

Random Forest

Bagging, Subbagging, Bragging, Out-of-Bag Error

Fernando Arias-Rodríguez

Banco Central de Bolivia

30 de agosto de 2024



- ① Introducción
- ② Métodos de agregación de modelos
- ③ Random Forest
- ④ Inferencia

- 1 Introducción
- 2 Métodos de agregación de modelos
- 3 Random Forest
- 4 Inferencia

- *Random Forest* surge como una manera de solucionar el problema de bajo poder predictivo de los modelos de árboles o *Regression Trees*.
- El método más básico sobre el que *Random Forest* funciona es mediante la agregación de muchos *Regression Trees*.
- Así, antes de explicar *Random Forest*, se hace necesario introducir algunos métodos de agregación de modelos: *Bagging*, *Subbagging*, *Bragging*.
- Se introducirá un método que sirve para evaluar el error fuera de muestra de los modelos estimados, conocido como *Out-of-Bag Error*.

1 Introducción

2 Métodos de agregación de modelos

Bagging

Subagging

Bragging

Out-of-Bag Error

3 Random Forest

4 Inferencia

- *Bagging* es una contracción de *Bootstrap Aggregating*.
- Dada una muestra y un método de estimación, *Bagging* puede disminuir la varianza de un estimador, comparado con el que se calcula a partir de solo la muestra original.
- Considere una muestra $\{(y_1, x_1), \dots, (y_N, x_N)\}$ donde $y_i \in \mathbb{R}$ es la variable dependiente y $x_i \in \mathbb{R}^p$ son las p variables explicativas.
- Suponga que el proceso generador de datos es $y = E(y|x) + u = f(x) + u$ donde $E(u|x) = 0$ y $Var(u|x) = \sigma^2$.

- Para estimar la media condicional de y dado x , $E(y|x) = f(x)$ se escoge una función $\hat{f}(x)$ tal que minimice la función de pérdida dada por:

$$\min_{\hat{f}} \sum_{i=1}^N (y_i - \hat{f}(x_i))^2 \quad (1)$$

- Si partimos de que $\hat{f}(x)$ es una **función no lineal**, esta puede sufrir del riesgo de **sobreajuste**.
- Considere la descomposición del Error Cuadrático Medio entre sesgo y varianza:

$$\begin{aligned} MSE &= E(y - \hat{f}(x))^2 \\ &= E[y - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x)]^2 \end{aligned}$$

- Agrupando términos:

$$\begin{aligned} MSE = E \left[(y - E[\hat{f}(x)])^2 \right] &+ E \left[(\hat{f}(x) - E[\hat{f}(x)])^2 \right] \\ &+ 2E \left[(y - E[\hat{f}(x)]) (E[\hat{f}(x)] - \hat{f}(x)) \right] \end{aligned}$$

- El último término de esta expresión se puede reducir a cero, dado que $\hat{f}(x)$ & y son independientes.
- Reemplazando y por su proceso generador de datos $f(x) + u$, el MSE puede descomponerse en:
 - 1 Sesgo cuadrático, medido por $E \left[(f(x) - E[\hat{f}(x)])^2 \right]$.
 - 2 La varianza, medida como $E \left[(\hat{f}(x) - E[\hat{f}(x)])^2 \right]$.
 - 3 La varianza del error u , σ^2 .

- En resumen, el error cuadrático medio se puede descomponer así:

$$MSE = Sesgo^2 + Var + \sigma^2 \quad (2)$$

- Nótese que mientras más elaborada sea $\hat{f}(x)$, mejor será el pronóstico y más bajo será el sesgo.
- Sin embargo, al mismo tiempo la varianza será más grande.
- Así, no siempre el conjunto de parámetros que minimicen la función de pérdida serán los que logren la menor varianza, por lo que el MSE será alto. **Esto se conoce como el riesgo de sobreajuste.**
- *Bagging* es una alternativa para controlar la varianza de $\hat{f}(x)$.

