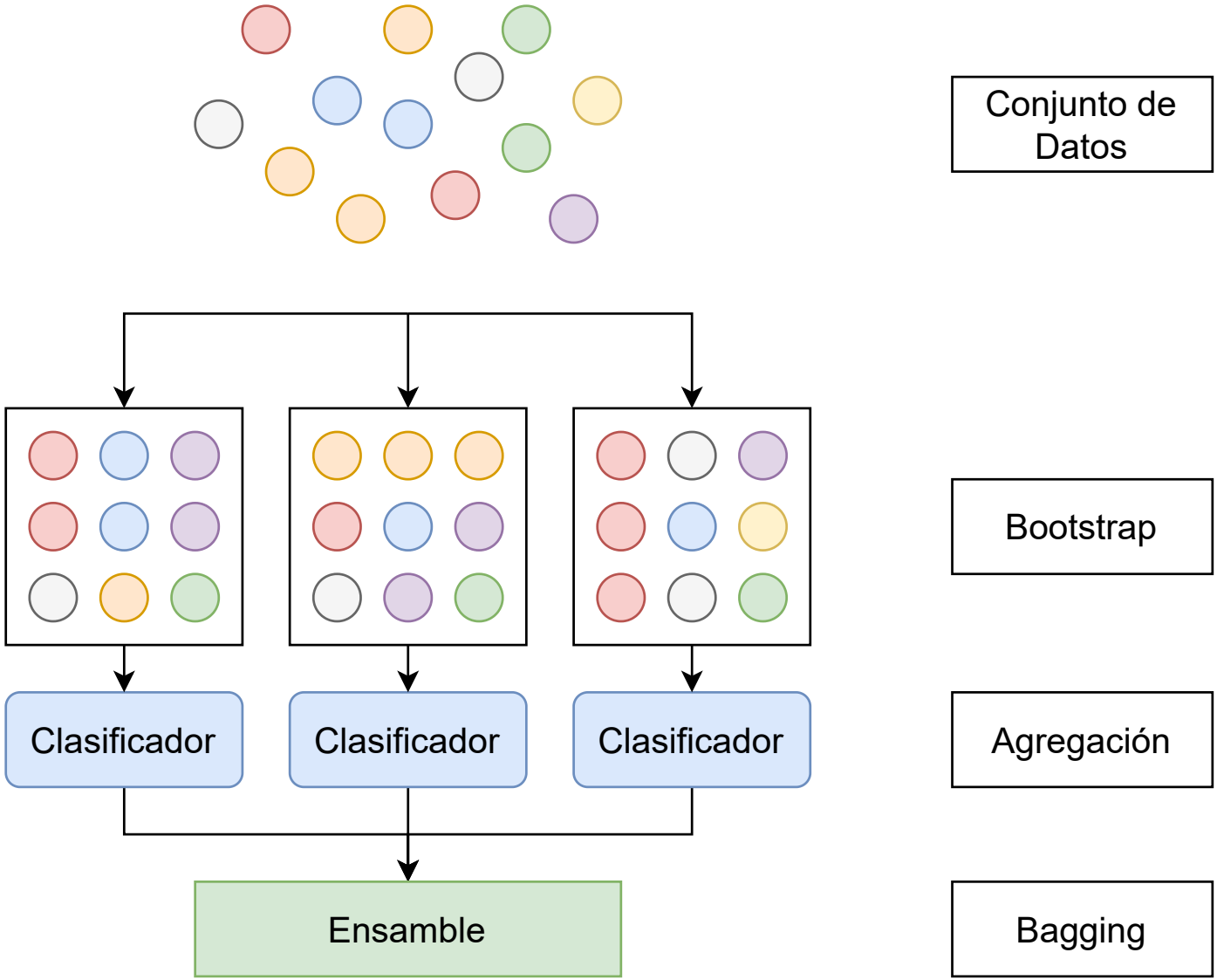
Ensemble Learning

Bosques Aleatorios

Luis Norberto Zúñiga Morales

8 de marzo de 2022

En clases anteriores...



En clases anteriores...

