Table of Contents

Semana 1

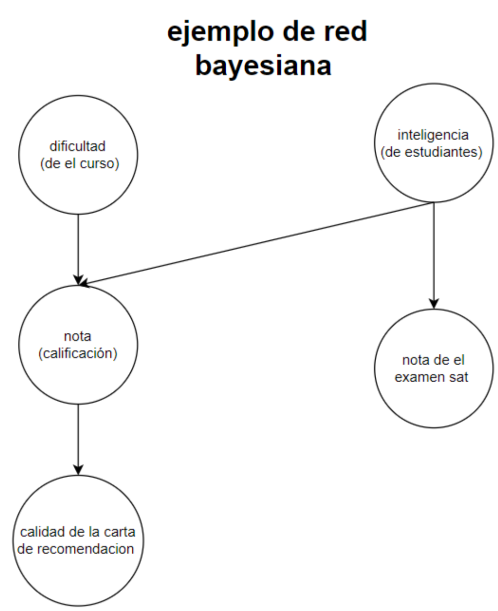
Introducción y visión general

En teoría de probabilidades y en estadística, un **modelo en grafo (MG)** representa las dependencias entre variables aleatorias como un grafo en el que cada variable aleatoria es un nodo.

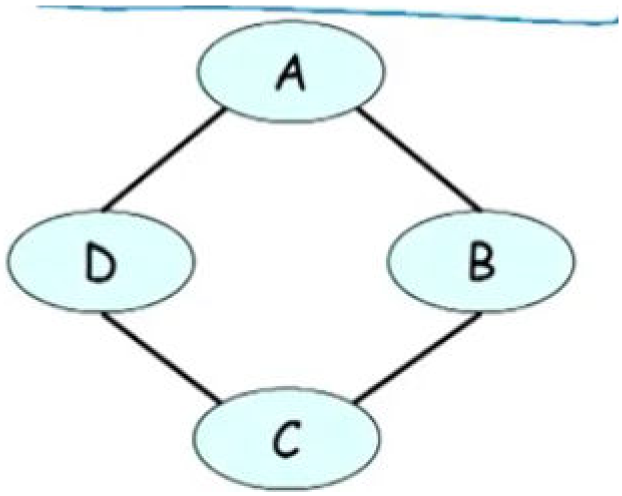
- tienen muchas aplicaciones

ejemplo de una red bayesiana (BN) simple

donde podemos ver que la nota depende de la inteligencia de el alumano y dificultad de el curso, la nota el examen de el sat depende de la inteligencia de el alumno, y la carta de recomendación que da el profesor depende de la nota de la calificación



otro ejemplo es la red de markov de graficos no dirigidos



Un ejemplo muy tipico es el de el tipo de sangre de cada persona, el tipo de sangre de los hijos depende unicamente de los padres, el de los padres unicamente de los abuelos y así sucesivamente.

Importancia de la representyación gráfica

- estructura de datos intuitiva y compacta
- eficiente razonamiento usando algoritmos de proposito general
- reducción de el número de parametros

distribución conjunta

En probabilidad, dados dos eventos aleatorios *X* y *Y*, la **distribución conjunta** de *X* e *Y* es la distribución de probabilidad de la intersección de eventos de *X* e *Y*, esto es, de los eventos *X* e *Y* ocurriendo de forma simultánea. En el caso de solo dos variables aleatorias se denomina una distribución bivariada, pero el concepto se generaliza a cualquier número de eventos o variables aleatorias.

Teorema de bayes

El **teorema de Bayes**, en la teoría de la probabilidad, es una proposición planteada por el matemático inglés Thomas Bayes (1702-1761)1� y publicada póstumamente en 1763, que expresa la probabilidad condicional de un evento aleatorio **A** dado **B** en términos de la distribución de probabilidad condicional del evento **B** dado **A** y la distribución de probabilidad marginal de solo **A**

Sea $\{A_1, A_2, ..., A_i, ..., A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos tales que la probabilidad de cada uno de ellos es distinta de cero $(P[A_i] \neq 0 \text{ para } i = 1, 2, \dots, n)$. Si B es un suceso cualquiera del que se conocen las probabilidades condicionales $P(A|B_i)$ entonces la probabilidad $P(A|B_i)$ viene dada por:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori,
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i
- $P(A_i|B)$ son las probabilidades a posteriori.

