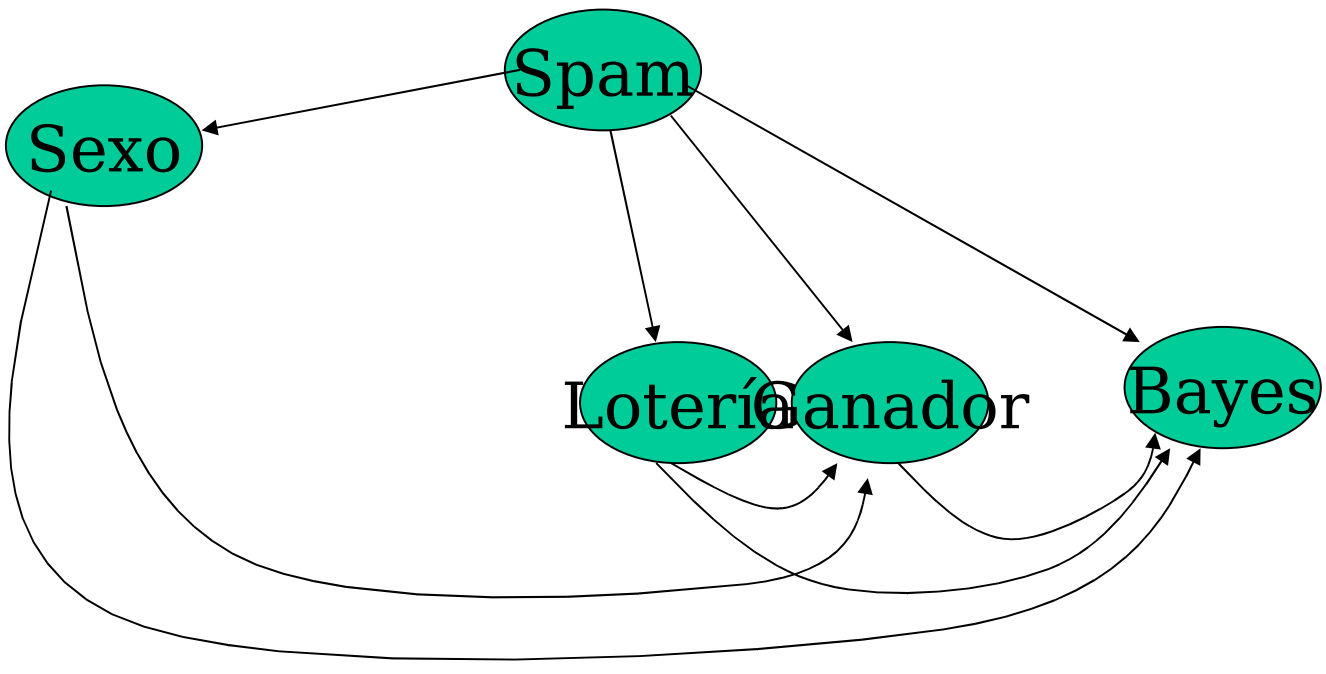
Eliminamos arcos por P(Loteria | Spam,Sexo) = P(Loteria | Spam)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

