Aprendizaje en Redes Bayesianas

Juan Fernando Pérez
Departamento de Ingeniería Industrial
Universidad de los Andes

Febrero de 2023

jf.perez33@uniandes.edu.co
juanfperez.com

Aprendizaje en redes bayesianas

Dado un conjunto de datos $\mathcal{D} = \{d_1, \ldots, d_M\}$ con M muestras de una población P desconocida, queremos aprender el modelo $\mathcal{M} = (\mathcal{K}, \mathbf{\Theta})$ que genera P, compuesto por un grafo $\mathcal{K}$ y un conjunto de parámetros $\mathbf{\Theta}$.

Problemas de Aprendizaje en RBs

Estimación de densidades (density estimation)

Buscamos estimar una distribución de probabilidad $\tilde{P}$ que sea cercana a P.

Problemas de Aprendizaje en RBs

Predicción

Buscamos predecir el valor de un conjunto de variables $\boldsymbol{Y}$ dadas observaciones para el conjunto $\boldsymbol{X}$, a través de $P(\boldsymbol{Y}|\boldsymbol{X})$.

Clasificación: asignación de categoría a Y dadas las características X

Problemas de Aprendizaje en RBs

Descubrimiento de conocimiento

Buscamos identificar el modelo $\mathcal{M} = (\mathcal{K}, \mathbf{\Theta})$ que genera la distribución P, especialmente las relaciones que definen $\mathcal{K}$ (e.g., posibles relaciones causales).

Aprendizaje como Optimización

Medida de desempeño: función objetivo

- Cercanía entre densidad real y estimada
- Exactitud de la predicción (% categorías bien asignadas)
- Cercanía entre modelo real y estimado

Aprendizaje como Optimización

Espacio de búsqueda: variables de decisión y restricciones

- Familia de modelos y sus densidades asociadas
- Familia de modelos para predicción
- Familia de modelos

Proceso de aprendizaje

Usar los datos para el proceso de aprendizaje o **entrenamiento**.

Evaluar la capacidad del modelo aprendido para **generalizar**.

Datos de **prueba**.

Datos D

Proceso de aprendizaje

Usar los datos para el proceso de aprendizaje o **entrenamiento**.

Evaluar la capacidad del modelo aprendido para **generalizar**.

Datos de **prueba**.

Datos de entrenamiento	Datos de prueba

Proceso de aprendizaje

Usar todos los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.

Datos de entrenamiento | Datos de prueba

Proceso de aprendizaje

Usar todos los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.

Datos de prueba	Datos de entrenamiento

Proceso de aprendizaje

Usar todos los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.

Datos de entrenamiento	Datos de prueba	Datos de entrenamiento

Redes de Markov - Factores

Sea $\boldsymbol{D}$ un conjunto de variables aleatorias, $\boldsymbol{D} \subset \mathcal{X}$. Un factor ϕ es una función de $\mathrm{Val}(\boldsymbol{D})$ en $\mathbb{R}$. El conjunto de variables $\boldsymbol{D}$ es el alcance/scope de ϕ, denotado como $\mathrm{Scope}[\phi]$.

Una CPD y una distribución conjunta son factores (no al revés).

Proceso de aprendizaje

Comparar métodos de entrenamiento.

Comparar modelos.

Comparación sobre datos de prueba, no de entrenamiento.

Evitar el sobreajuste (**overfitting**) a los datos de entrenamiento.

Espacio de hipótesis

Familia de posibles modelos.

