IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

Cierre de la clase 8 – Dependencias entre variables

Dependencias:

¿Cuál es la diferencia entre asociación y dependencia?

Redes Bayesianas:

¿Qué representan las relaciones en una red Bayesiana?

¿Qué representa la red Bayesiana?

¿Qué es un razonamiento basado en evidencia?

Cierre de la clase 8 – Actividad formativa

¿Qué representan estas relaciones?

$(A \perp F, E, G, D)$
$(B \perp F, E, D, G, H \mid C, A)$
$(C \perp F, E, D, G, H \mid A)$
$(D \perp F, C, A, B, H \mid G)$
$(E \perp F, C, A, B, H \mid G, D)$
$(F \perp C, E, D, G, A, B)$
$(G \perp F, C, A, B, H)$
$(H \perp C, E, D, G, B \mid F, A)$

- Causalidad -

Dependencia no es lo mismo que causalidad

La escala de la causalidad

Rung	Action	Question
Association (1)	*Observing*	How does observing X change my belief in Y?
Intervention (2)	*Doing*	What will happen to Y if I do X?
Counterfactual (3)	*Imagining*	If I had done X, what would Y be?

¿Qué entendemos por independencia entre variables?

$$P(Y) = P(Y|X)$$

$$P(X) = P(X|Y)$$

La probabilidad marginal de Y es igual a la probabilidad condicional de Y|X. Análogamente, la probabilidad marginal de X es igual a la probabilidad condicional de X|Y.

Una consecuencia importante de independencia es que su distribución conjunta es el producto de las marginales:

$$P(X, Y) = P(X)P(Y)$$

Usaremos como notación de independencia:

$$X \perp\!\!\!\perp Y$$

¿Qué entendemos por independencia condicional entre variables?

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

X e Y son condicionalmente independientes en relación con Z si su distribución conjunta, dado Z, es el producto de las condicionales X|Z e Y|Z.

La notación será:

$$X \perp\!\!\!\perp \mathrm{Y}|\mathrm{Z}$$

Distinguiremos además entre independencia en el modelo (grafo) y en la distribución conjunta:

$$X \perp\!\!\!\perp_{\mathrm{P}} Y \qquad\qquad X \perp\!\!\!\perp_{\mathrm{G}} Y$$

Independencia en G

Dos variables X e Y serán independientes en G si no existe un camino que las conecte directa o indirectamente.

Dos variables X e Y serán condicionalmente independientes en G dado Z cuando Z **bloquee** todos los caminos entre X e Y.

Condiciones para inferencia causal

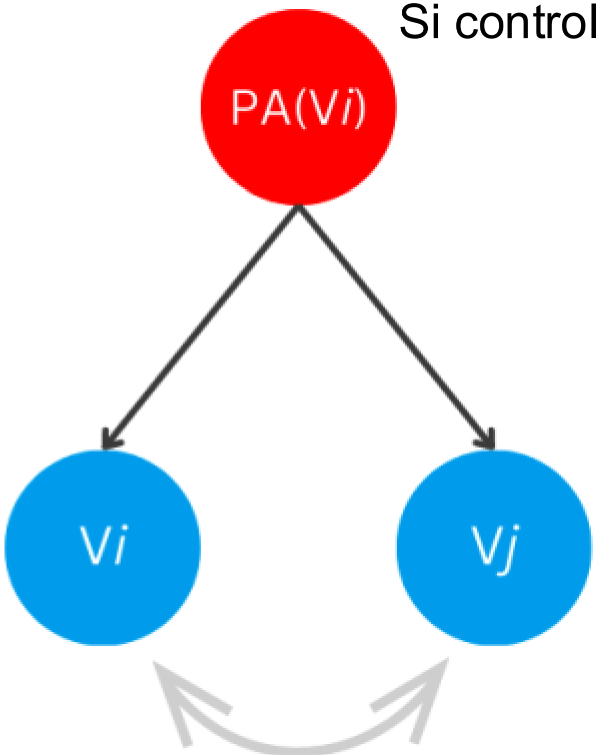
Necesitamos asegurar que podemos mapear las independencias condicionales de G a independencias estadísticas (p). Para esto se necesita la condición causal de Markov, la cual indica que toda variable debe ser independiente de sus no descendientes si conozco a sus padres:

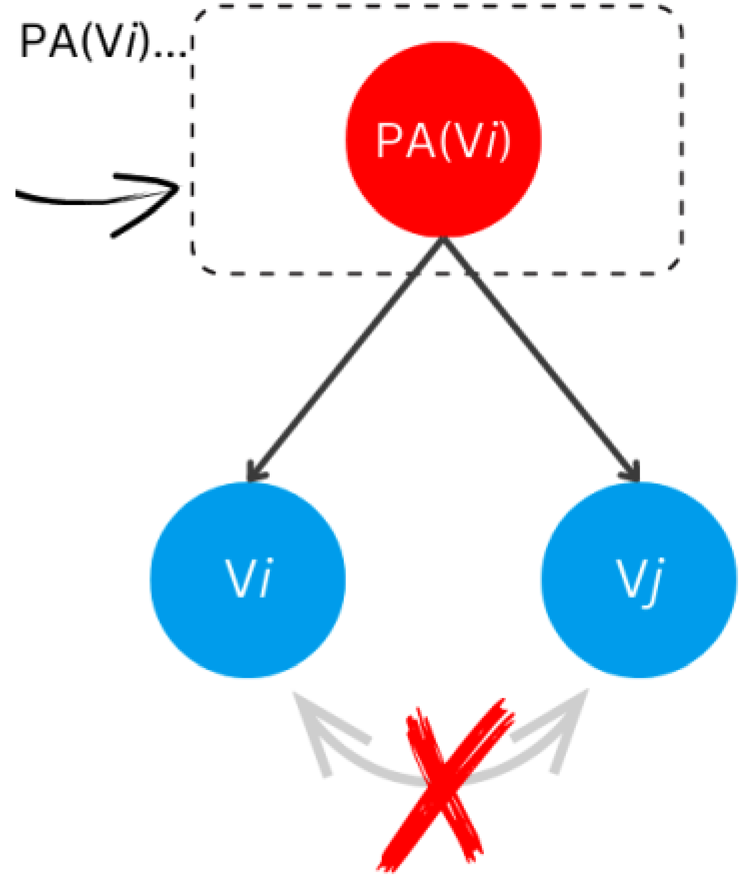
$$\boxed{V_i \perp\!\!\!\perp_{\mathrm{G}} V_j \,\middle|\, PA(V_i)\ \forall_{j \neq i \in G(V,E) \setminus \{DE(V_i),\, PA(V_i)\}}}$$

Condición causal de Markov

La clave está en si la variable es observada o no

Si controlamos la causa PA(V*i*)...





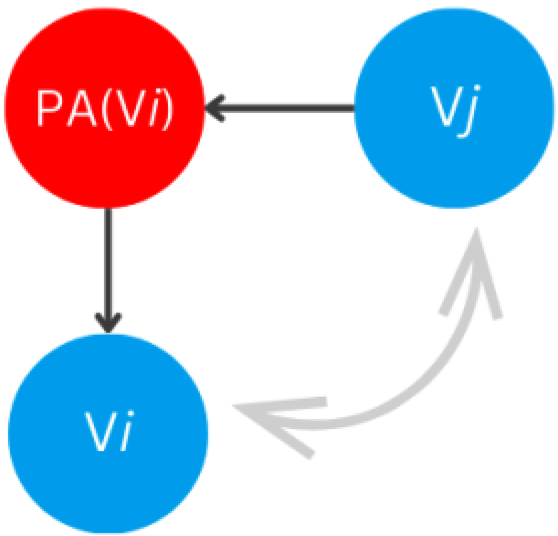
Vi y Vj son dependientes porque tienen una causa común

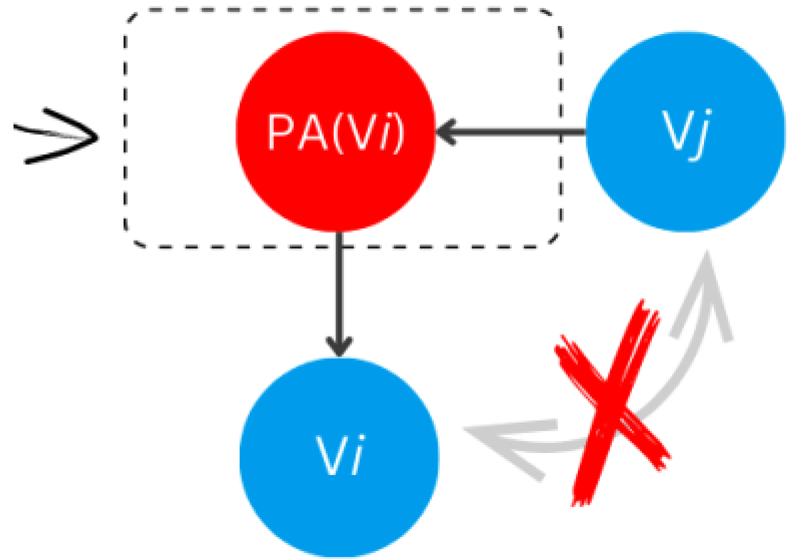
… esperamos que Vi y Vj se vuelvan **independientes**

Condición causal de Markov

La clave está en si la
variable es observada o no

Si controlamos PA(V*i*)...