Con base en la definición de probabilidad condicionada se obtiene la Fórmula de Bayes, también conocida como Regla de Bayes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)} \cdots [1]$$

Esta fórmula nos permite calcular la probabilidad condicional $P(A_i|B)$ de cualquiera de los eventos A_i dado B. La fórmula $[1]$ «ha originado muchas especulaciones filosóficas y controversias».[3]�

Factores

Es una función o una tabla que a partir de un conjunto de variables aleatorias devuelve un valor para cada asignación a esas variables aleatorias

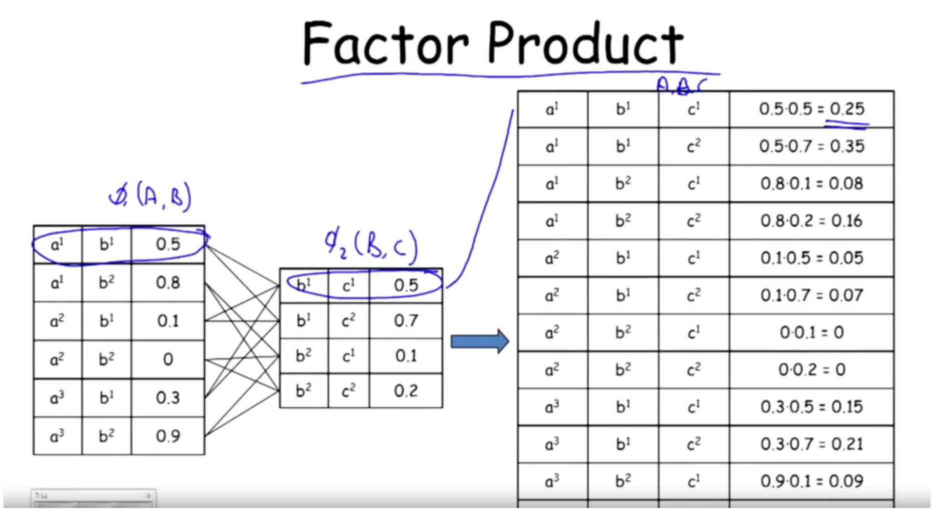
A factor $\phi(x_1, \ldots, x_k)$

- $\phi : \text{val}(x_1, \ldots, x_k) \longrightarrow \mathbb{R}$
- $\text{scope} : \{x_1, \ldots, x_k\}$ (valores que acepta el factor)