- Se conoce la estructura del grafo K. Estimar parámetros.
- No se conoce el grafo. Estimar estructura y parámetros.
- No se conocen todas las variables. Variables ocultas (hidden).
- Datos incompletos

Estimación de parámetros

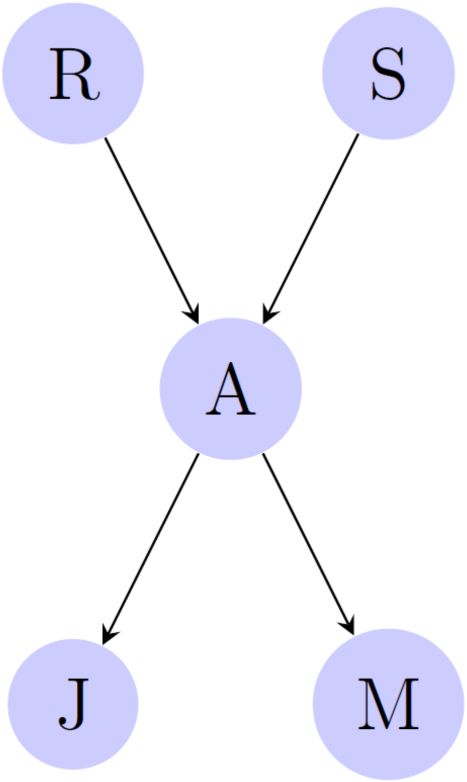
Estimación de parámetros

Se conoce la estructura del grafo $\mathcal{K}$.

Se quiere estimar los parámetros $\mathbf{\Theta}$

Estimación a partir de los datos $\mathcal{D} = \{d_1, \ldots, d_M\}$

Ejemplo: alarma



Probabilidad de robo.

Probabilidad de Sismo.

Probabilidad de activación la alarma, dado robo y sismo.

Probabilidad de llamada de Juan, dado que suena o no la alarma.

Probabilidad de llamada de María, dado que suena o no la alarma.

Estimación de parámetros

Dos grandes métodos:

- Máxima verosimilitud.
- Estimación Bayesiana.

Estimación por máxima verosimilitud

Estimación por máxima verosimilitud

Función de verosimilitud: $L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$

- Dado un conjunto de datos **D**, ¿qué tan probable/verosímil es que provengan de un distribución **P** con parámetros **Θ**?
- Seleccionar valor de parámetros **Θ** que maximicen la probabilidad de observar los datos **D**.

Ejemplo: moneda

- Se lanza una moneda al aire
- Cae en Cara con probabilidad θ
- Suponga 5 lanzamientos: C,S,C,S,C
- Función de verosimilitud:

$$P(\mathcal{D}|\theta) = P((C,S,C,S,C)|\theta) = \theta^3(1-\theta)^2$$

Ejemplo

En general

- Función de verosimilitud:

$$L(\theta|\mathcal{D}) = \theta^{mc}(1-\theta)^{ms}$$

m_C: número de caras

m_S: número de sellos

Ejemplo

En general

- Función de log-verosimilitud (logaritmo de la verosimilitud):

$$l(\theta|\mathcal{D}) = m_C \log\theta + m_S \log(1-\theta)$$

- Tiene máximo en:

$$\hat{\theta} = \frac{m_C}{m_C + m_S}$$

Ejemplo – Resumen - Bernuolli

Espacio de parametros $\Theta = [0, 1]$

Modelo de probabilidad

$$P(x|\theta) = \begin{cases} \theta, & x = C, \\ 1 - \theta, & x = S. \end{cases}$$

Función de verosimilitud

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta) = \theta^{m_C}(1-\theta)^{m_S}$$

Estimador de máxima verosimilitud

$$\hat{\theta} = \frac{m_C}{m_C + m_S}$$

Ejemplo: multinomial

Una variable aleatoria toma uno de K valores con probabilidad θ_k

Modelo de probabilidad

$$P(x|\theta) = \theta_k, \;\; x = x_k, \;\; k = 1, \ldots, K.$$

Espacio de parámetros

$$\Theta = \{\boldsymbol{\theta} \in [0,1]^K : \textstyle\sum_{i=1}^{K} \theta_i = 1\}$$

