SISTEMAS INTELIGENTES

Lecture 07:Regresión y Clasificación con Random Forests

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024

Cómo surge Random Forest?

- ✓Uno de los problemas que aparecía con la creación de un árbol de decisión es que si le damos la profundidad suficiente, el árbol tiende a “memorizar” las soluciones en vez de generalizar el aprendizaje. Es decir, a padecer de overfitting. La solución para evitar esto es la de crear muchos árboles y que trabajen en conjunto.

Ensembles

El término ensemble significa grupo. Los métodos tipo ensemble están formados de un grupo de modelos predictivos que permiten alcanzar una mejor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión.

Ensembles

- ✓ **Por qué surgen los ensembles de árboles?**
  - ➢ Así como todos los modelos, un árbol de decisión también sufre de los problemas de sesgo(bias) y varianza. Es decir, 'cuánto en promedio son los valores predecidos diferentes de los valores reales' (sesgo) y 'cuan diferentes serán las predicciones de un modelo en un mismo punto si muestras diferentes se tomaran de la misma población' (varianza).
  - ➢ Al construir un árbol pequeño se obtendrá un modelo con baja varianza y alto sesgo. Normalmente, al incrementar la complejidad del modelo, se verá una reducción en el error de predicción debido a un sesgo más bajo en el modelo. En un punto el modelo será muy complejo y se producirá un sobreajuste del modelo el cual empezará a sufrir de varianza alta.

Ensembles

✓Por qué surgen los ensembles de árboles?

- ➢El modelo óptimo debería 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 ensembles es una forma de aplicar este “trade-off”.

