

Separation and D-Separation.

- Los bordes de un grafo codifican las interacciones directas entre variables.
- Se requiere codificar interacciones indirectas.
- Grupos de variables con independencia condicional.
- Para grafos no dirigidos → Separation
- Para grafos dirigidos → D-Separation.

Separation: Si a y b conectados por caminos que incluyen variables observables (medibles)



camino activo
(No separado)



camino inactivo
(separable)
● : observable .

D-Separation (Dependence): : camino inactivo en grafo dirigido .

NOTA: Algunas distribuciones contienen independencias que no se pueden representar mediante representación gráfica.

Muestreo desde modelos gráficos.

- Generar muestras desde un modelo.
 - Para modelos dirigidos se utiliza muestreo ancestral.
 - **Ancestral sampling**: organizar variables x_i en el grafo y se muestra $x_1 \sim P(x_1)$; $x_2 \sim P(x_2 | P_{\text{par}}(x_2))$
y $x_n \sim P(x_n | P_{\text{par}}(x_n))$.
- NOTA:** Solo aplicable a grafos dirigidos.
- Para modelos no dirigidos se puede primero pasar a modelos dirigidos (tarea costosa, intractable).

Gibbs Sampling. Alternativa para grafos no dirigidos.

- Se analiza cada variable x_i
- Se muestra sobre el condicional $p(x_i | x_{-i})$
 x_{-i} : todas las variables menos x_i .
- Por análisis de separación se puede restringir la condicional solo sobre los vecinos de x_i .
- Proceso iterativo hasta convergencia.

Inferencia e Inferencia aproximada.

- **Ventaja de modelos probabilísticos:** permiten revelar como las variables se relacionan entre si.
- En modelos de variable latente, se pretenden extraer las características: $E\{h|v\}$.
 v : variables visibles
 h : variables latentes
- Generalmente, se trabaja sobre máxima verosimilitud:
 $p(h, v) = p(h|v)p(v)$
 $p(v) = p(h, v) / p(h|v)$
 $\log(p(v)) = \log(p(h, v)) - \log(p(h|v))$
- Se pretende estimar $p(h|v)$ para implementar regla de aprendizaje.
- **Inferencia:** predecir el valor de unas variables dado otras. Predecir la función de distribución de probabilidad sobre unas variables dado otras.
- NOTA:** En DL tiende a ser intractable el problema de inferencia incluso para modelos estructurados.

- Los modelos gráficos nos permiten representar problemas en altas-dimensiones en pocos parámetros
- En DL, dichos modelos no son lo suficientemente restrictivos
- Se requiere trabajar con inferencia aproximada.
- En DL inferencia aproximada se refiere a inferencia variacional.
- Inferencia Variacional aproxima $p(h|v)$ con $q(h|v)$.

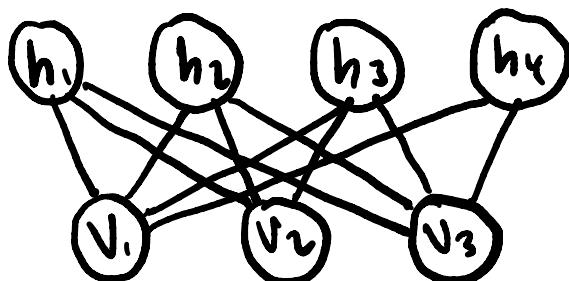
DL como modelo probabilístico estructurado

Restricted Boltzmann Machine. (RBM)

- Modelos basados en energía con variables binarias.

$$E(v, h) = -b^T v - c^T h - v^T W h$$

b, c, W : parámetros en los reales (no restringidos)



- Las restricciones en RBM se modelan desde:

$$p(h|v) = \prod_i p(h_i|v)$$

$$p(v|h) = \prod_i p(v_i|h)$$

dónde:

$$p(h_i=1|v) = \sigma(v^T w_{:,i} + b_i)$$

$$p(h_i=0|v) = 1 - p(h_i=1|v)$$

$\sigma(\cdot)$: función sigmoid.

→ RBM se resuelve mediante muestras de Gibbs.

$$\frac{\partial}{\partial w_{ij}} E(v, h) = -v_i h_j$$

Métodos de Monte Carlo (MC)

→ Métodos Las Vegas: retornan respuesta correcta
 → precisa.

→ Métodos Monte Carlo: respuesta con cantidad arbitraria
 de error → rpt. aproximada.

→ Machine Learning: modelos complejos y difícilmente per-
miten obtener la respuesta correcta.

→ ML requiere algoritmos determinísticos approximadores
o aprox. Monte Carlo.

Muestreo y MC.

→ Generar muestras desde alguna distribución de proba-
bilidad y dichas instancias se usan sobre modelos MC
para estimar cantidad deseada.

Muestreo: aprox. sumas e integrales con costo reducido.

↳ Ej. cercano: submuestreos de la función de costo por mini-lotes.