Las distribuciones de probabilidad conjunta son un factor

la funcion de probabilidad condicionada tambien es un factor

Multiplicacion de factores

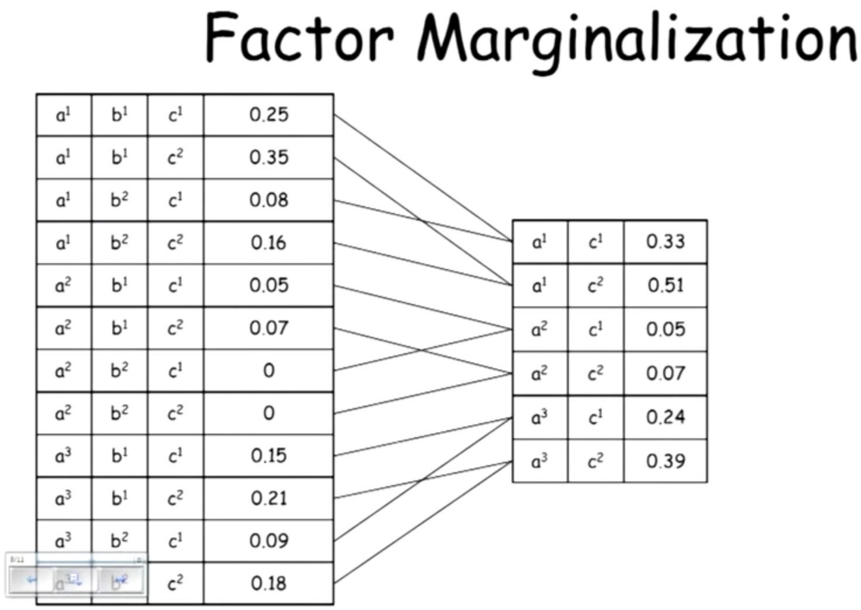


a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	0.5·0.5 = 0.25
a^1	b^1	c^2	0.5·0.7 = 0.35
a^1	b^2	c^1	0.8·0.1 = 0.08
a^1	b^2	c^2	0.8·0.2 = 0.16
a^2	b^1	c^1	0.1·0.5 = 0.05
a^2	b^1	c^2	0.1·0.7 = 0.07
a^2	b^2	c^1	0·0.1 = 0
a^2	b^2	c^2	0·0.2 = 0
a^3	b^1	c^1	0.3·0.5 = 0.15
a^3	b^1	c^2	0.3·0.7 = 0.21
a^3	b^2	c^1	0.9·0.1 = 0.09