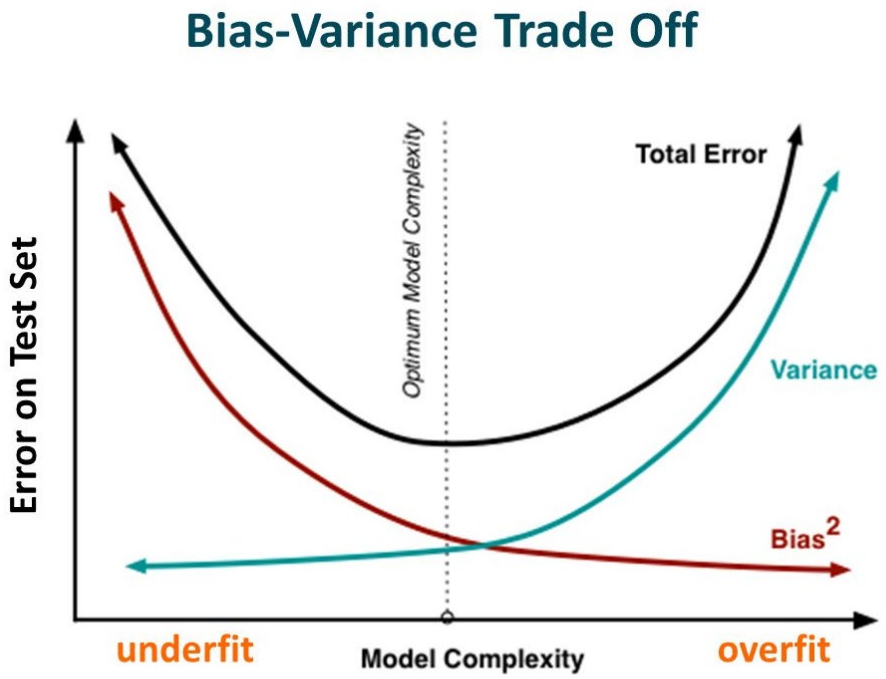


image credit: scott.fortmann-roe.com

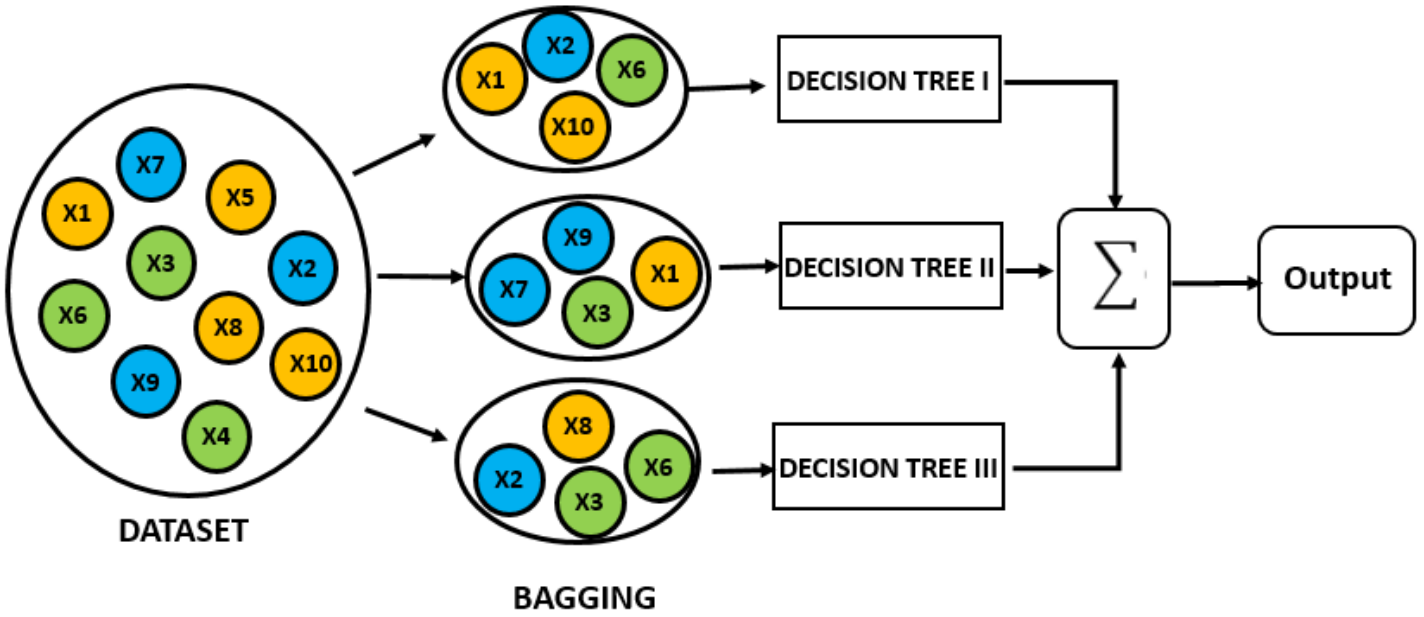
Ensembles comunes:

- Bagging
- Boosting
- Stacking.

Random Forest es del primer tipo.

Qué es el proceso de bagging y cómo funciona?

✓ **Bagging es una técnica usada para reducir la varianza de las predicciones a través de la combinación de los resultados de varios clasificadores, cada uno de ellos modelados con diferentes subconjuntos tomados de la misma población.**



En resumen:

