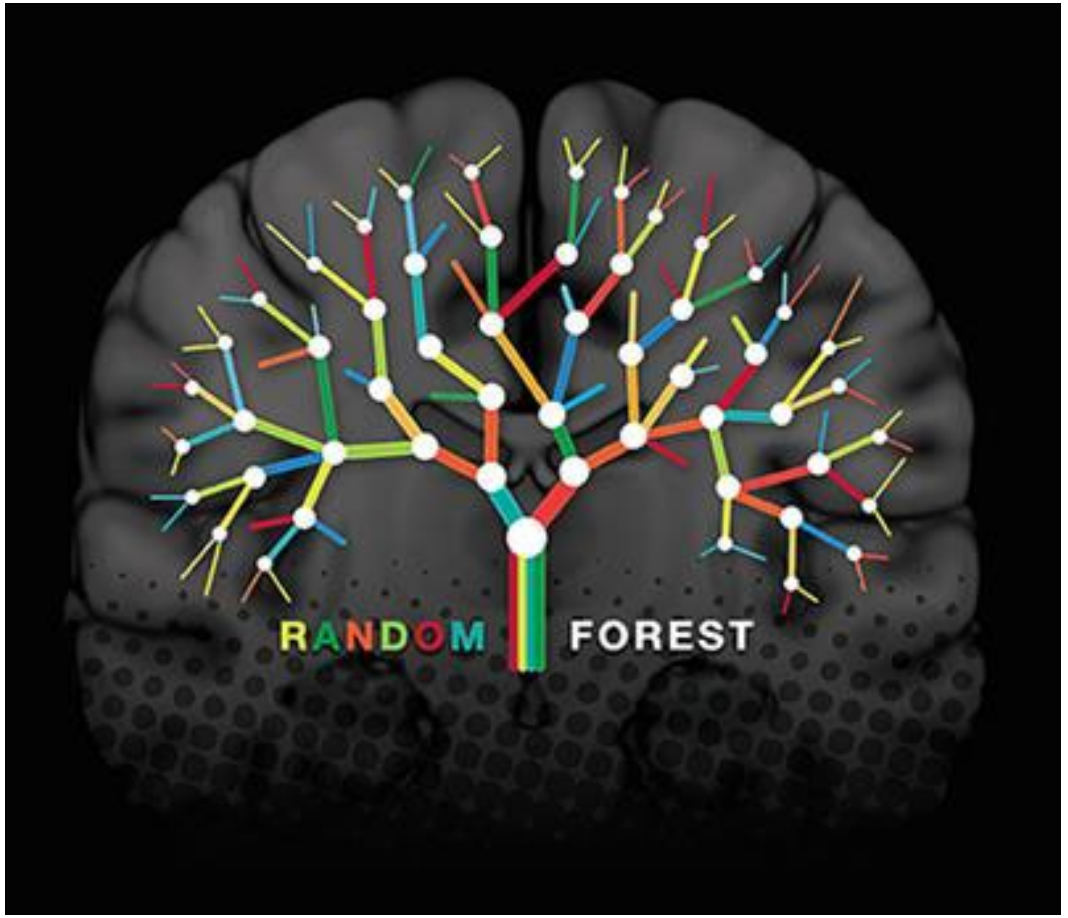


Bosques Aleatorios

Jonnatan Arias Garcia – Jonnatan.arias@utp.edu.co
David Cardenas peña - dcardenasp@utp.edu.co
Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Contenido

Introducción

Como surgieron?

Visión de los BA

Como se crea un BA

Muestro

Hiperparámetros

Ventajas – Desventajas

Bosques Aleatorios
Random Forest

Algoritmo de ml de uso común, registrado por Leo Breiman y Adele Cutle.

Es considerado un ensamblador, pues combina la salida de múltiples árboles de decisión para alcanzar un solo resultado.

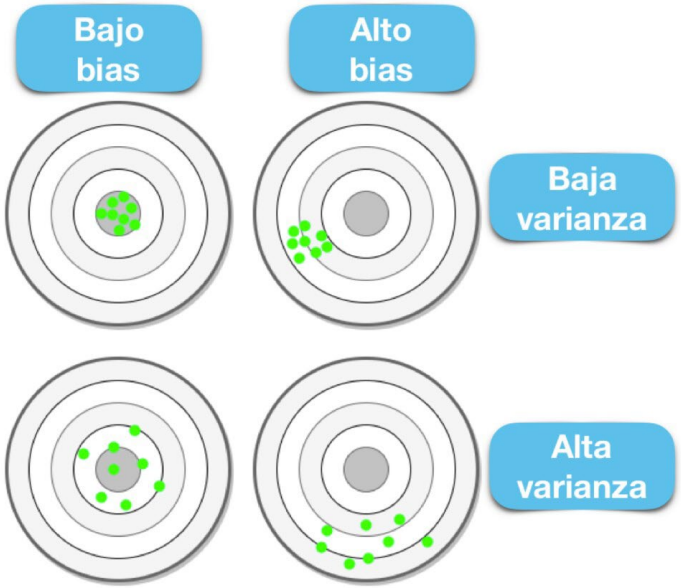
Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión.