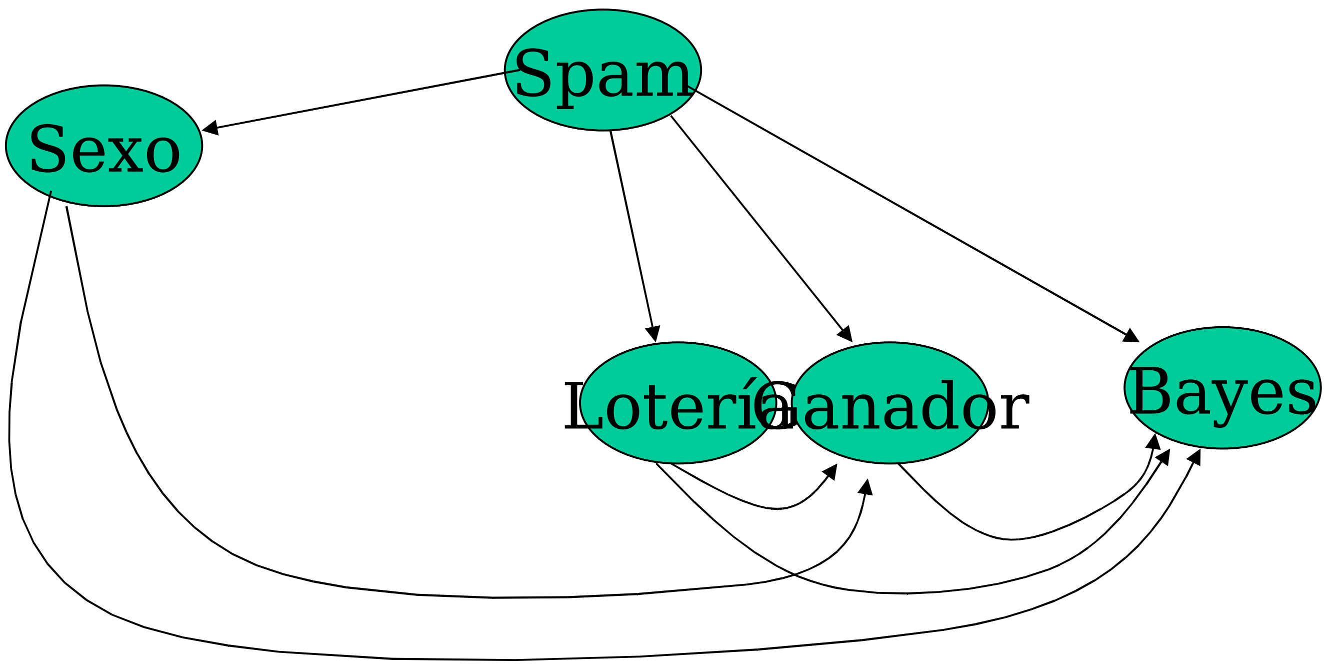
Eliminamos arco por P(Loteria | Spam,Sexo) = P(Loteria | Spam)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

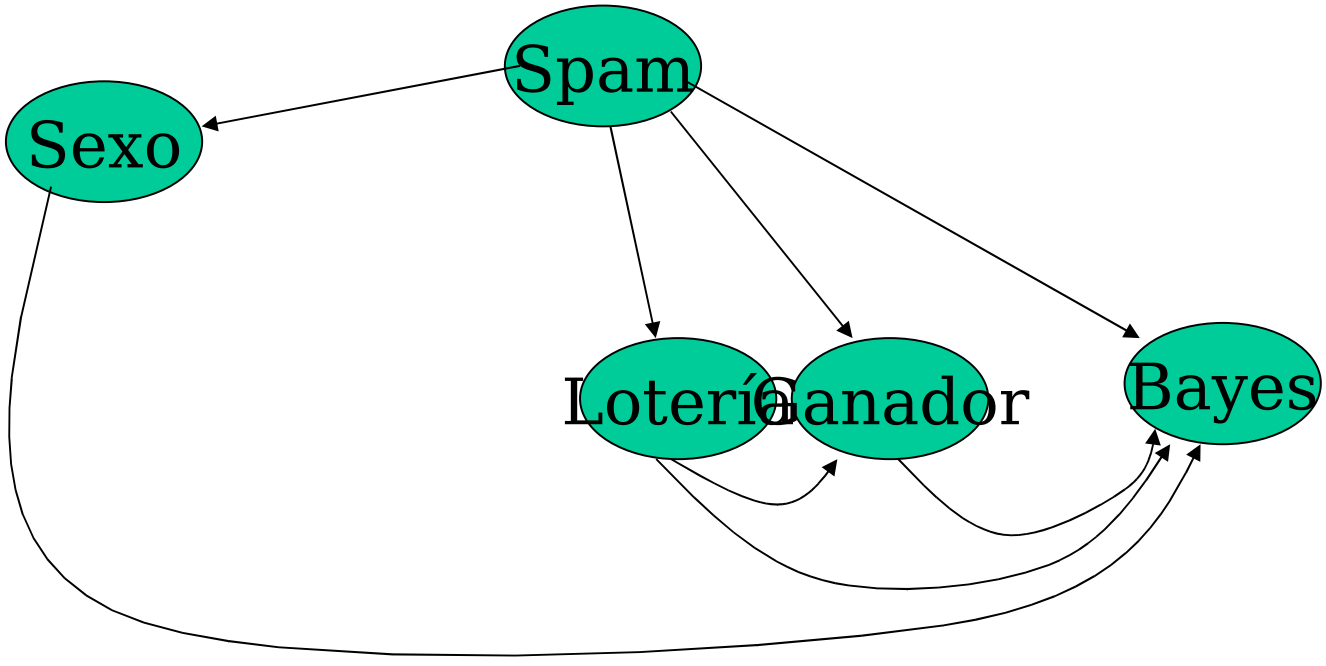
Eliminamos arco por
P(Ganador | Spam,Sexo,Loteria) = P(Ganador | Spam,Loteria)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