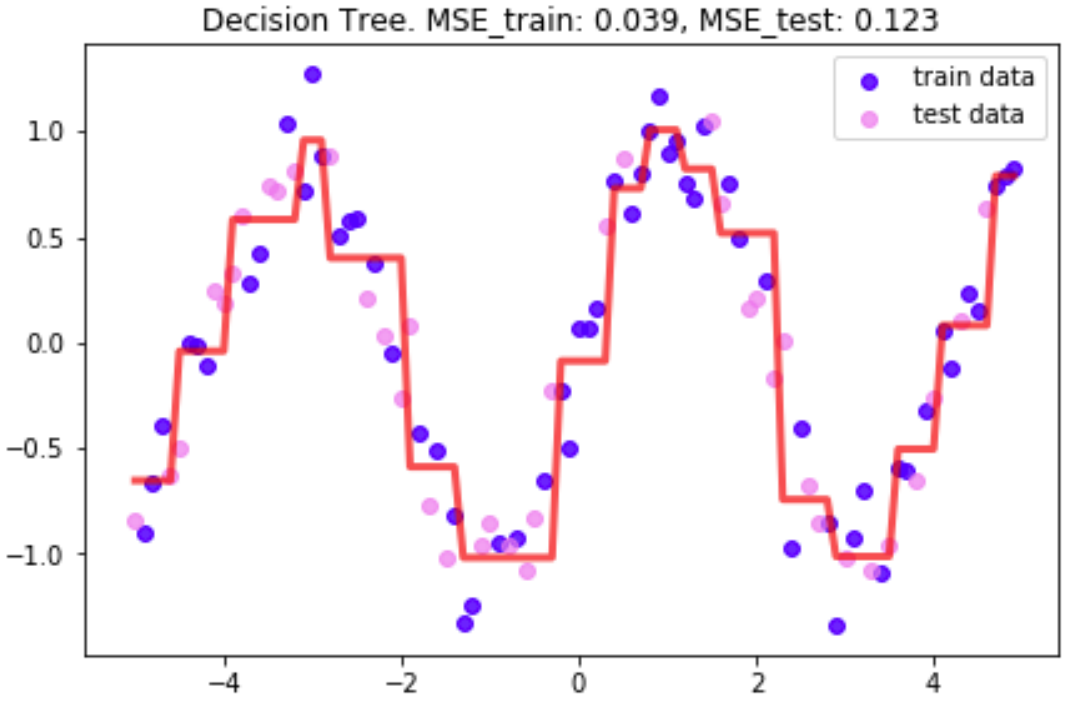
- Crear múltiples subconjuntos de datos
- Construir múltiples modelos
- Combinar los modelos.