A diferencia de los arboles de decisión que usan todos los datos, para cada árbol el bosque selecciona un subconjuntos, teniendo así una baja correlación entre arboles.

Bosques Aleatorios
Random Forest

¿Por qué surgen los BA?

- Un árbol suele sufrir de sesgo y varianza. Es decir, 'en promedio son los valores predichos diferentes de los valores reales' (sesgo) y 'cuan diferentes serán las predicciones de un modelo en un mismo punto si muestras' (varianza).
- Un árbol pequeño tiene baja varianza y alto sesgo. Normalmente, al incrementar la complejidad del modelo, se verá una reducción en el sesgo. Si el modelo es muy complejo se producirá sobre-ajuste empezando a sufrir de varianza alta.



Bosques Aleatorios
Random Forest

- El modelo óptimo debe mantener un balance entre estos dos tipos de errores. A esto se le conoce como "trade-off" (equilibrio) entre errores de sesgo y varianza.

El uso de ensambladores es una forma de aplicar este "trade-off".

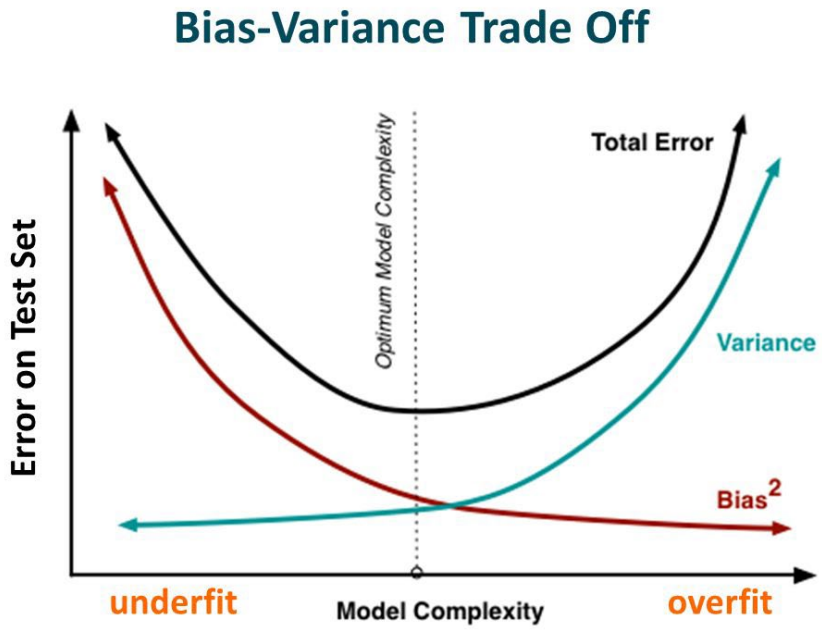


image credit: scott.fortmann-roe.com

Visión de los Bosques A.

- Útil en regresión y clasificación
- Un grupo de modelos “Débiles” se combinan en un modelo robusto
- Sirve como técnica de reducción de dimensionalidad
- Se generan múltiples árboles
- Cada árbol da una clasificación (vota por una clase) y el resultado es la clase con mayor numero de votos.
- Para regresión, se toma el promedio de las salidas de los arboles

Como se crea un Bosque A.

Cada Árbol se construye así:

1. Muestreo con reemplazo

- Dado un conjunto de entrenamiento con N casos, se toma una muestra aleatoria de estos casos con reemplazo. Esto significa que un mismo caso puede aparecer múltiples veces en la muestra.
- Esta muestra será utilizada como el conjunto de entrenamiento para construir un árbol en el bosque.

2. Selección aleatoria de características:

- Supongamos que tenemos M variables de entrada. Se elige un número m < M, que representa la cantidad de características que se considerarán en cada nodo al construir el árbol.
- En cada nodo del árbol, se seleccionan aleatoriamente m variables de las M disponibles.
- Luego, se evalúan todas estas variables para encontrar la mejor división posible en el nodo actual.

Como se crea un Bosque A.

3. Construcción de un árbol

- Una vez seleccionadas las variables, se utiliza un algoritmo de árbol de decisión (generalmente Gini) para dividir los datos en cada nodo.
- Este proceso continúa hasta que se alcanza una condición de parada, como una profundidad máxima o se usen todas las instancias (n variables)