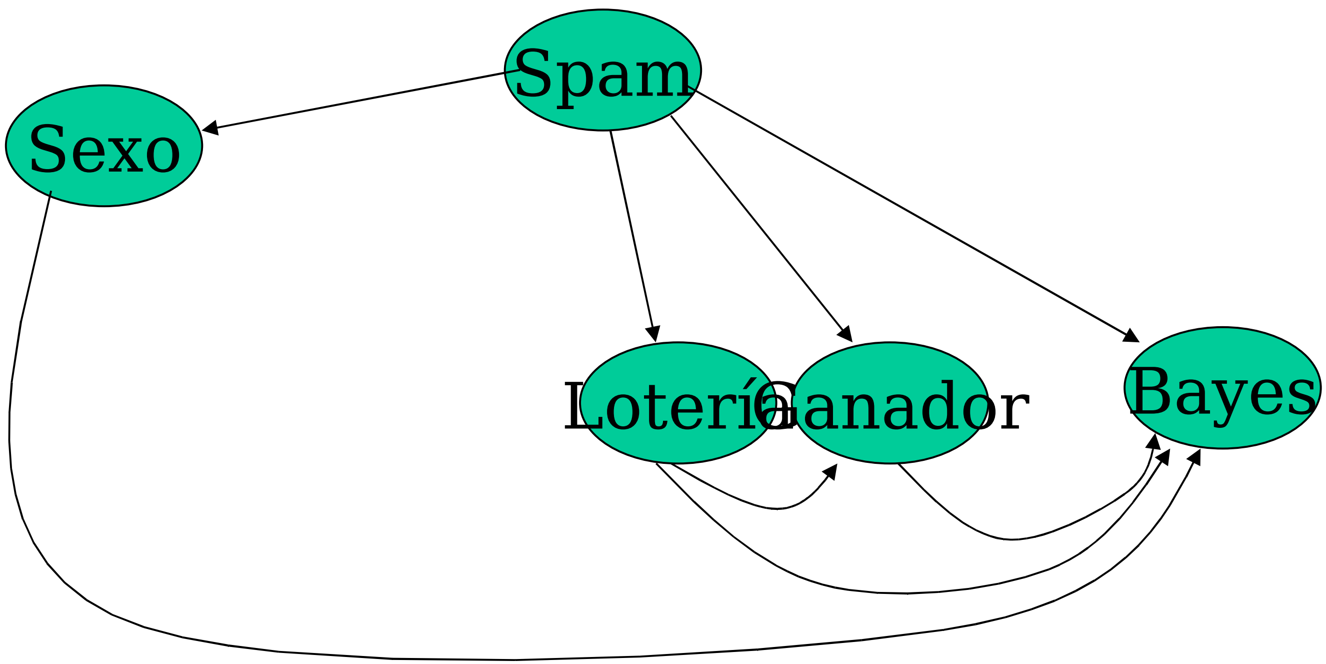
Eliminamos arco por
P(Ganador | Spam,Sexo,Loteria) = P(Ganador | Spam,Loteria)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