¿Quién ofrece un resumen del algoritmo de Bagging?

Random Forests

- Durante el proceso del Bagging se construyen árboles similares con conjuntos de datos.

Random Forests

- Durante el proceso del Bagging se construyen árboles similares con conjuntos de datos.
- Los datos se obtienen de forma aleatoria mediante muestreo con reemplazo.

Random Forests

- Durante el proceso del Bagging se construyen árboles similares con conjuntos de datos.
- Los datos se obtienen de forma aleatoria mediante muestreo con reemplazo.
- De esta forma, se reduce la varianza de cada modelo base.

Random Forests

- Durante el proceso del Bagging se construyen árboles similares con conjuntos de datos.
- Los datos se obtienen de forma aleatoria mediante muestreo con reemplazo.
- De esta forma, se reduce la varianza de cada modelo base.
- ¿Notan algún problema con este procedimiento?

Random Forests

- Al momento de generar los árboles, estos son idénticamente distribuidos.

Random Forests

- Al momento de generar los árboles, estos son idénticamente distribuidos.
- ¿Qué implica?

Random Forests

- Al momento de generar los árboles, estos son idénticamente distribuidos.
- ¿Qué implica?
  - El sesgo es igual en todo el ensamble.

Random Forests

- Al momento de generar los árboles, estos son idénticamente distribuidos.
- ¿Qué implica?
  - El sesgo es igual en todo el ensamble.
  - Sólo se puede reducir la varianza.

Random Forests

- Un promedio de B v.a.i.i.d, cada una con varianza σ^2 , tiene varianza $\frac{1}{B}\sigma^2$.

Random Forests

- Un promedio de B v.a.i.i.d, cada una con varianza σ^2 , tiene varianza $\frac{1}{B}\sigma^2$.
- ¿Qué pasa si no son independientes?

Random Forests

- Un promedio de B v.a.i.i.d, cada una con varianza σ^2 , tiene varianza $\frac{1}{B}\sigma^2$.
- ¿Qué pasa si no son independientes?
- ¡Existe correlación!

Random Forests

- Un promedio de B v.a.i.i.d, cada una con varianza σ^2 , tiene varianza $\frac{1}{B}\sigma^2$.
- ¿Qué pasa si no son independientes?
- ¡Existe correlación!
- Si las variables son i.d. con correlación positiva ρ, la varianza del promedio esta dada por

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{1}$$

Random Forests

Varianza con correlación

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2}$$

Conforme B aumenta, el segundo término de la expresión va a cero, por lo que la correlación entre pares de árboles en el bagging afecta el beneficio del premedio, i.e., no reduce la varianza .

Random Forests

- La idea de los Bosques Aleatorios yace en mejorar la reducción de la varianza propuesta por bagging al reducir la correlación entre los árboles.

Random Forests

- La idea de los Bosques Aleatorios yace en mejorar la reducción de la varianza propuesta por bagging al reducir la correlación entre los árboles.
- Lo anterior se logra al momento de crecer los árboles eligiendo de forma aleatoria los valores del vector de entrada.

Random Forests

- La idea de los Bosques Aleatorios yace en mejorar la reducción de la varianza propuesta por bagging al reducir la correlación entre los árboles.
- Lo anterior se logra al momento de crecer los árboles eligiendo de forma aleatoria los valores del vector de entrada.
- Después de realizar el Bootstrap, se seleccionan $m \leq p$ del total de las variables de entrada al azar como candidatos para el corte de los árboles. Algunos valores populares para m son $\sqrt{p}$ o 1.