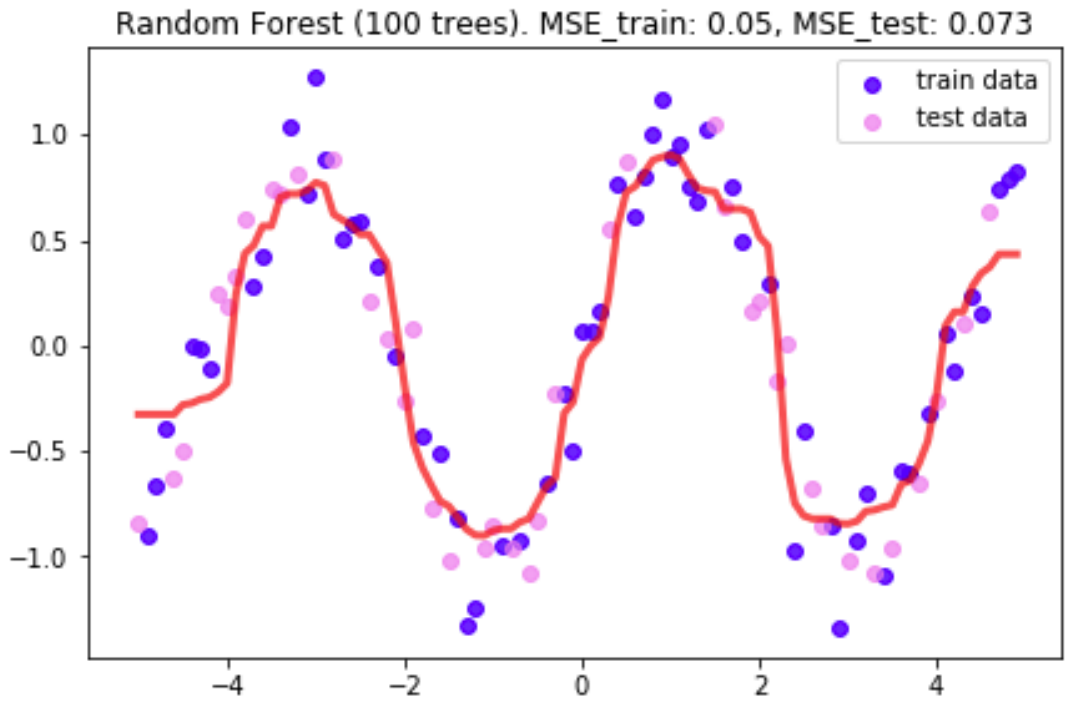
Random Forest

- ✓Random Forest es un algoritmo de aprendizaje basado en ensembles (conjunto) de decision trees, inventado por Leo Breiman y Adele Cutler
- ✓Combina los principios de bagging con selección de características aleatorias para añadir diversidad a los árboles de decisión
- ✓Consiste en una combinación de árboles de decisión entrenados de manera independiente y cuya salida se combina (mediante votación para clasificación o promedio para regresión) para producir una predicción más robusta y precisa, que será la respuesta del <<Bosque Aleatorio>>.
- ✓El modelo combina versatilidad y potencia en un enfoque
- ✓Como el ensemble solo utiliza una porción pequeña y aleatoria de todo el conjunto de características pueden trabajar con conjuntos de datos bastante grandes.

Random Forest

✓Random Forest son buenos para la generalización





- Random Forest proporciona un error de entrenamiento similar (same bias) que un árbol individual.
- Random Forest proporciona un error de prueba más bajo (lower variance)) como un árbol individual.

Decision Trees or Random Forest?

Decision Tree

- Estás interesado en un modelo de caja blanca
- La comprensión es más importante

Random Forest

- No te importa usar un modelo de caja negra
- Predecir (generalizar bien con nuevos datos) es más importante que entender

Tree + Random Forest

- Árbol de decisión para la comprensión
- Bosque aleatorio para predecir

Cómo se construye un modelo Random Forest?

Cada árbol se construye así:

1. Dado que el número de casos en el conjunto de entrenamiento es N. Una muestra de esos N casos se toma aleatoriamente pero CON REEMPLAZO. Esta muestra será el conjunto de entrenamiento para construir el árbol i.
2. Si existen M variables de entrada, un número m<M se especifica tal que para cada nodo, m variables se seleccionan aleatoriamente de M. La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.
3. Cada árbol crece hasta su máxima extensión posible y NO hay proceso de poda.
4. Nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación, promedio para regresión)

Por qué es aleatorio?

- ✓Contamos con una <<doble aleatoriedad>>: tanto en la selección del valor k de características para cada árbol como en la cantidad de muestras que usaremos para entrenar cada árbol creado.
- ✓Es curioso que para este algoritmo la aleatoriedad sea tan importante y de hecho es lo que lo “hace bueno”, pues le brinda flexibilidad suficiente como para poder obtener gran variedad de árboles y de muestras que en su conjunto aparentemente caótico, producen una salida concreta.
- ✓El modelo de random forests combina los principios de bagging con selección de variables aleatorias para añadir diversidad a los árboles de decisión. Una vez generado el ensemble de árboles (forest) el modelo utiliza el mecanismo de votación o la media para generar las predicciones.

