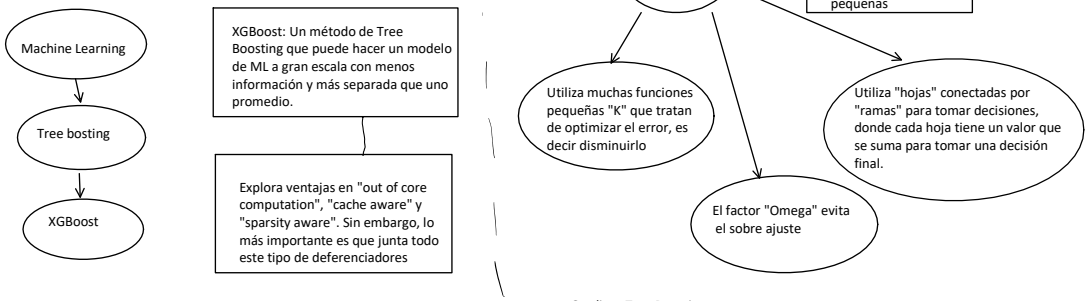
XGBoost

Friday, May 27, 2022 12:45 PM

Resumen: El algoritmo XGBoost es un algoritmo que apunta a optimizar el modelo **Tree Boosting** al punto máximo, a la vez utilizando todos los componentes del dispositivo que lo corre. Esto lo hace a través de múltiples ventajas como la reunión de datos en bloques para optimizar el **parallel learning** y la compresion de los bloques para correr el algoritmo **out of core**. Igualmente mientras se corre, este código optimiza su funcionamiento al ser consciente de cuando los datos no muestran correlación para encontrar una solución fácil, usar soluciones distintas cuando trabaja con valores que se benefician de **aproximate local** o **aproximate global** y trabajar con **exact greedy**.



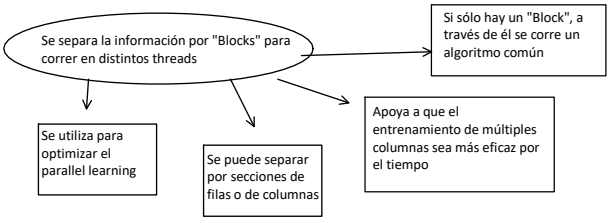
Gradient Tree Boosting

- Se va agregando una función ft(xi), por cada instancia en una iteración
- Con esto se calcula el peso "wj" ideal.
- Para el siguiente para de ramas del árbol se crea un "split" y se calcula el error para este split (para de esta forma calcular el split adecuado)
- Se utiliza el shrinkage para que decisiones viejas no afecten el futuro del árbol

Split Finding Algorithm

1. Organizar Valores (optimizar)
   a. Poner los valores por serie de magnitud para hacerlos más razonables de procesar.

2. Hacer "exact greedy algorithm"
   a. Es un proceso intensivo
   b. Asigna posibles candidatos en relación al "parallel learning"
   c. Revisa todos los posibles splits
   d. Empieza a hacer variantes ya sean locales o globales
      i. El global ayuda a cuando se tiene muchos candidatos
      ii. El local es más preciso pero más tardado
   e. Se integra un "weight" algorythm que utiliza mezcla y corte de operaciones
   f. Se hace consciente el algoritmo de la separación de los datos. En los casos en que los datos no expresen suficiente información, el algoritmo agarra un camino por default
      i. Esto hace considerablemente más rápido el algoritmo



*Al hacer el algoritmo "Cache Aware" y mezclarlo con la división apropiada de los "Block", se puede ahorrar una cantidad impresionante de tiempo.

Diferenciador:

System	exact greedy	approximate global	approximate local	out-of-core	sparsity aware	parallel
XGBoost	yes	yes	yes	yes	yes	yes
pGBRT	no	no	yes	no	no	yes
Spark MLLib	no	yes	no	no	partially	yes
H2O	no	yes	no	no	partially	yes
scikit-learn	yes	no	no	no	no	no
R GBM	yes	no	no	no	partially	no