Ejemplo: multinomial

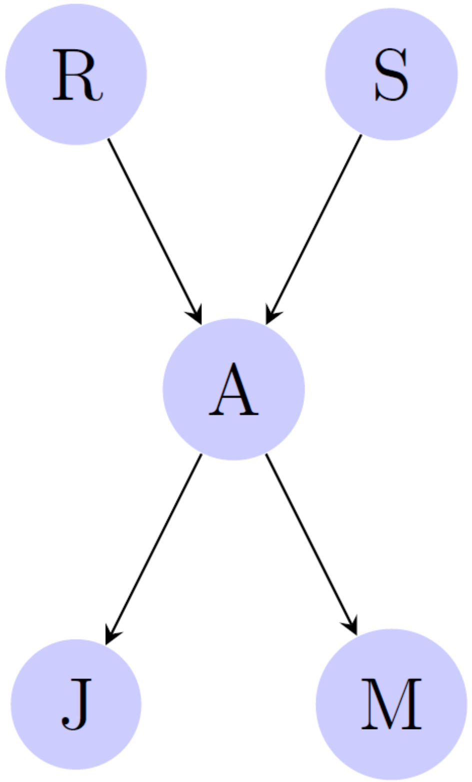
Función de verosimilitud

$$L(\boldsymbol{\theta}|\mathcal{D}) = P(\mathcal{D}|\boldsymbol{\theta}) = \theta_1^{m_1}\theta_2^{m_2}\cdots\theta_K^{m_K} = \prod_{k=1}^{K}\theta_k^{m_k}$$

Estimador de máxima verosimilitud

$$\hat{\theta}_k = \frac{m_k}{\sum_{i=1}^{K} m_i} = \frac{m_k}{m}$$

Ejemplo: alarma



R: multinomial

R puede ser V o F con probabilidades θ_V y θ_F

$\theta_V + \theta_F = 1$

$m_{R,V} = 2$

$m_{R,F} = 98$

$\theta_V = m_{R,V} / (m_{R,V} + m_{R,F}) = 0{,}02$ $\theta_F = m_{R,F} / (m_{R,V} + m_{R,F}) = 0{,}98$

MLE para redes Bayesianas

Datos $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$

Modelo de probabilidad dado un grafo $\mathcal{G}$ sobre un conjunto de variables $\mathcal{X} = \{X_1, \ldots, X_n\}$:

$$P_{\mathcal{G}}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i|\mathrm{Pa}_{X_i}^{\mathcal{G}}, \boldsymbol{\theta})$$

Función de verosimilitud

$$L(\boldsymbol{\theta}|\mathcal{D}) = P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{m=1}^{M} P_{\mathcal{G}}(\boldsymbol{d}_m|\boldsymbol{\theta})$$

MLE para redes Bayesianas

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} \left[\prod_{m=1}^{M} P(d_{m,i}|\mathrm{Pa}_{X_i}^{\mathcal{G}}, \boldsymbol{\theta})\right]$$

Función de verosimilitud local

$$L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{D}) = \prod_{m=1}^{M} P(d_{m,i}|\mathrm{Pa}_{X_i}^{\mathcal{G}}, \boldsymbol{\theta})$$

Función de verosimilitud con descomposición global

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{D})$$

MLE para RBs con CPDs en tabla

Función de verosimilitud con descomposición global

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{D})$$

Función de verosimilitud local

$$L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{D}) = \prod_{m=1}^{M} P(d_{m,i}|\mathrm{Pa}_{X_i}^{\mathcal{G}}, \boldsymbol{\theta})$$

MLE para RBs con CPDs en tabla

$$L_i(\boldsymbol{\theta}_{X_i|\boldsymbol{U}}|\mathcal{D}) = \prod_{m=1}^{M} P(d_{m,i}|\boldsymbol{U},\boldsymbol{\theta})$$

$\theta_{x|\boldsymbol{u}}$: probabilidad de $X_i = x$ dado $\boldsymbol{U} = \boldsymbol{u}$.

