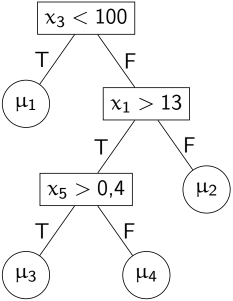
BART: Bayesian Additive Regression Trees

Luis D. Suárez

Estimación Bayesiana, Noviembre 2024

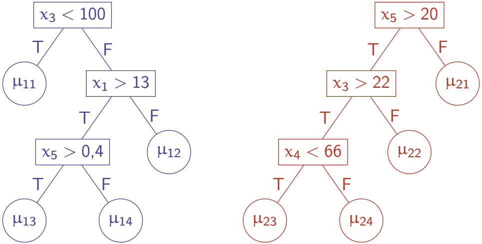
Modelo de Árbol



- Soluciones algorítmicas *"greedy"*.
- Requiere *prunning*.
- Ensembles o boosting para mejorar performance.

$$
\begin{aligned}
y = & \mu_1 I_{\{x_3<100\}} + \mu_2 I_{\{x_3 \geqslant 100\}} I_{\{x_1 \leqslant 13\}} \\
& + \mu_3 I_{\{x_3 \geqslant 100\}} I_{\{x_1 > 13\}} I_{\{x_5 > 0{,}4\}} \\
& + \mu_4 I_{\{x_3 \geqslant 100\}} I_{\{x_1 > 13\}} I_{\{x_5 \leqslant 0{,}4\}} + \varepsilon
\end{aligned}
$$

Suma de árboles



$$Y = \sum_{j=1}^{m} g_j(x) + \varepsilon$$

$$\begin{aligned} y = & \mu_{11} I_{\{x_3 < 100\}} + \mu_{12} I_{\{x_3 \geqslant 100\}} I_{\{x_1 \leqslant 13\}} + \mu_{13} I_{\{x_3 \geqslant 100\}} I_{\{x_1 > 13\}} I_{\{x_5 > 0{,}4\}} \\ & + \mu_{14} I_{\{x_3 \geqslant 100\}} I_{\{x_1 > 13\}} I_{\{x_5 \leqslant 0{,}4\}} \\ & + \mu_{21} I_{\{x_5 \leqslant 20\}} + \mu_{22} I_{\{x_5 > 20\}} I_{\{x_3 \leqslant 22\}} + \mu_{23} I_{\{x_5 > 20\}} I_{\{x_3 > 22\}} I_{\{x_4 < 66\}} \\ & + \mu_{24} I_{\{x_5 > 20\}} I_{\{x_3 > 22\}} I_{\{x_4 \geqslant 66\}} + \varepsilon \end{aligned}$$

BART: un modelo bayesiano de suma de árboles

Modelo de árboles aditivos:

$$Y = \sum_{j=1}^{m} g(x; \mathcal{T}_j, \mathcal{M}_j) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Prior:

$$p\Big((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma\Big)$$

Se propone utilizar un priores bajo el supuesto de independencia de cada $(\mathcal{T}_j, \mathcal{M}_j)$ entre sí y con respecto a σ. Es decir que

$$\text{Prior} = \left[\prod_j p(\mathcal{M}_j \mid \mathcal{T}_j,)p(\mathcal{T}_j)\right] p(\sigma), \quad \text{con } p(\mathcal{M}_j \mid \mathcal{T}_j,) = \prod_i^{b_j} p(\mu_{ij} \mid \mathcal{T}_j)$$

Especificación de los priores de BART

- $\mathcal{T}_j$ está especificado por tres aspectos:
  - La probabilidad de que un nodo a profundidad d sea no-terminal.
  - La distribución de las variables de división.
  - La distribución de los valores de división.

Especificación de los priores de BART

- $\mathcal{T}_j$ está especificado por tres aspectos:
  - La probabilidad de que un nodo a profundidad d sea no-terminal.
  - La distribución de las variables de división.
  - La distribución de los valores de división.
- $p(\mu_{ij} \mid \mathcal{T}_j)$
  Se puede utilizar un prior conjugado Normal.
  Si se reescala el target, e.g. (-05, 05), se puede simplificar a

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu^2)$$

Especificación de los priores de BART

- $\mathcal{T}_j$ está especificado por tres aspectos:
  - La probabilidad de que un nodo a profundidad d sea no-terminal.
  - La distribución de las variables de división.
  - La distribución de los valores de división.
- $p(\mu_{ij} \mid \mathcal{T}_j)$
  Se puede utilizar un prior conjugado Normal.
  Si se reescala el target, e.g. (-05, 05), se puede simplificar a

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu^2)$$

- σ
  Un prior conjugado Gamma inversa (típicamente χ^2 inversa).

$$\sigma \sim \nu\lambda / \chi_\nu^2$$

Especificación de los priores de BART

- $\mathcal{T}_j$ está especificado por tres aspectos:
  - La probabilidad de que un nodo a profundidad d sea no-terminal.
  - La distribución de las variables de división.
  - La distribución de los valores de división.
- $p(\mu_{ij} \mid \mathcal{T}_j)$
  Se puede utilizar un prior conjugado Normal.
  Si se reescala el target, e.g. (-05, 05), se puede simplificar a

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu^2)$$

- σ
  Un prior conjugado Gamma inversa (típicamente χ^2 inversa).

$$\sigma \sim \nu\lambda/\chi_\nu^2$$

- m

Muestreo del posterior

Bayesian Backfitting MCMC

- Objetivo: extraer m muestras de $(\mathcal{T}_j, \mathcal{M}_j) \mid \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \sigma, y$

$$R_j \equiv y - \sum_{k \neq j} g(x; \mathcal{T}_k, \mathcal{M}_k)$$

El problema se reduce a obtener un árbol por vez usando R_j en lugar de y.

Muestreo del posterior

Bayesian Backfitting MCMC

- Objetivo: extraer m muestras de $(\mathcal{T}_j, \mathcal{M}_j) \mid \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \sigma, y$

$$R_j \equiv y - \sum_{k \neq j} g(x; \mathcal{T}_k, \mathcal{M}_k)$$

El problema se reduce a obtener un árbol por vez usando R_j en lugar de y.

  - Generar un nuevo árbol $\mathcal{T}_j^*$ a partir de un set de reglas:

Grow Prune Swap Change

Muestreo del posterior

Bayesian Backfitting MCMC

- Objetivo: extraer m muestras de $(\mathcal{T}_j, \mathcal{M}_j) \mid \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \sigma, y$

$$R_j \equiv y - \sum_{k \neq j} g(x; \mathcal{T}_k, \mathcal{M}_k)$$

El problema se reduce a obtener un árbol por vez usando R_j en lugar de y.

  - Generar un nuevo árbol $\mathcal{T}_j^*$ a partir de un set de reglas:

    Grow Prune Swap Change

    - Metropolis-Hastings para seleccionar entre $\mathcal{T}_j$ y $\mathcal{T}_j^*$

Conclusión

- BART puede capturar interacciones complejas entre variables predictivas.
- Tiene pocos hiperparámetros que pueden fácilmente determinarse con valores default.
- Puede ser extendido fácilmente (e.g. para clasificación).
- El outcome es un *posterior de sumas de árboles*:
  - Predicciones con intervalos de credibilidad.
  - Produce medidas directas para *Feature Selection*.

¡Muchas Gracias!

…preguntas?