Marginalizacion de factores

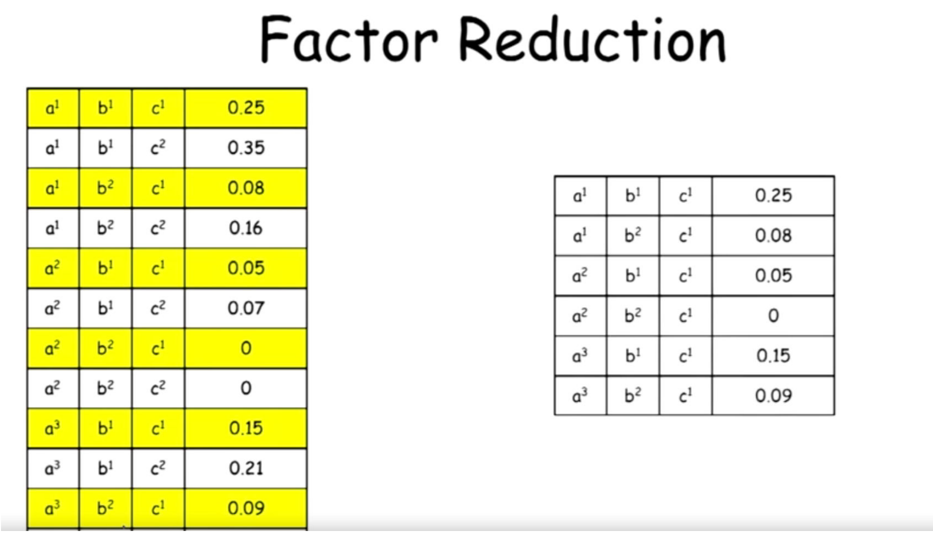


a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

a^1	c^1	0.33
a^1	c^2	0.51
a^2	c^1	0.05
a^2	c^2	0.07
a^3	c^1	0.24
a^3	c^2	0.39

Reducción de factores

C=C^1



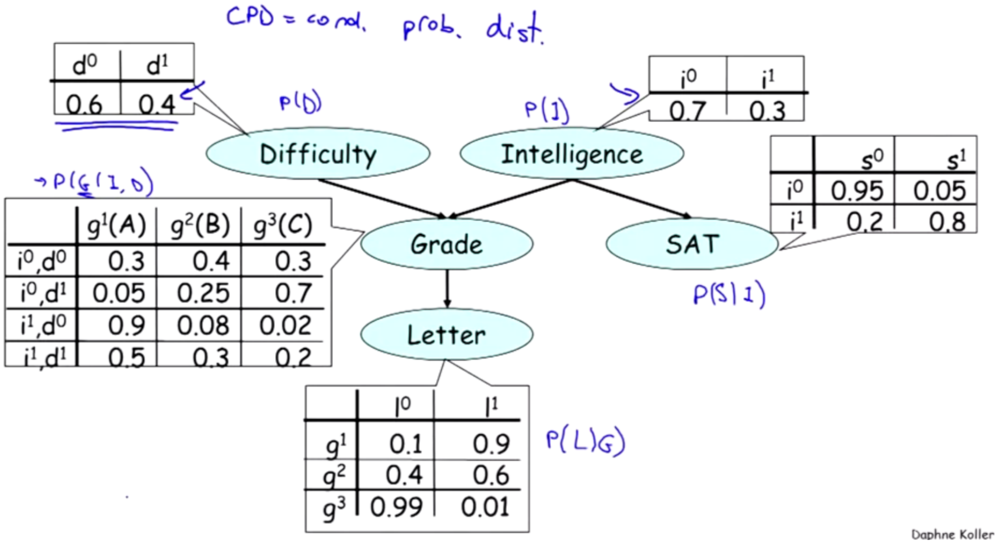
a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Fundamentos de redes bayesianas

semántica y factorizacion

retomando nuestro ejemplo de el estudiante tenemos que nuetstro factor es $P(D, I, G, S, L)$,
también podemos asignar distribuciones de probabilidad (CPD) acondicionadas a cada nodo



existe algo llamada **regla de cadena para redes bayesianas,** esta regla multiplica todas las CPD:

$$P(D, I, G, S, L) = P(D)P(I)P(\mathrm{G}|\mathrm{I}, D)P(\mathrm{S}|\mathrm{I})P(\mathrm{L}|\mathrm{G})$$

por ejemplo

$$P(D^0, I^1, G^3, S^1, L^1) = 0.6 * 0.3 * 0.02 * 0.8 * 0.01$$

entonces una red bayesiana es:

- una grafo dirigido acíclico (DAG), es decir que no tiene ciclos,
- para cada nodo x_i tenemos una CPD, y depende de sus nodos padres

Una BN representa una distribucion de probabilidad conjunta mediante la regla de la cadena para redes bayesianas

Demostración :

- $P > 0$: P es una multiplicacion de CPD las cuales son > 0, por lo tanto P >0
- $\sum P = 1$

$$\sum_{D,I,G,S,L} P(D,I,G,S,L) = \sum_{D,I,G,S,L} \underbrace{P(D)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)}_{\text{chain rule}}$$

$$= \sum_{D,I,G,S} P(D)\,P(I)\,P(G|I,D)\,P(S|I) \sum_L \overset{=1}{\cancel{P(L|G)}}$$

$$= \sum_{D,I,G,S} P(D)\,P(I)\,P(G|I,D)\,P(S|I)$$

$$= \sum_{D,I,G} P(D)\,P(I)\,P(G|I,D) \sum_S \overset{1}{\cancel{P(S|I)}}$$

$$= \sum_{D,I} P(D)\,P(I) \sum_G \cancel{P(G|I,D)}$$

Se dice P es factorizable sobre G si se puede aplicar la regla de la cadena

patrones de razonamiento

si se sabe la informacion de el nodo hijo, esta puede afetar la probabilidad de el nodo padre: si se tiene una nota alta es probable que el estudiante sea inteligente

razonamiento intercasual: dos nodos no conectados pueden afectarse entre si, por, ejemplo si una nota es alta y el grado de dificultad es alto, es muy probable que el alumno sea inteligente,

si se tienen dos nodos conectados por un padre se pueden afectar entre si, si se tuvo un puntaje alto en SAT es probable que tenga una nota alta

flujo de influencia probabilística

sendero activo: los nodos por los que pude psar la infliencia en una red

¿X influye en Y?