$m_{x,\boldsymbol{u}}$: número de veces que se observa $(x, \boldsymbol{u})$ en los datos.

$$L_i(\boldsymbol{\theta}_{X_i|\boldsymbol{U}}|\mathcal{D}) = \prod_{\boldsymbol{u}\in \mathrm{Val}(\boldsymbol{U})} \prod_{x\in \mathrm{Val}(X_i)} \theta_{x|\boldsymbol{u}}^{m_{x,\boldsymbol{u}}}$$

MLE para RBs con CPDs en tabla

Estimador de máxima verosimilitud

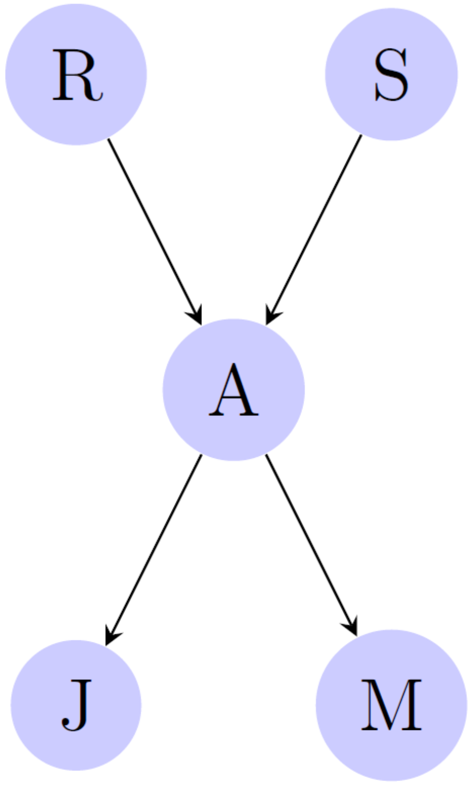
$$\hat{\theta}_{x|\boldsymbol{u}} = \frac{m_{x,\boldsymbol{u}}}{m_{\boldsymbol{u}}}$$

$$m_{\boldsymbol{u}} = \sum_{x \in \mathrm{Val}} m_{x,\boldsymbol{u}}$$

Casos **favorables**: en los que aparece el valor **x** del **nodo** y los valores **u** de los **padres**

Casos **totales**: en los que aparecen los valores **u** de los **padres**, y **cualquier valor** del **nodo**

Ejemplo: alarma



J: puede ser V o F, dado A = V o dado A = F

$m_{A=V,\ J=V} = 10$

$m_{A=V,J=F} = 30$

$m_{A=F,\ J=V} = 20$

$m_{A=F,\ J=F} = 40$

$\theta_{J=V|\ A=V} = 10\ /\ (10+30) = 0{,}25$

$\theta_{J=F|\ A=V} = 30\ /\ (10+30) = 0{,}75$

$\theta_{J=V|\ A=F} = 20\ /\ (20+40) = 1/3$

$\theta_{J=F|\ A=F} = 40\ /\ (20+40) = 2/3$

Estimación Bayesiana

Estimación Bayesiana

- Estimación por verosimilitud: totalmente basada en los datos
- Queremos incluir conocimiento previo/**a priori** de los parámetros
- Combinar conocimiento **a priori** con datos para obtener una estimación **a posteriori**
- Valorar (ponderar) conocimiento **a priori** vs. datos

Ejemplo: moneda

- Se lanza una moneda al aire
- Cae en Cara con probabilidad θ
- Suponga 5 lanzamientos: C,S,C,S,C
- Función de verosimilitud:

$$P(\mathcal{D}|\theta) = P((C, S, C, S, C)|\theta) = \theta^3(1-\theta)^2$$

Ejemplo: moneda

- Modelo de probabilidad dado θ

$$P(x_i|\theta) = \begin{cases} \theta, & x_i = C \\ 1-\theta, & x_i = S \end{cases}$$

- Densidad de θ: $P(\theta)$

- Probabilidad de observar la muestra y el valor de θ:

$$P(x_1, x_2, \ldots, x_M, \theta) = P(x_1, x_2, \ldots, x_M|\theta)P(\theta)$$

Ejemplo: moneda

- Modelo de probabilidad dado θ

$$P(x_i|\theta) = \begin{cases} \theta, & x_i = C \\ 1-\theta, & x_i = S \end{cases}$$

- Probabilidad de observar la muestra y el valor de θ:

$$P(x_1, x_2, \ldots, x_M, \theta) = P(x_1, x_2, \ldots, x_M|\theta)P(\theta)$$

$$= P(\theta) \prod_{m=1}^{M} P(x_i|\theta) \ = P(\theta)\theta^{mc}(1-\theta)^{ms}$$