Vi y Vj son dependientes porque la información fluye desde Vj a Vi pasando por PA(Vi)

… esperamos que Vi y Vj se vuelvan independientes

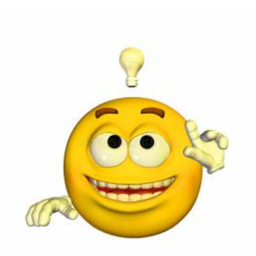
Condición causal de Markov

La condición causal de Markov es relevante ya que:

$$\boxed{X \perp\!\!\!\perp_{G} Y \mid Z \Rightarrow X \perp\!\!\!\perp_{P} Y \mid Z}$$

Observemos que el reverso no es necesariamente cierto:

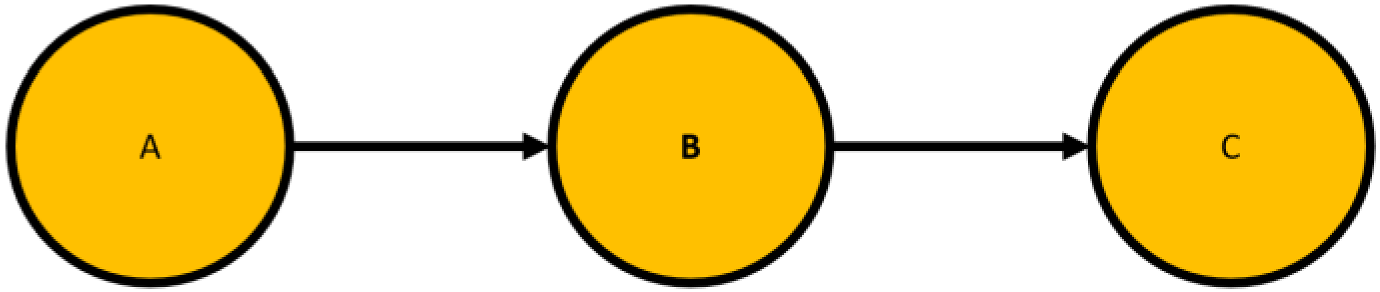
$$X \perp\!\!\!\perp_{P} Y \mid Z \Rightarrow X \perp\!\!\!\perp_{G} Y \mid Z$$



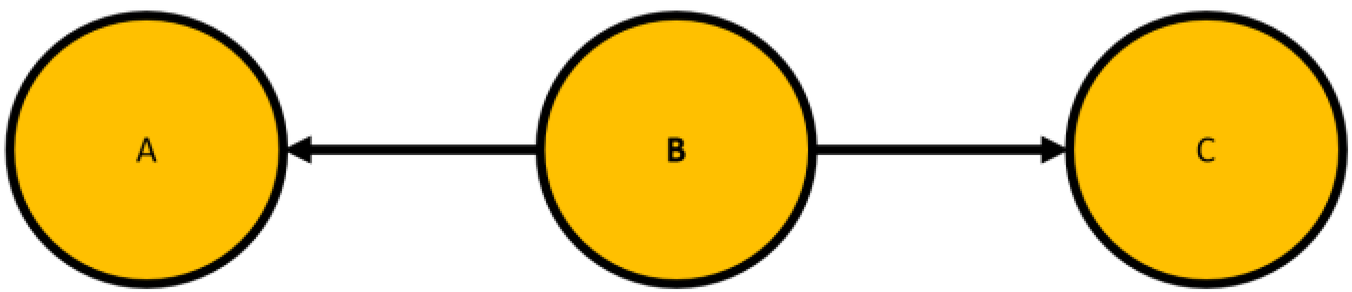
¿Por qué? Al testear independencia condicional sobre la base de una muestra finita, se introduce un error de estimación.

