Redes Bayesianas: Cómo se aprenden? (I)

Aritz Pérez[1] Borja Calvo[2]

Basque Center for Applied Mathematics

UPV/EHU

Donostia, Febrero de 2015

Bibliografía

Koller09: D. Koller y M. Friedman (2009). Probabilistic Graphical Models. MIT Press.

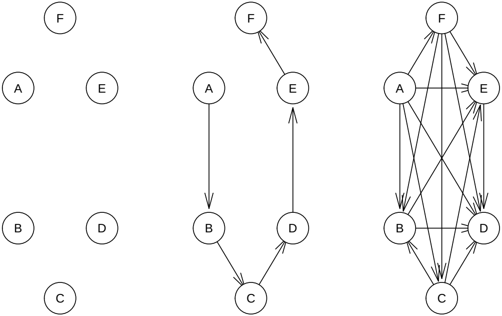
Castillo97: E. Castillo, J.M. Gutiérrez, y A.S. Hadi (1997). Sistemas Expertos y Modelos de Redes Probabilísticas. Academia de Ingeniería.

Objetivo del aprendizaje

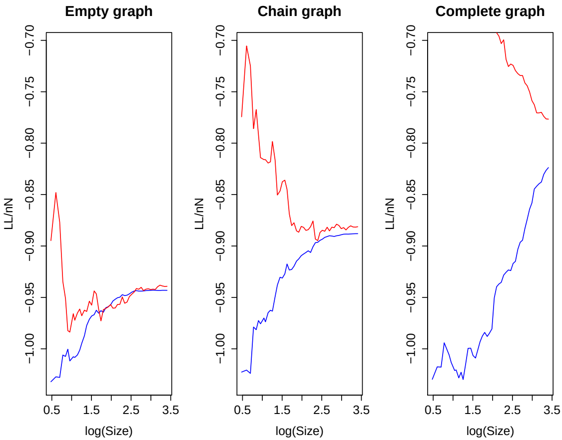
Aproximar p mediante una red Bayesiana aprendida a partir de un conjunto de datos $\mathcal{D}$ independiente e identicamente distribuido (i.i.d.) conforme a p

- Maximizar la **generalización**
- Solo conocemos el **ajuste**
- Hacer que el ajuste sea un **buen estimador** de la generalización
- Equilibrio entre el número de **parámetros** y de **casos** disponibles

Ajuste y generalización



Ajuste y generalización



Aprendizaje

- **Estructural** $\mathbf{G}$
- **Paramétrico** $\boldsymbol{\Theta}$

Aprendizaje estructural

- Aproximación **cuantitativa**: Función de evaluación
- Aproximación **cualitativa**: Test de la independencia condicionada
- Problema dificil: **NP-completo**
- **Heurísticos** de búsqueda

Aprendizaje paramétrico: máxima verosimilitud

- Forma **cerrada**: $\hat{p}(\mathbf{x}_S) = \frac{N(\mathbf{x}_S)}{N}$
- Maximiza la probabilidad del conjunto de entrenamiento
- Problemas de **sobreajuste**: Cuando el número de parámetros de $p(\mathbf{X}_S)$ es grande en relación al número de casos

Aprendizaje paramétrico: Estadística Bayesiana

- Emplear **información a priori**
- Empleando la **distribución conjugada** de la multinomial: Dirichlet
- Permite **controlar el sobreajuste**: a prioris **no informativos**
- Corrección de **Laplace**

Aproximación cuantitativa

- Maximización de una **función de evaluación**
- La función de evaluación es un **estadístico** de los datos
- Criterio que cuantifica la **calidad** de la estructura
- Log. Verosimilitud, BIC, BDe,...

Funciones aditivamente descomponibles

$$S(\mathbf{G}; \mathcal{D}) = \sum_{i=1}^{n} S(X_i, \mathbf{Pa}(X_i); \mathcal{D})$$

- **Cambios locales** en el grafo afectan localmente a la función
- Posibilita heurísticos de **busqueda eficientes**

Log Verosimilitud

Aditívamente descomponible

$$
\begin{aligned}
LL(p_M|\mathcal{D}) &= \sum_{i}^{n} -\hat{H}(X_i|\mathbf{Pa}(X_i)) \\
&= \sum_{i}^{n} -\hat{H}(X_i) + \hat{I}(X_i; \mathbf{Pa}(X_i)) \qquad (1)
\end{aligned}
$$

- Disminución lineal con n y N
- Maximizar LL equivale a **maximizar** $\sum_{i=1}^{n} \hat{I}(X_i; \mathbf{Pa}(X_i))$

Monótono creciente

Cómo cambia LL si añadimos (X_j, X_i)?

$$\hat{I}(X_i; \mathbf{Pa}(X_i), X_j) - \hat{I}(X_i; \mathbf{Pa}(X_i)) = \hat{I}(X_i; X_j | \mathbf{Pa}(X_i))$$

- $\hat{I}(X_i; X_j | \mathbf{Pa}(X_i))$ es mayor o igual que cero
- Es cero si y solo si para todo $x_i, x_j, \mathbf{pa}(X_i)$:
  $\hat{p}(x_i; x_j | \mathbf{pa}(X_i)) = \hat{p}(x_i | \mathbf{pa}(X_i)) \hat{p}(x_j | \mathbf{pa}(X_i))$
- Los datos rara vez muestran **independecias**
- Al añadir un arco la **verosimilitud es igual o mayor**

Problemas y soluciones

Favorece modelos con **muchos parámetros**, $dim(\mathbf{G})$

- EL **grafo completo** maximiza la verosimilitud
- Soluciones:
  - **Limitar** la **complejidad**
  - **Penalizar** la complejidad

Limitar la complejidad

Restringir la busqueda a un subconjunto de estructuras que acoten la **complejidad**

- Limitar el **número de padres**
- Establecer el límite en funcion del **número de casos**

Verosimilitud penalizada

$$s(\mathbf{G}) = LL(\mathbf{G}) - f(dim(\mathbf{G}))$$

- Busca un **equilibrio** entre la verosimilitud y el número de parámetros
- **Eficiencia de los paramátros** del modelo

Estadística Bayesiana

- Información **a priori**
- A prioris **no informativos**
- Interpretable como **verosimilitud penalizada**
- BIC, BDe,...

Ajuste y generalización