Ejemplo: moneda - Predicción

$$P(x_{M+1}|x_1,\ldots,x_M) = \int_0^1 P(x_{M+1}|x_1,\ldots,x_M,\theta)P(\theta|x_1,\ldots,x_M)d\theta$$

$$= \int_0^1 P(x_{M+1}|\theta)P(\theta|x_1,\ldots,x_M)d\theta$$

Suponiendo distribución *a priori* uniforme para θ

$$P(x_{M+1}|x_1,\ldots,x_M) = \frac{m_C+1}{m_C+m_S+2}$$

Distribución Beta

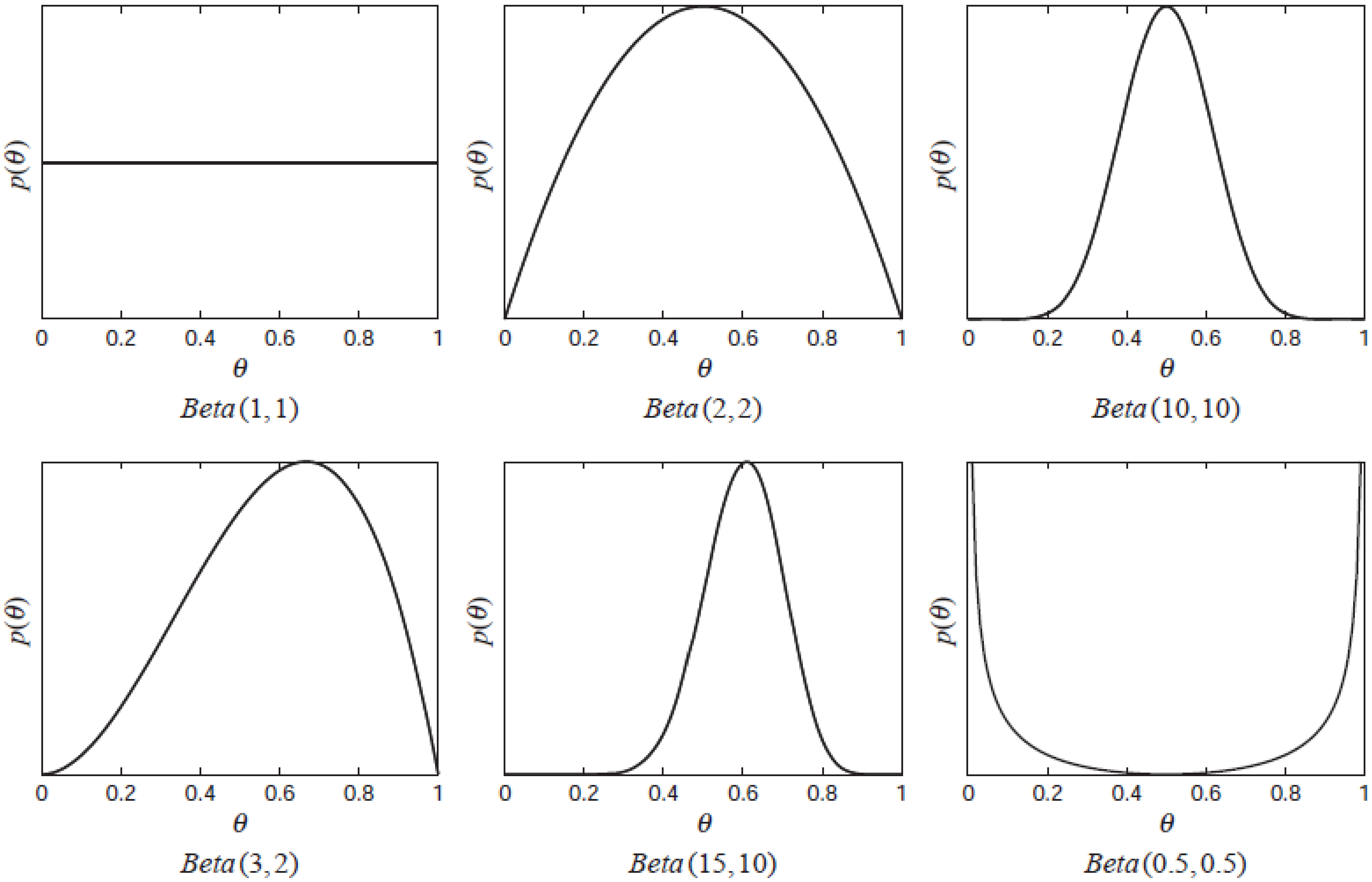
$$\theta \sim \text{Beta}(\alpha_1, \alpha_0)$$