- $X \longrightarrow Y$ si
- $X \longleftarrow Y$ si

Si W es un nodo intermedio

- $X \rightarrow W \rightarrow Y$ si
- $X \leftarrow W \leftarrow Y$ si
- $X \leftarrow W \rightarrow Y$ si
- $X \rightarrow W \leftarrow Y$ no

una senda $X_1 - \ldots - X_n$ es activa si no tiene estructuras en V $x_{i-1} \rightarrow x_i \leftarrow x_{i+1}$

Dada la evidencia sobre un nodo Z

si $W \notin Z$

- $X \rightarrow W \rightarrow Y$ si

- $X \leftarrow W \leftarrow Y$ si
- $X \leftarrow W \rightarrow Y$ si
- $X \rightarrow W \leftarrow Y$ no solo si los hijos de w no pertenecen a z

si $w \in z$

- $X \rightarrow W \rightarrow Y$ no
- $X \leftarrow W \leftarrow Y$ no
- $X \leftarrow W \rightarrow Y$ no
- $X \rightarrow W \leftarrow Y$ si incluso con los hijos de W

una senda $X_1 - \ldots - X_n$ es activa dada Z si:

- para cualquier estructura en V $x_{i-1} \rightarrow x_i \leftarrow x_{i+1}$ tenemos que x_i o uno de sus desendientes $\in Z$
- ningun otro $X \in Z$

Independencia de las redes bayesianas

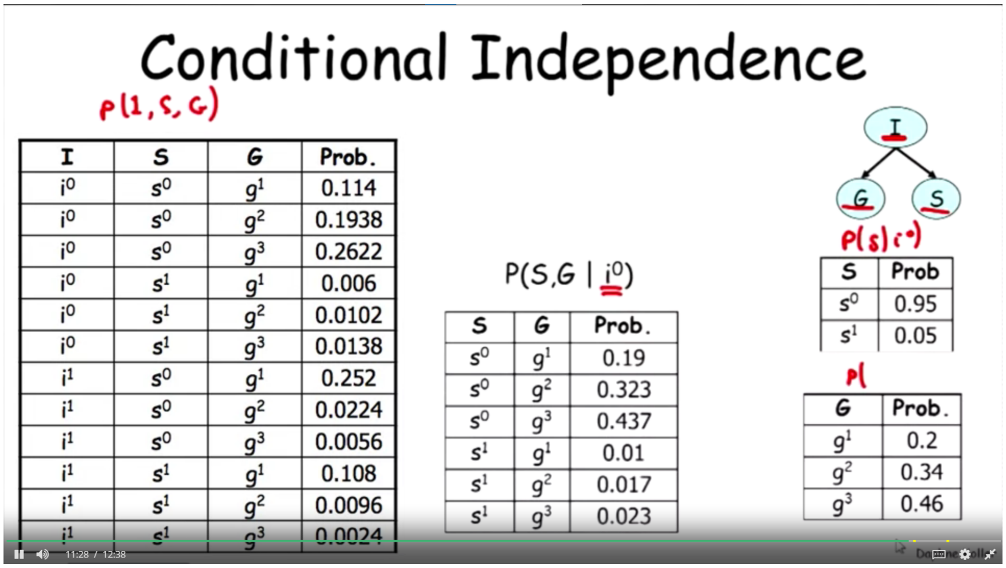
Una red bayesiana es un modelo grafo probabilístico (un tipo de modelo estático) que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un grafo acíclico dirigido (DAG por sus siglas en inglés). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades

independencia condicional

En probabilidad, dos acontecimientos R y B son **condicionalmente independientes** dado un tercer evento Y, si la ocurrencia o no ocurrencia de R junto con la de B se da en forma independiente dada Y. En otras palabras, R y B son condicionalmente independientes dado Y, si y sólo si el conocimiento que se tiene de Y provoca que el conocimiento sobre el estado de R no genere cambios sobre la probabilidad de que ocurra B, y de igual manera el conocimiento de si se produce B no proporciona información sobre la probabilidad de que ocurra R.