El procedimiento de *Bagging* se compone de los siguientes pasos:

- 1 De la muestra original, se genera una muestra *bootstrap*, $\{(y_1^b, x_1^b), \dots, (y_N^b, x_N^b)\}$ mediante un muestreo con reemplazamiento ($b = 1, \dots, B$).
- 2 Para cada muestra *bootstrap*, se estima $\hat{f}_b(x)$ mediante la minimización de la función de pérdida

$$\min_{\hat{f}_b(x)} \sum_{i=1}^N (y_i^b - \hat{f}_b(x_i^b))^2$$

- 3 Se combinan todos los pronósticos estimados $\hat{f}_1(x), \dots, \hat{f}_B(x)$ para construir la estimación *Bagging*:

$$\hat{f}_{bagging}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(x)$$

1 Introducción

2 Métodos de agregación de modelos

Bagging

Subagging

Bragging

Out-of-Bag Error

3 Random Forest

4 Inferencia

- Guarda el mismo principio del anterior método, pero en este caso se implementa muestreo sin reemplazamiento.
- Comparado con métodos de *Bootstrap*, este provee resultados similares sin involucrar largos procesos computacionales.
- Sea d el número de elementos de la muestra contenidos en cada submuestra.
- Como el muestreo es sin reemplazamiento, el número de submuestras es $M = \binom{N}{d}$.
- *Subagging*, entonces, agrega predictores que se desprenden de modelos entrenados con muestras derivadas del sub-sampleo propuesto arriba.

Subbagging se compone de los pasos:

- 1 Con la muestra original, se construyen $M = \binom{N}{d}$ diferentes submuestras $\{(y_1^m, x_1^m), \dots, (y_d^m, x_d^m)\}$ al tomar M muestras sin reemplazamiento, $m = 1, \dots, M$.
- 2 Para cada submuestra, se estima $\hat{f}_m(x)$:

$$\min_{\hat{f}_m(x)} (y_i^m - \hat{f}_m(x_i^m))^2$$

- 3 Se combinan todos los modelos resultantes, $\hat{f}_1(x), \dots, \hat{f}_M(x)$

$$\hat{f}_{\text{subbagging}}(x) = \frac{1}{M} \sum_{m=1}^M \hat{f}_m(x)$$

- 4 Se suele escoger $d = \alpha N$ con $0 < \alpha < 1$. Dado que d se relaciona al costo computacional de hacer *subbagging*, lo más usado es $d = N/2$.

1 Introducción

2 Métodos de agregación de modelos

Bagging

Subagging

Bragging

Out-of-Bag Error

3 Random Forest

4 Inferencia

- Cuando se considere que la muestra puede tener valores atípicos, se considera utilizar la mediana en lugar de la media.
- Este es precisamente el aporte de esta metodología: construir un estimador agregado que sea robusto a valores atípicos.
- *Bragging* es una contracción para *Bootstrap Robust Aggregating*.
- Los dos primeros pasos de *Bragging* son iguales a los de *Bagging*.
- La diferencia radica en la agregación de los resultados. En este caso, se combinan los modelos estimados así:

$$\hat{f}_{bragging}(x) = \text{mediana}(\hat{f}_b(x); b = 1, \dots, B)$$

1 Introducción

2 Métodos de agregación de modelos

Bagging

Subagging

Bragging

Out-of-Bag Error

3 Random Forest

4 Inferencia

- Cuando se aplican técnicas de *Bootstrapping*, hay datos que no se seleccionan en el remuestreo, con una probabilidad dada por:

$$P((x_i, y_i) \notin \text{Boot}_b) = \left(1 - \frac{1}{N}\right)^N \rightarrow e^{-1} \approx 37$$

- En otras palabras, el 37% de los datos de la muestra original no quedan en los remuestreos.
- Sin embargo, estas observaciones se convierten en una muestra de evaluación muy útil. Se conocen como muestra *Out-of-Bag* (OOB sample).
- El error de la estimación $\hat{f}_b(x)$ hecho con la OOB sample se conoce como *Out-of-Bag Error* (OOB Error), equivalente al error generado al utilizar una muestra de evaluación.

- El OOB error se calcula como:

$$\begin{aligned}\hat{err}_{OOB,b} &= \frac{\sum_{i=1}^N I((y_i, x_i) \notin Boot_b) \times Loss(y_i, \hat{f}_b(x_i))}{\sum_{i=1}^N I((y_i, x_i) \notin Boot_b)} \\ &= \frac{1}{N_b} \sum_{i=1}^{N_b} Loss(y_{i,b}^{OOB}, \hat{f}_b(x_{i,b}^{OOB}))\end{aligned}\tag{3}$$

