

Bosques Aleatorios (Random Forest)

Es un conjunto de árboles de clasificación (regresión) que se obtienen combinando la elección aleatoria de datos y la selección al azar de las variables predictoras.

- ▶ Generar muchos modelos diferentes usando conjuntos de entrenamiento diferentes.
- ▶ Elegir al azar y en cada nodo, un número pequeño de variables independientes.

Los bosques aleatorios sólo tienen interés si los árboles que se obtienen son muy diferentes entre sí.

Dada una muestra de N individuos con p variables independientes se construyen árboles de clasificación (regresión) de la siguiente forma:

1. Se elige al azar y con reemplazamiento una muestra de tamaño N (bootstrap). Al resto de individuos se les denomina OOB (Out Of Bag)
2. Se forma un árbol de clasificación con los siguientes criterios:
 - 2.1 Para dividir cada nodo se elige al azar un número pequeño de variables independientes (en general $\sqrt{p}$)
 - 2.2 Se hace una división binaria del conjunto de entrenamiento con la variable que mejor predice a la variable dependiente 'y'
 - 2.3 Cada árbol se desarrolle tanto como sea posible sin realizar ninguna poda
 - 2.4 Los nodos finales deben contener muy pocos individuos

Una vez contruidos los árboles de regresión se realiza la predicción combinando los árboles obtenidos de acuerdo a la naturaleza de la variable dependiente:

1. Problemas de clasificación

- ▶ Calcular la proporción de veces que un individuo del grupo OOB es asignado a cada clase.
- ▶ La clase predicha es la que tiene una proporción mayor.

2. Problemas de regresión

Promediando en cada individuo de OOB los valores obtenidos en los diferentes árboles.

Bosques Aleatorios. Importancia de las variables

La calidad de la predicción para un individuo i se mide como

$$\text{Margin}(i) = \hat{P}(i \in \text{Clase real}) - \text{Max}(\hat{P}(i \in \text{Otras clases}))$$

La importancia de la variable x_m se calcula de la siguiente forma:

1. Usar la muestra OOB para el árbol k
2. Calcular el 'Margin', M_0 para la muestra OOB
3. Permutar los valores de la muestra para x_m
4. Aplicar el árbol k a OOB con los valores permutados
5. Calcular la nueva 'Margin', M con x_m permutada
6. Calcular la diferencia $M_0 - M$

$$I(x_m) = \overline{M_0 - M}$$

Bosques Aleatorios. Proximidad entre los individuos

Los bosques aleatorios se pueden emplear para medir la 'proximidad' entre los individuos

- ▶ Se aplica cada árbol a todos los individuos de la muestra y se determina en que nodo final se encuentran
- ▶ La proximidad entre los individuos i, j se mide como la proporción de veces que se encuentra en el mismo nodo final

Esta medida es invariante frente a transformaciones monótonas

Un individuo se etiqueta como 'outlier' cuando está alejado de todos los que pertenecen a su mismo grupo o clase.

Para detectarlos se puede emplear el siguiente algoritmo

- ▶ Para el caso i de la clase C , se calcula la suma de los cuadrados de su similitud con los sujetos su clase

$$\sum_{j \in C} prox(i, j)^2$$

- ▶ Se invierte ese valor y se tipifica usando la mediana y la desviación típica
- ▶ En general un valor superior a 10 indica un comportamiento 'anómalo'

Bosques Aleatorios. Imputación de valores perdidos

Para imputar los valores perdidos se suelen utilizar-combinar dos procedimientos

1. Se les asigna la mediana (variables cuantitativas) o la moda (variables cualitativas)
2. Se utiliza el siguiente algoritmo:
 - ▶ Se asigna la mediana o la moda
 - ▶ Se aplica el Random Forest para predecir esos valores
 - ▶ Se replica el paso anterior dos o tres veces

Parámetros del Random Forest:

- ▶ Número de predictores en cada nodo $mtry$:
 - ▶ Regresión: $mtry = \sqrt{p}$,
 - ▶ Clasificación: $mtry = \frac{p}{3}$, para la
 - ▶ Probar con los valores recomendados por defecto, la mitad y el doble. Quedarse con el mejor.
- ▶ Número de árboles B (muestras bootstrap) a construir
 - ▶ Usualmente $B \simeq 500$.
 - ▶ Construir árboles hasta que el error no decrezca

Bosques Aleatorios. Ventajas

Las principales ventajas de los bosques aleatorios son:

- ▶ Es muy sencillo de usar.
- ▶ El único parámetro es el número de variables que intervienen en la división de cada nodo
- ▶ En general no tiene problemas de sobreajuste
- ▶ Suele conseguir menor tasa de error que los árboles de clasificación
- ▶ Tienen poco coste computacional y se puede emplear con grandes bases de datos

Bosques Aleatorios. Ventajas

- ▶ Permite medir la importancia que tiene cada variable en la predicción
- ▶ ¿Tiene un método eficaz para estimar datos perdidos y mantener la información cuando hay muchas observaciones perdidas?
- ▶ Permite calcular una distancia entre las observaciones y detectar valores atípicos.
- ▶ ¿Ofrece un método experimental para detectar las interacciones de las variables?

Bosques Aleatorios. Inconvenientes

Entre sus puntos débiles se pueden citar los siguientes aspectos:

- ▶ A veces sobreajusta las predicciones en problemas con mucho 'ruido'.
- ▶ No es fácil determinar cómo se realizan las predicciones (es una caja negra)
- ▶ Tiende a sobrevalorar la importancia de las variables categóricas con muchos niveles de respuesta.
- ▶ Si los datos contienen grupos de variables correlacionadas y con una importancia similar respecto a la variable objetivo, los grupos con frecuencias menores suelen obtener mejores resultados que los más grandes.

Bosques Aleatorios. Código R

<https://www.statmethods.net/advstats/cart.html>

Ejemplo de análisis con Bosques Aleatorios

```
# Random Forest prediction of Kyphosis data
```

```
library(randomForest)
```

```
fit<- randomForest(Kyphosis ~ Age+Number+Start, data=kyphosis)
```

```
print(fit) ## view results
```

```
importance(fit) ## importance of each predictor
```