Estructuras frecuentes en modelos causales

Cadenas: $\boxed{A \perp\!\!\!\perp_G C \mid B}$

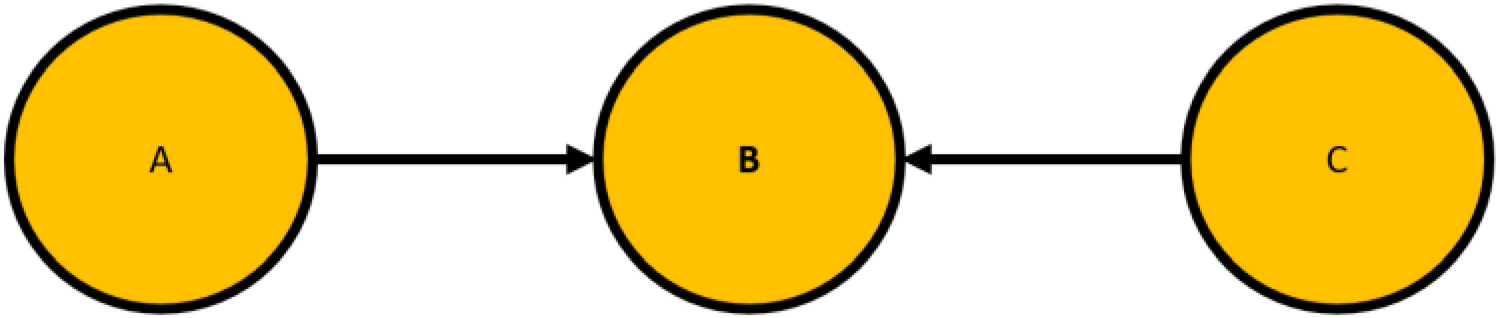


Forks: $\boxed{A \perp\!\!\!\perp C \mid B}$



Estructuras frecuentes en modelos causales

Colliders (estructuras v): $\boxed{A \perp\!\!\!\perp C}$



Sin embargo: $\boxed{A \not\perp\!\!\!\perp C \mid B}$

d-separación

Diremos que dos variables en G están *d*-separadas si todos los caminos entre ellas están bloqueados.

¿Cuándo se bloquea un camino entre dos variables?

Se bloquea el camino cuando hay un collider en el camino entre ellas y no podemos controlar el consecuente o si hay un fork o cadena que contiene una variable en el medio que podemos controlar.

Variables que podemos controlar: $\boxed{\mathcal{Z}}$

Condiciones para bloqueo entre *i* y *k*:

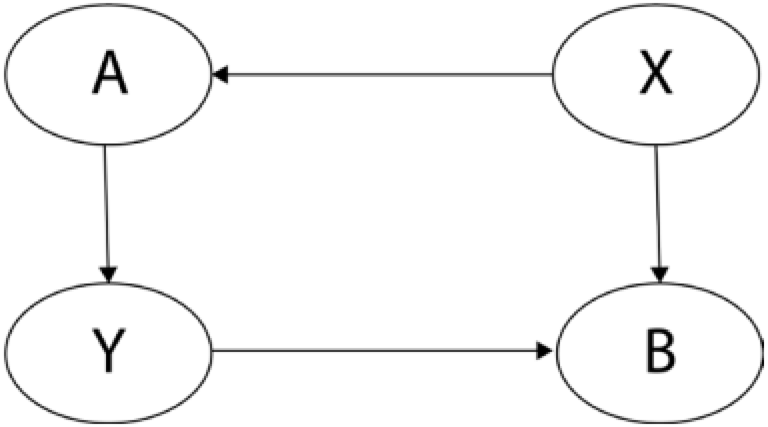
- Existen $i \leftarrow j \rightarrow k$ o bien $i \rightarrow j \rightarrow k$ tal que $j \in \mathcal{Z}$
- o bien existe $i \rightarrow j \leftarrow k$ tal que $j \notin \mathcal{Z}$

La *d* es de direccional.

d-separación

… entendiendo el concepto

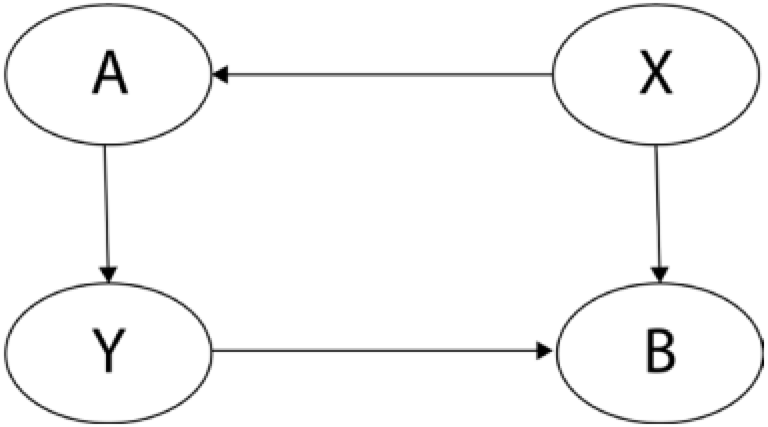
¿Cuáles variables debiera controlar para que X e Y sean *d*-separables?



d-separación

… entendiendo el concepto

¿Cuáles variables debiera controlar para que X e Y sean *d*-separables?



R.: **A**