El procedimiento de implementar OOB error es el siguiente

- 1 Basado en la muestra original se generan B muestras *bootstrap* $\{(y_1^b, x_1^b), \dots, (y_N^b, x_N^b)\}$.
- 2 Para cada muestra *bootstrap*, se estima $\hat{f}_b(x)$ con la minimización de la función de pérdida:

$$\min_{\hat{f}_b(x)} \sum_{i=1}^N \text{Loss}(y_i^b - \hat{f}_b(x_i^b))$$

- 3 Comparar la b – esima muestra *bootstrap* con la muestra original, para generar la b – esima OOB sample $\{(y_{1,OOB}^b, x_{1,OOB}^b), \dots, (y_{N_b,OOB}^b, x_{N_b,OOB}^b)\}$.
- 4 Calcular el *Out-of-Bag error* de $\hat{f}_b(x)$ de todas las OOB samples, como en la Ecuación 3

- 1 Introducción
- 2 Métodos de agregación de modelos
- 3 Random Forest**
- 4 Inferencia

- *Random Forest* es una combinación de muchos *Regression Trees*, utilizando *Bagging* como método de agregación.
- Recuérdese que métodos como *Bagging* tienen el atractivo de disminuir la varianza del Error Cuadrático Medio.

$$MSE = \text{Sesgo}^2 + \text{Var}^2 + \sigma^2$$

- Sin embargo, si se combina un conjunto de estimadores insesgados, pero correlacionados, la varianza no disminuirá como se espera.

- Considere B estimadores **insesgados**, f_1, \dots, f_B , con la misma varianza, σ^2 . Si dichos estimadores son *i.i.d.*, la varianza del promedio de los estimadores es

$$\text{Var}(g) = \text{Var}\left(\frac{1}{B} \sum_{b=1}^B f_b\right) = \frac{1}{B} \sigma^2$$

- Si los estimadores insesgados están correlacionados, la varianza del promedio de estimadores es ahora:

$$\begin{aligned} \text{Var}(g) &= \frac{1}{B^2} \text{Var}\left(\sum_{b=1}^B f_b\right) \\ &= \frac{1}{B^2} \left(\sum_{b=1}^B \text{Var}(f_b) + 2 \sum_{b \neq c} \text{cov}(f_b, f_c) \right) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{B^2} (B\sigma^2 + (B^2 - B)\rho\sigma^2) \\ &= \rho\sigma^2 + \frac{(1 - \rho)}{B}\sigma^2 \end{aligned} \tag{4}$$

- En este caso, ρ es el coeficiente de correlación entre dos estimadores. En general, la varianza del promedio de estimadores depende del **número de estimadores y la correlación entre estos**.
- Aun si se aumenta el número de estimadores (mayor B , lo que hace más pequeño el término $\frac{(1-\rho)}{B}\sigma^2$), el término $\rho\sigma^2$ se mantendrá inalterado, impidiendo disminuir la varianza.

- En la práctica, el promedio de estimadores derivados de *Regression Trees* similares son más robustos, pero al estar altamente correlacionados no se desempeñan mejor que un solo modelo.
- *Random Forest* propone un camino para disminuir ambos términos de la Ecuación 4.
- Para disminuir la correlación entre estimadores, este introduce el procedimiento conocido como *random subset projection* (RSP) o *random feature projection* durante el proceso de crecimiento de un árbol:
 - En lugar de usar todas las variables en todos los árboles, cada árbol se construye con un subconjunto aleatorio de variables en cada ramificación.
 - En lugar de podar el árbol, este se deja "crecer" hasta alcanzar el criterio de detención estipulado.

- RSP puede disminuir las correlaciones entre los árboles, dado que crecerán con diferentes atributos (variables), llevando a un $\rho\sigma^2$ menor.
- Sin embargo, puede afectar el término $\frac{(1-\rho)}{B}\sigma^2$, en la medida en que cada árbol no está usando toda la información disponible para predecir.
- Así, es necesario seleccionar el número de variables en cada ramificación de manera que se pueda balancear la minimización de ambos términos.

El procedimiento para implementar *Random Forest* es el siguiente:

- Generar B conjuntos de muestras *bootstrap*.
- En cada muestra, implementar un árbol sin penalización.
- Durante el crecimiento del árbol, seleccionar **aleatoriamente** m variables en cada potencial ramificación (RSP).
- Combinar los B árboles resultantes para crear un *Random Forest*. Se toma el promedio del resultado para todos los árboles (regresión).

Podría escogerse m a partir de validación cruzada, pero es demasiado demandante en tiempo.

Se escoge m tal que $1 \leq m \leq p/3$, con p igual al número de variables.