$$p(\theta) = \gamma \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_0 - 1}$$

$$\gamma = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \qquad \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

Distribución Beta

$$\theta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$$



Distribución Beta – Ejemplo moneda

Probabilidad de obtener cara con probabilidad *a priori* Beta

$$P(X_1 = C) = \int_0^1 P(X_1 = C|\theta)P(\theta)d\theta$$

$$= \int_0^1 \theta P(\theta)d\theta = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Distribución Beta – Ejemplo moneda

$$P(X_1 = C) = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Predicción con probabilidad *a priori* Beta

$$P(x_{M+1}|x_1, \ldots, x_M) = \frac{\alpha_1+m_C}{\alpha_1+\alpha_0+m_C+m_S} = \frac{\alpha_1+m_C}{\alpha+m}$$

Casos a favor: caras observadas + caras virtuales (prior)

Casos totales: muestras totales + muestras virtuales (prior)

Estimación de Bayes en general

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})}$$

Verosimilitud marginal de observar datos D:

$$P(\mathcal{D}) = \int_{\Theta} P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\theta$$

Distribución Dirichlet

$$\theta \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

$$p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Si *a priori* θ sigue distribución Dirichlet, *a posteriori* también, con parámetros

$$\theta | \mathcal{D} \sim \mathrm{Dirichlet}(\alpha_1 + m_1, \ldots, \alpha_K + m_K)$$

Estimación de Bayes con prior Dirichlet para la multinomial

$$P(x_{m+1} = k|\mathcal{D}) = \frac{\alpha_k + m_k}{\alpha + m}$$

$$\alpha = \sum_k \alpha_k$$

$$m = \sum_k m_k$$

Casos a favor: observaciones de k reales + virtuales (prior)

Casos totales: observaciones reales + virtuales (prior)

Estimación de Bayes para CPDs en Tabla con prior Dirichlet

$$P(X_{i,(m+1)} = x_{i,k} | U_{m+1} = u, \mathcal{D}) = \frac{\alpha_{i,k|u} + m_{ik,u}}{\sum_k (\alpha_{i,k|u} + m_{i,u})}$$

Casos a favor: observaciones de k y u reales + virtuales (prior)

Casos totales: observaciones de u y cualquier k reales + virtuales (prior)

Estimación de Estructura en Redes Bayesianas

Estimación de Estructura

Dos grandes métodos:

- Aprendizaje basado en restricciones
- Aprendizaje basado en puntajes

Aprendizaje basado en Restricciones

Búsqueda de estructura que capture independencias y dependencias en los datos

Estimar (in)dependencias en los datos

Buscar estructuras (parametrizadas) que mejor la representen

Aprendizaje basado en Restricciones

Para todo nodo X_i, evaluar independencias de la forma

$$(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U} \mid \boldsymbol{U})$$

Considerando diferentes conjuntos de nodos padre U

Dificultad: número exponencial de combinaciones de nodos

Solución: búsqueda usando grafos acíclicos dirigidos (DAG)

Tiempo polinomial en el número de variables

Aprendizaje basado en Restricciones

Para cada todo nodo X_i, evaluar independencias de la forma

$$(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U} \mid \boldsymbol{U})$$

Usando prueba estadística de independencia entre variables

Dificultad: alto número de pruebas de hipótesis simultáneas (1 de cada 20 falla con valor p = 0.05)