Interpretación de out-of-bag error

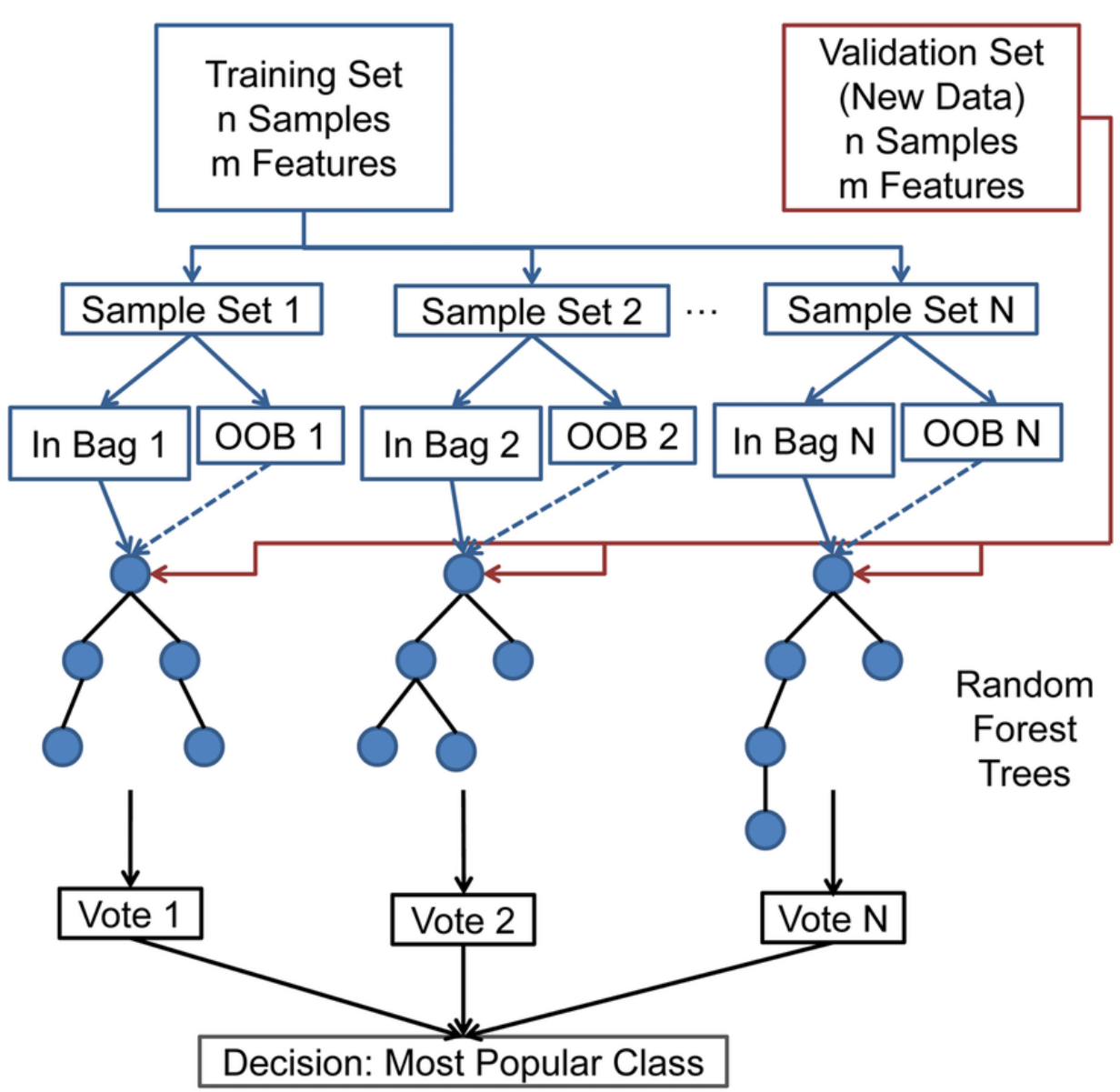
- ✓Una característica importante de los modelos Random Forests es el out-of-bag error. El out-of-bag error es una forma sencilla de estimar el error de test en un modelo bagged.
- ✓Se puede demostrar que de media cada árbol construido con un modelo bagged utiliza 2/3 de las observaciones del conjunto de entrenamiento. El 1/3 restante de las observaciones de entrenamiento no se utilizan para generar el árbol bagged y son llamadas out-of-bag.

Interpretación de out-of-bag error

- ✓El proceso de muestreo de los datos con reemplazo se denomina bootstrap.
- ✓Un tercio de los datos no se usan para el entrenamiento y pueden ser usados para test.
- ✓Este conjunto se denomina out of bag (OOB)samples.

Interpretación de out-of-bag error

- ✓El error estimado en estos out of bag samples se conoce como out of bag error (OOB error)
- ✓Usar este conjunto de test (OOB) es tan preciso como si se usara un conjunto de test del mismo tamaño que el de entrenamiento.
- ✓Sería posible no usar un conjunto de test adicional.