Cada árbol crecerá hasta que en cada nodo o rama tenga mínimo 5 observaciones (estándar).

- 1 Introducción
- 2 Métodos de agregación de modelos
- 3 Random Forest
- 4 Inferencia**
 - Importancia de una variable
 - Aplicaciones y extensiones

- 1 Introducción
- 2 Métodos de agregación de modelos
- 3 Random Forest
- 4 Inferencia**
 - Importancia de una variable
 - Aplicaciones y extensiones

- Como *Random Forest* es una combinación lineal de árboles, aquí puede usarse la medida de importancia relativa promedio:

$$I_j^2 = \frac{1}{B} \sum_{b=1}^B I_j^2(b) \quad (5)$$

con $I_j^2(b)$ igual a la importancia relativa del b – *esimo* árbol,

$$I_j^2(b) = \sum_{t=1}^{T_b-1} e_t^2 I(v(t)_b = j)$$

- Aplicar esta medida implica revisar cada nodo de un árbol, lo que no es eficiente dada la gran cantidad de árboles detrás de *Random Forest*.

- Se puede usar permutaciones aleatorias para sortear con esta dificultad.
- La idea es: para una variable, se permutan las muestras usando permutación aleatoria.
- Partiendo de un *Random Forest*, para una variable j a lo largo de **todas las muestras** $x_j = (x_{j,1}, x_{j,2}, \dots, x_{j,N})$, se reordenan aleatoriamente todos los x s para generar una nueva serie de muestras $x_j^* = (x_{j,1}^*, x_{j,2}^*, \dots, x_{j,i}^*, \dots, x_{j,N}^*)$.
- Por ejemplo, para x_j puede tenerse un ordenamiento aleatorio igual a $(x_{j,2}, x_{j,10}, \dots, x_{j,N-4}, \dots, x_{j,i+5})$.

- Una forma de estimar el error es subdividir la muestra entre de entrenamiento y evaluación y estimar el error con esta última. Esto no es eficiente, dada la pérdida de datos en el remuestreo *Bootstrap*.
- Alternativa: usar el *Out-of-Box error*.
- Con cada remuestreo, se puede tomar tanto lo que queda en muestra *Bootstrap* como lo que queda fuera.
- La medida de importancia de una variable j se calcula como:

$$\begin{aligned} VI_j^{OOB} &= \frac{1}{B} \sum_{b=1}^B \left(\hat{err}_{OOB,b}^* - \hat{err}_{OOB,b} \right) \\ &= \frac{1}{B} \sum_{b=1}^B \Delta \hat{err}_{OOB,b} \end{aligned} \quad (6)$$

En últimas, la implementación de esta metodología es:

- 1 Para la b – *esima* muestra *bootstrap*, se pone a crecer un árbol de regresión.
- 2 Se hallan los puntos no incluidos en la muestra de estimación y con ellos se construye la b – *esima* muestra OOB.
- 3 Se computa el OOB error para el árbol de regresión b , basado en la muestra OOB con y sin permutación aleatoria.
- 4 Se calcula VI_j^{OOB} para medir la importancia de la variable j .

- 1 Introducción
- 2 Métodos de agregación de modelos
- 3 Random Forest
- 4 Inferencia**
 - Importancia de una variable
 - Aplicaciones y extensiones

- Principalmente, tanto *Bagging* y sus variaciones como *Random Forest* son metodologías enfocadas a mejorar el pronóstico de variables.
- Para *Bagging* se documentan aplicaciones con LARS, Modelos de Factores Dinámicos, Regresiones Ridge, LASSO.
- *Random Forest* resulta atractivo por la fácil calibración de sus parámetros y su mejor desempeño, comparado con otros métodos más complejos como Redes Neuronales. Además, son particularmente efectivos cuando se tiene una gran cantidad de variables (*features*) no relacionadas con el resultado (*settings with sparsity*).

- Wager & Athey (2017) derivan una variación de *Random Forest* que puede estimar parámetros con una distribución normal y una varianza finita, por lo que puede derivarse intervalos de confianza.
- Athey *et al.* (2016) derivan variaciones de *Random Forest* en los cuales se puede utilizar técnicas de GMM o de máxima verosimilitud en la estimación de los parámetros en cada rama (nodo).
- Una debilidad de *Random Forest* es que no son muy eficientes en capturar efectos lineales o cuadráticos o en explotar la información cuando esta no presenta cambios bruscos o muestra patrones suaves.