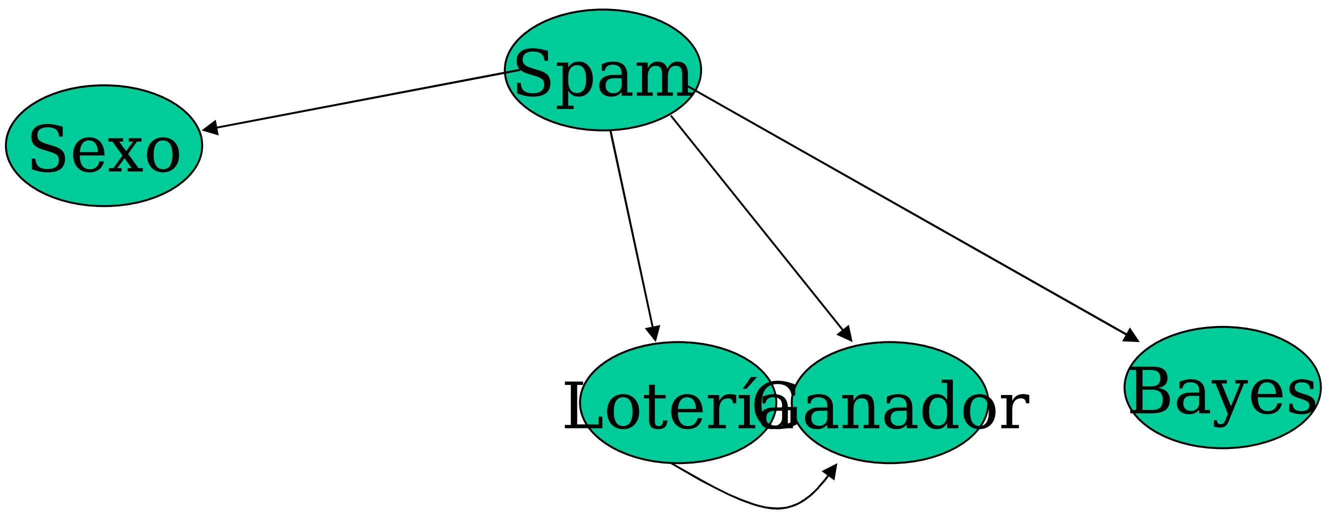
Eliminamos arcos por
P(Bayes | Spam,Sexo,Loteria, Ganador) = P(Bayes | Spam)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

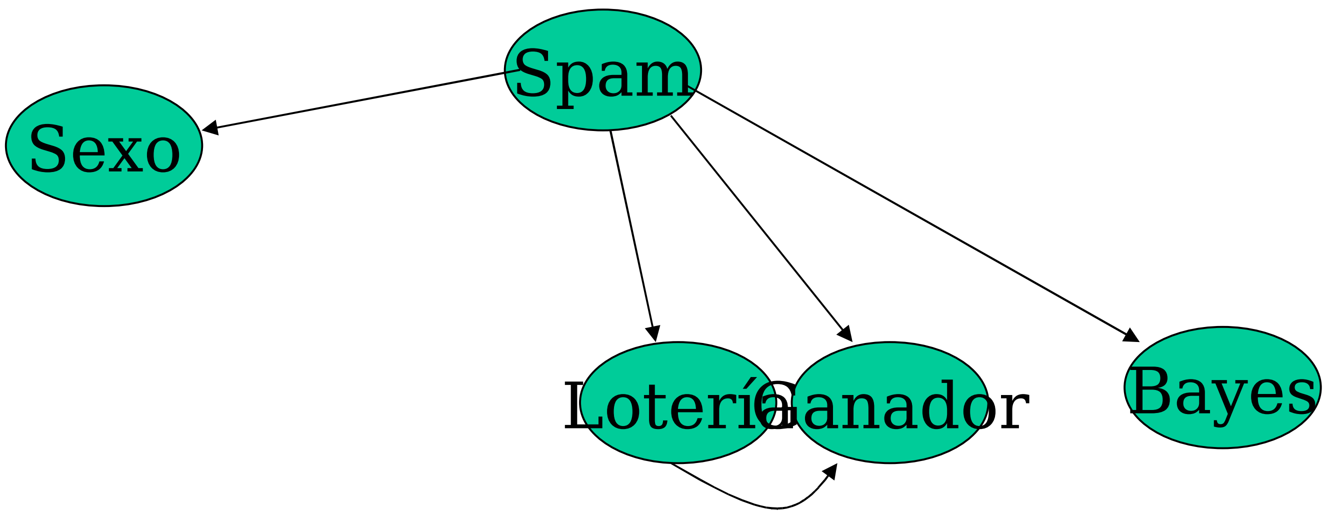
Eliminamos arcos por
P(Bayes | Spam,Sexo,Loteria, Ganador) = P(Bayes | Spam)



Ejemplo: Spam

Orden: Spam, Sexo, Loteria, Ganador, Bayes

Red final:



Ejemplo: Spam

- Tamaño de tabla de probabilidad conjunta: 2^5=32
- Tamaño de tablas de probabilidades locales de la red final: 2+2+4+2+2=12
- Luego podemos decir que la red codifica la prob. conjunta en forma eficiente.

Problema

Si elegimos un orden inadecuado de variables la red resultante no reflejará muchas independencias entre variables

Ejemplo: J,G,S,E,F

$$P(g|j)$$

$$P(s|j,g)$$

$$P(e|j,g,s)$$

$$P(f|j,g,s,e)$$

Ninguna se puede simplificar.