$P \models (X \perp Y | Z)$ si: (P satisface que X es independiente de Y dado Z)

- $P(X, Y|Z) = P(X|Z) * P(Y|Z)$
- $P(X|Y, Z) = P(X|Z)$
- $P(Y|X, Y) = P(Y|Z)$
- $P(X, Y, Z) \alpha\, \phi_1(X, Z)\ \phi(Y_2, Z)$(la probabilidad de x,y,z es proporcional a el producto de los factores $\phi_1(X, Z)$ y $\phi(Y_2, Z)$)



en este caso S y G eran depenedientes pero son condiconalmente independientes dado I, un caso contrario pasa con la estructura en V pues en nuestro ejemplo D & & I son condicionalmente dependientes dado G

ejemplo: si se escogen dos monedas, una alterada para que salga el 90% cara y otra normal y se escoge una al azar y se lanzan dos veces, si en el primer tiro sale cara entonces es más probable que en el segundo tiro salga cara porque no se sabe cual de las dos se escogio, si en cambio se sabe de antemano cual moneda se escogio entonces la probabilidad de las monedas no se ve afectada

independencias de redes bayesianas

nota: aqui G se refiere a el grafo en general

- La factorización de una distribución P sobre G implica interdependencias que mantienen P en G

definicion: X y Y estan d-separados en G dado un conjunto de observaciones Z si no hay un sendero activo en G entre X y Y dado Z: d-sep(X,Y|Z)

teorema: si P factoriza sobre G y $d-\text{sep}_G(X,Y|Z)$ entonces P satisface $(X \perp Y|Z)$

cualquier nodo en el gráfico es d-separado de sus no descendientes dado sus padres, o de otra forma si P factoriza sobre G, etnonces en P, cualquier variable es independiente de sus no desendientes dado sus padres, esto es por los casos donde si $W \in Z$ entonces se cancela el sendero activo de Y a Z y por el teorema dicho anteriormente

definamos las independencias en G como :

$$I(G) = \{(X \perp Y|Z) : d-\text{sep}_G(X,Y|Z)\}$$

definicion: si P satisface $I(G)$entonces decimos que G es un I-map (mapa de independencia, es decir todas las posibles) de P, si P no satisface todo las independencias de G(I) entonces G no es un mapa, en este caso P es la distribucion y todas las independencias de G tienen que ser independencias de P, Sin embargo, esto no significa que todas las independientes en P también estén en G.

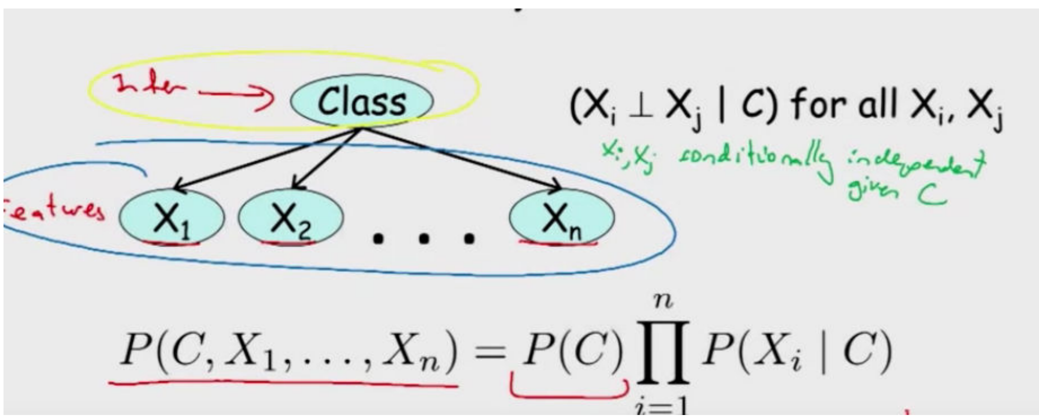
teorema: si P factoriza sobre G, entonces G es un I-Map de P, si G es un Imap para P, entonces P factoriza sobre G

Naive bayes (bayes ingenuo)

Es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de *naive*, es decir, *ingenuo*.

- Es una subclase de las redes bayesianas
- Asume que dada la variable de clase, cada variable observada es independiente de las demás variables observadas.

En términos simples, un clasificador de Naive Bayes asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable,Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Naive Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.



- Enfoque simple para clasificacion
- computacionalmente efectivo
- facil de construir
- muy efectivo en dominios con caracteristicas efectivas
- se reduce el rendimiento cuando muchas caracteristicas estan correlacionadas