4. No hay poda

- El árbol puede crecer hasta la máxima extensión posible

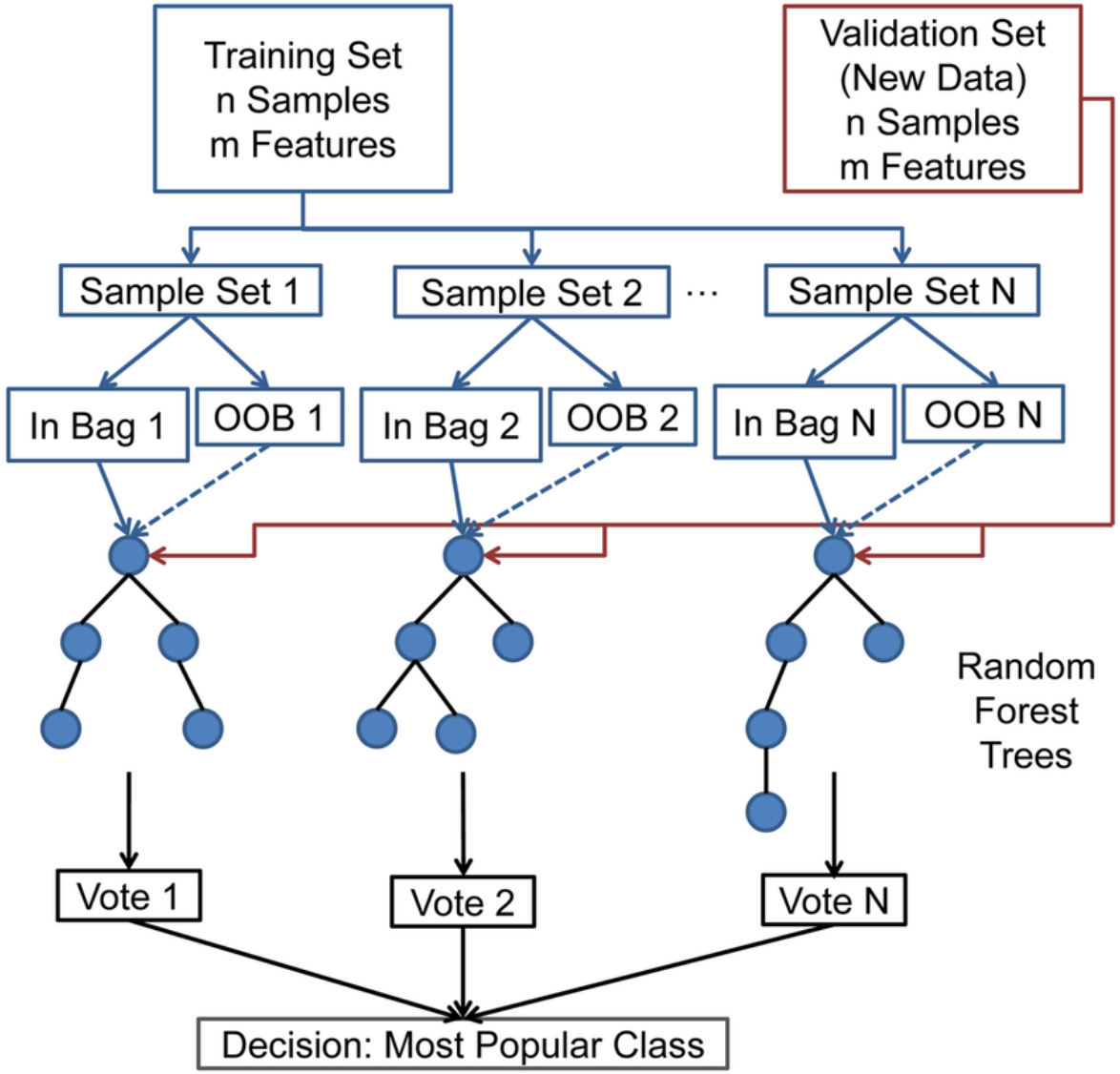
5. Predicción

- Para predecir una nueva instancia, se pasa a través de cada árbol individual.
- En clasificación, se realiza una votación mayoritaria entre todos los árboles para determinar la clase final. En regresión, se toma el promedio de las predicciones de todos los árboles como la predicción final.

Muestreo

- El muestreo con reemplazo se denomina Bootstrap
- Un tercio de los datos de entrenamiento pueden ser usados para test Este conjunto es denominado **out of bag (OOB) samples**
- El error estimado en OOB es conocido como OOB error
- OOB es tan preciso como usar datos Test de igual tamaño que entrenamiento.
- Seria posible no usar conjunto de test adicional

Random Forest



Hiperparámetros

El hiperparámetro mas importante seria el numero de variables candidatas para hacer cada ramificación, sin embargo, existen algunos adicionales.

- **ntree**: # de arboles
- **mtree**: # de variables aleatorias candidatas para ramificación
- **nodesize**: mínimo numero de muestras dentro de los nodos terminales
- **maxnodes**: máximo numero de nodos terminales
- **sampsize**: # de muestras sobre las cuales entrenar (por defecto 63.25%)
  - Valores menores bajos podría inducir sesgo pero reducir el tiempo
  - Valores altos podrían dar rendimiento pero llegar al overfitting
  - recomendado entre (60-80%)

Ventajas y Desventajas

Ventajas	Desventajas
Alta precisión	Falta de interpretabilidad
Robustes al sobreajuste	Tiempos de entrenamientos largos
Manejo automático de variables	Menos efectividad a datos muy dispersos
Eficiente para muchos datos	Sensible a parámetros como profundidad o numero de arboles.
Flexible al tipo de datos (categóricas o numéricas)	