Random Forests

Después de que B árboles $\{T(\boldsymbol{x}, \blacksquare_b)\}_{b=1}^{B}$ son creados, el bosque aleatorio para el caso de regresión se encuentra dado por

$$\hat{f}^B(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} T(\boldsymbol{x}, \blacksquare_b) \tag{3}$$

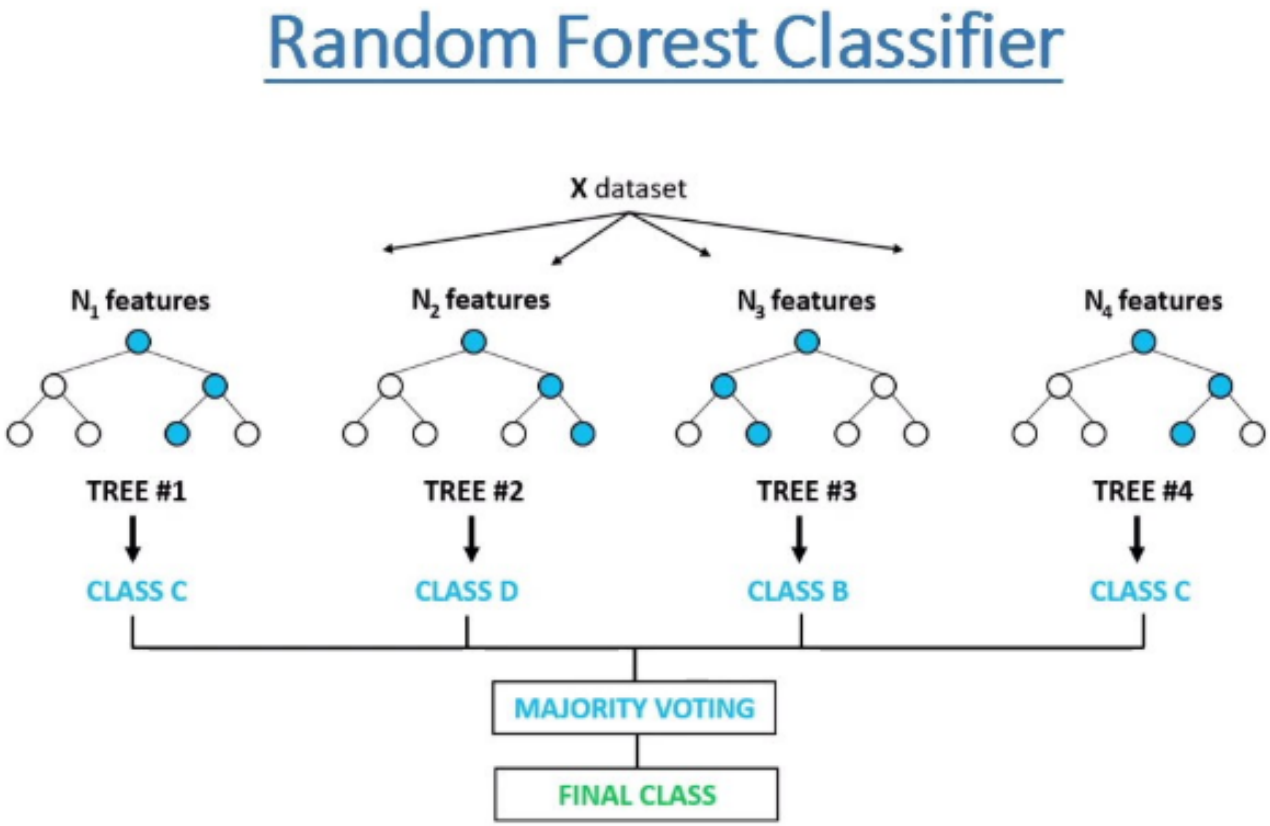
donde $\blacksquare_b$ caracteriza el b-ésimo árbol del bosque aleatorio: valores de corte, cortes en cada nodo y los valores de los nodos terminales.

Random Forests

Para $b = 1$ hasta B:

1. Formar una muestra aleatoria Z con reemplazo (Boostrap) de tamaño N del conjunto de datos.
2. Construir un árbol aleatorio T_b utilizando el conjunto Z de la siguiente manera:
    1. Seleccionar m de las p variables de entrada de cada dato.
    2. Elegir el mejor punto de corte de las variables entre las m.
    3. Dividir el nodo en dos, creando dos ramas.
3. Para predecir el valor de un punto x: realizar un promedio de los resultados para el caso de regresión, y un voto mayoritario para el caso de clasificación.

Random Forests



Bibliografía