Idea: Suma o integral como valor esperado y aprox. el valor esperado por media muestral.

$$S = \sum_x p(x) f(x) = \mathbb{E}_p \{ f(x) \}$$

$$S = \int p(x) f(x) dx = \mathbb{E}_p \{ f(x) \}$$

$x_i \sim p(x) \rightarrow$ muestreo.

$$\hat{S}_N = \frac{1}{N} \sum_{i=1}^N f(x_i)$$

Ricordando el concepto de sesgo y varianza de estimadores:

$$mse(\hat{\theta}) = \mathbb{E} \{ (\hat{\theta} - \theta)^2 \} = b^2(\hat{\theta}) + var(\hat{\theta})$$

$$b(\hat{\theta}) = \mathbb{E}\{\hat{\theta}\} - \theta ; \quad var\{\hat{\theta}\} = \mathbb{E}\{(\hat{\theta} - \mathbb{E}\{\hat{\theta}\})^2\}$$

sesgo

varianza

$$\mathbb{E}\{\hat{S}_N\} = \mathbb{E}\left\{\frac{1}{N} \sum_{i=1}^N f(x_i)\right\} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}\{f(x_i)\} = \frac{1}{N} \sum_{i=1}^N S = S.$$

$$var\{\hat{S}_N\} = var\left\{\frac{1}{N} \sum_{i=1}^N f(x_i)\right\} = \frac{1}{N^2} \sum_i var\{f(x_i)\} = \frac{1}{N} var\{f(x)\}$$

S: $var\{f(x)\} < \infty$; Además:

$$\lim_{N \rightarrow \infty} \hat{S}_N = S$$

NOTA: Para estimar \hat{S}_n y $\text{Var}\{\hat{S}_n\}$:

$$\hat{S}_n = \frac{1}{N} \sum_{i=1}^N f(x_i); \quad \text{var}\{\hat{S}_n\} = \frac{1}{N} \sum_{i=1}^N (f(x_i) - \hat{S}_n)^2$$

→ $n-1$ para estimador sin sesgo.

Teorema Límite Central: $\hat{S}_n \sim N(\hat{S}_n, \text{var}(f(x))/n)$.
permite estimar intervalos de confianza desde la cumbre.

CLAVE EN MC: Muestreo desde $x \sim p(x)$.

Si no es posible muestrear sobre $p(x)$ se puede utilizar:

Importance Sampling: Determinar que parte en la suma o integral es $p(x)$ y $f(x)$.

→ La descomposición no es única:

$$p(x)f(x) = q(x) \frac{p(x)f(x)}{q(x)}$$

→ Así:

$$\hat{S}_p = \frac{1}{N} \sum_{x_i \sim p(x)} f(x_i)$$

→ Se transforma en Importance Sampling:

$$\hat{S}_q = \frac{1}{N} \sum_{x_i \sim q(x)} f(x_i) \frac{p(x_i)}{q(x_i)}$$

$$\begin{aligned} \rightarrow \mathbb{E}_q\{\hat{S}_q\} &= \frac{1}{N} \sum_{x_i \sim q} \int \frac{q(x_i)}{q(x_i)} f(x_i) p(x_i) dx = \frac{1}{N} \sum_{x_i \sim q} \mathbb{E}_p\{f(x_i)\} \\ &= S \end{aligned}$$

$$\text{Var}\{\hat{f}_q\} = \text{Var}\left\{ \frac{p(x)f(x)}{q(x)} \right\} / N$$

→ La mínima $\text{Var}\{\hat{f}_q\}$ se obtiene para:

$$q^*(x) = \frac{p(x)|f(x)|}{Z}$$

Z : factor de normalización.

NOTA: q^* generalmente no es factible → se proponen q que reduzcan la varianza.

Markov Chain Monte Carlo. (MCMC)

- Si no se cuenta con método tratable para muestrear de $p_{\text{model}}(x)$ ni se logra obtener un $q(x)$ que minimice la var del estimador en Importance Sampling.
- MCMC requiere que el modelo no asigne cero a ninguna probabilidad de estado.
- Se sugiere la forma EBM $p(x) \propto \exp(-E(x))$
- **Idea de Markov Chain:** x estado inicial en valor arbitrario. → x se itera aleatoriamente.

Markov Chain → x : random state

$T(x'|x)$: transition distribution.

Reparametrizando:

$g^{(t)}(x)$: distribución para muestra de estados.

$$g(x=i) = v_i$$

$$g^{(t+1)}(x') = \sum_x g^{(t)}(x) T(x'|x)$$

Definiendo matriz de transición A :

$$A_{ij} = T(x'=i|x=j)$$

$$\begin{aligned} v^{(t)} &= A v^{(t-1)} \\ v^{(t)} &= A^t v^{(0)} \end{aligned} \quad \begin{array}{l} \text{se multiplica } A \\ t \text{ veces.} \end{array}$$

/A: matrices estocásticas

$$v^{(t)} = (V \operatorname{diag}(\lambda) V^{-1})^t v^{(0)}$$

$$v^{(t)} = V (\operatorname{diag}(\lambda))^t V^{-1} v^{(0)}$$

En general, un MCMC con operador de transición de T converge, a un punto fijo:

$$q'(x') = \sum_{x \sim q} \{ T(x'|x) \}$$

NOTA: MCMC debe resolverse de forma heurística
y es difícil determinar su convergencia.

Gibbs Sampling

- En MCMC $x \leftarrow x' \sim T(x'|x)$
- En DL se usa Markov Chain para muestrear datos desde un modelo de energía a partir de $p_{\text{model}}(x)$.
- Se quiere aproximar $p(x)$ mediante $q(x)$ approximando $T(x'|x)$.
- En Gibbs Sampling $T(x'|x)$ se muestrea escogiendo una variable x_i , y muestreando restringiendo a sus vecinos sobre el grafo no dirigido \mathcal{G} .

NOTA: En algunos casos se parametriza los modelos de energía para controlar los picos de las distribuciones:

$$p_\beta(x) \propto \exp(-\beta E(x))$$