Hyperparámetros

- ✓ **El hyperparámetro más importante para ajustar es el número de variables candidatas a seleccionar para evaluar cada ramificación. Sin embargo, existen algunos adicionales que deben considerarse. Independientemente de los nombres en las distintas librerías, al menos los siguientes deberían estar presentes.**
  - ➢ **ntree: número de árboles en el bosque. Se quiere estabilizar el error, pero usar demasiados árboles puede ser innecesariamente ineficiente.**
  - ➢ **mtry: número de variables aleatorias como candidatas en cada ramificación.**
  - ➢ **sampsize: el número de muestras sobre las cuales entrenar. El valor por defecto es 63.25%.**
    - Valores más bajos podrían introducir sesgo y reducir el tiempo.
    - Valores más altos podrían incrementar el rendimiento del modelo pero a riesgo de causar overfitting. Generalmente se mantiene en el rango 60-80%.
  - ➢ **nodesize: mínimo número de muestras dentro de los nodos terminales. Equilibrio entre bias-varianza**
  - ➢ **maxnodes: máximo número de nodos terminales.**

Ventajas y Desventajas del uso de Random Forest

Ventajas

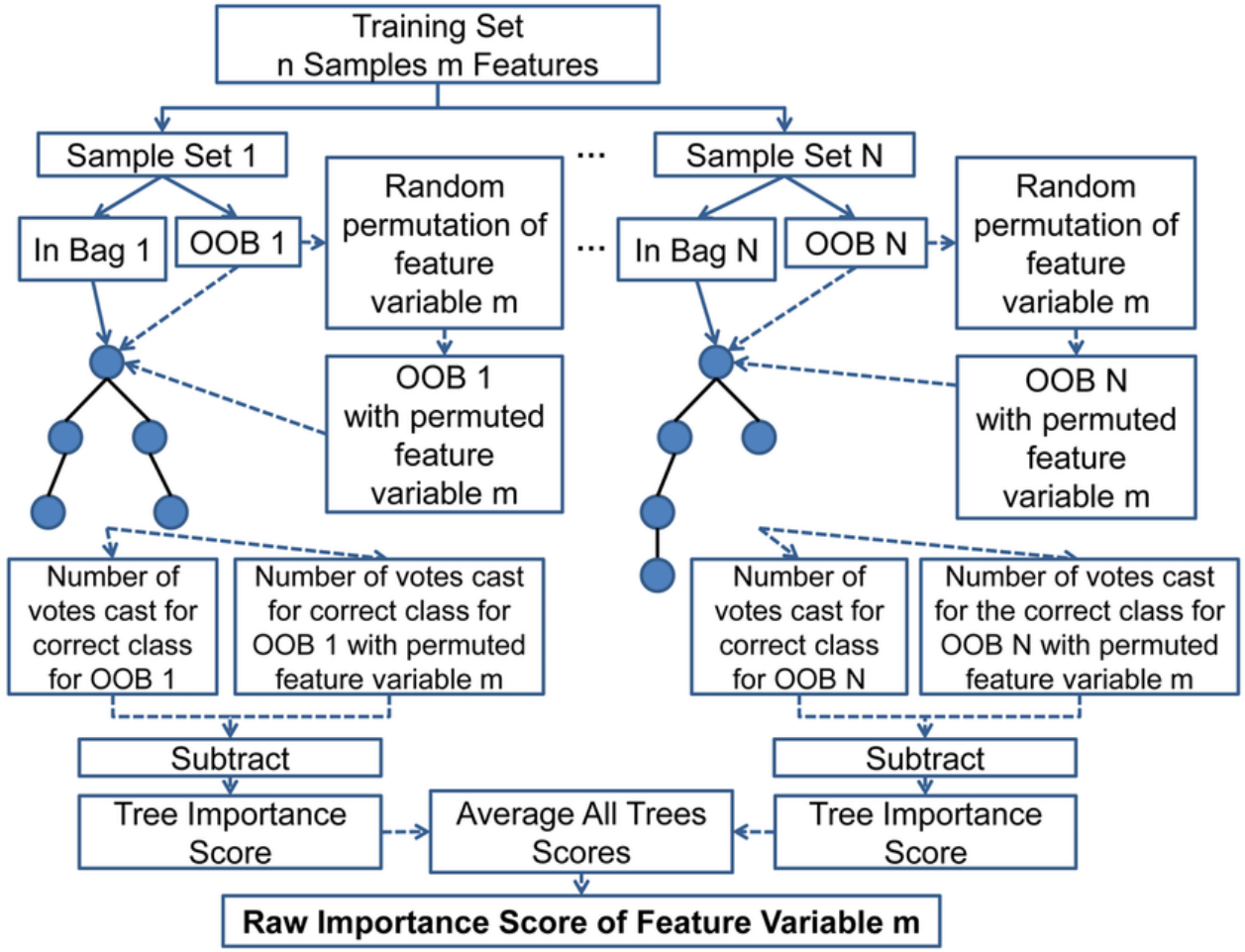
- ✓ Funciona bien -aún- sin ajuste de hiperparámetros
- ✓ Funciona bien para problemas de clasificación y también de regresión.
- ✓ Al utilizar múltiples árboles se reduce considerablemente el riesgo de overfiting
- ✓ Se mantiene estable con nuevas muestras puesto que al utilizar cientos de árboles sigue prevaleciendo el promedio de sus votaciones.
- ✓ Existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima.
- ✓ Puede manejar hasta miles de variables de entrada e identificar las más significativas. Método de reducción de dimensionalidad.
- ✓ Una de las salidas del modelo es la importancia de variables.
- ✓ Incorpora métodos efectivos para estimar valores faltantes.
- ✓ Es posible usarlo como método no supervisado (clustering) y detección de outliers.