Funciona bien con número moderado de variables, buen número de observaciones, dependencias fuertes

Aprendizaje basado en Puntaje

Tres grandes tareas:

- Explorar estructuras
- Parametrizar una estructura dada
- Evaluar la estructura

Problema de optimización

Explorar Estructuras

Buscar buenas soluciones (óptimas de ser posible) sobre:

- Grafos generales
- Restringir posibles estructuras:

  Árboles: máximo un padre por nodo

  Cardinalidad restringida: máximo número de hijos o padres

  Orden conocido: algunos nodos ancestros de otros

Explorar Estructuras

Algoritmos de búsqueda:

- Probar diferentes soluciones
- Mejorar la solución de manera iterativa
- Heurísticas de búsqueda de buenas soluciones

Parametrizar Estructura

Para una estructura candidata.

- Estimar parámetros
- Máxima verosimilitud
- Bayes

Evaluar Estructura

Determinar un indicador de la bondad de la estructura.

¿Qué tan bien describe la estructura los datos observados?

¿Qué tanto se acerca a la distribución poblacional?

Puntajes (*scores*)

Evaluar Estructura – Puntajes (*scores*)

Puntaje de **máxima verosimilitud**.

$$\mathrm{score}_L(\mathcal{G} : \mathcal{D}) = \ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$$

Log-verosimilitud.

Usando parámetros estimados por máxima verosimilitud dados los datos.

Evaluar Estructura – Puntajes (*scores*)

Puntaje de **máxima verosimilitud**.

- Al agregar un enlace se incrementa la verosimilitud.

Consecuencia (desventaja):

- Prefiere redes más complejas.
- Solo prefiere una red más simple si desde los datos se revela una independencia clara.
- Sobreajuste a los datos de entrenamiento.

Evaluar Estructura – Puntajes (*scores*)

Puntaje de **máxima verosimilitud**.

- Al agregar un enlace se incrementa la verosimilitud.

Posibles correcciones:

- Restringir espacio de búsqueda.
- Considerar redes con complejidad limitada (máximo número de padres o hijos).

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

Como no se conoce el grafo G, se define una probabilidad a priori P(G)

$$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}$$

Puntaje (numerador):

$$\text{score}_B(\mathcal{G} \; : \; \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

Como no se conoce el grafo G, se define una probabilidad a priori P(G)

$$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}$$

Puntaje (numerador):

$$\text{score}_B(\mathcal{G} \ : \ \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

$$\text{score}_B(\mathcal{G} \ : \ \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

Domina el primer término:

$$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}) P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}$$

Promedio sobre posibles valores de parámetros (no máximo)

- Evita sobreajuste al considerar efecto de parámetros.

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

$$\text{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

- Evita sobreajuste al considerar efecto de parámetros.
- Prefiere estructuras más simples, especialmente con pocos datos.
- Cuando hay más datos con evidencia, incluye más enlaces.

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

Usando una distribución *a priori* para los parámetros

$$\log P(\mathcal{D} \mid \mathcal{G}) = \ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D}) - \frac{\log M}{2}\mathrm{Dim}[\mathcal{G}] + O(1)$$

- Primer término es máxima log verosimilutud.
- Dim[G]: número de parámetros en la red.
- Segundo término: disminuye el puntaje para redes más complejas.

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

Criterio de Información Bayesiano

$$\text{score}_{BIC}(\mathcal{G} \ : \ \mathcal{D}) = \ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D}) - \frac{\log M}{2}\text{Dim}[\mathcal{G}]$$

Evaluar Estructura – Puntajes (*scores*)

Puntaje **Bayesiano**.

¿Cómo definir distribuciones *a priori* para cada posible configuración de red?

- K2: se usa Dirichlet(α, α,…,α), con $\alpha=1$
- BDe: Se define una distribución global *a priori* P', un tamaño de muestra virtual α y se usa una Dirichlet en cada caso con parámetros

$$\alpha_{x_i|\mathrm{pa}_{X_i}} = \alpha \cdot P'(x_i, \mathrm{pa}_{X_i})$$