Desventajas

- ✓ En algunos datos de entrada “particulares” random forest también puede caer en overfitting
- ✓ Es mucho más “costo” de crear y ejecutar que “un sólo árbol” de decisión.
- ✓ Puede requerir muchísimo tiempo de entrenamiento
- ✓ Random Forest no funciona bien con datasets pequeños.
- ✓ Es muy difícil poder interpretar los ¿cientos? de árboles creados en el bosque, si quisiéramos comprender y explicar a un cliente su comportamiento.
- ✓ Pérdida de interpretación
- ✓ Bueno para clasificación, no tanto para regresión. Las predicciones no son de naturaleza continua.
- ✓ En regresión, no puede predecir más allá del rango de valores del conjunto de entrenamiento.
- ✓ Poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos)

Importancia de atributos

- ✓ **¿Cómo se calcula?**
  - Para cada árbol en el bosque, calcular el número de votos por la clase correcta en la muestra (OOB)
  - Efectuar una permutación aleatoria de los valores de un predictor (e.g., variable-k) en la muestra (OOB) y verificar el número de votos por la clase correcta. Por permutuación aleatoria de una variable nos referimos a "barajar" (shuffling).
  - Sustraer el número de votos de la clase correcta en los datos de variable-k permutada, del número de votos por la clase correcta en la muestra OOB original.
  - El promedio de este número sobre todos los árboles en el bosque es el "score" de importancia sin normalizar. Este score es normalizado a partir de la desviación estándar.
  - Las variables que tiene valores altos para este score son clasificadas (ranked) como las más importantes.
  - Esto significa que, si el modelo fuera construido sin los valores originales de una variable específica, las predicciones serían peores. Por lo tanto, la variable es importante.
- ✓ **En conclusión: La importancia de un atributo es el incremento en el error del modelo de predicción luego de que el valor de dicho atributo ha sido permutado (se rompe la relación entre el atributo y la salida del modelo).**



PREGUNTAS??

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024