Machine Learning aplicado a las Ciencias Sociales

Clase 7. Análisis supervisado. Ensamble Learning - Bagging / Random Forest

factor~data
EIDAES_UNSAM

Random Forest

- Random forest es muy similar a un bagging de árboles de decisión
- La diferencia: además de generar variabilidad sobre los registros, se genera variabilidad sobre los predictores.
- Bagging genera B predicciones a partir de B remuestras bootstrap del TrS original y de M predictores del TrS original
- De esta forma, en bagging entran el total de los M predictores.

Random Forest

- Esto puede generar árboles muy correlacionados… ¿por qué?

- Si una o algunas variables son predictores muy fuertes para la variable target, estas variables serán seleccionadas en muchos de los árboles base del bagging, haciendo que queden correlacionados.

- Seleccionando un subconjunto aleatorio de las variables en cada división, contrarrestamos esta correlación entre los árboles base, fortaleciendo el modelo final.

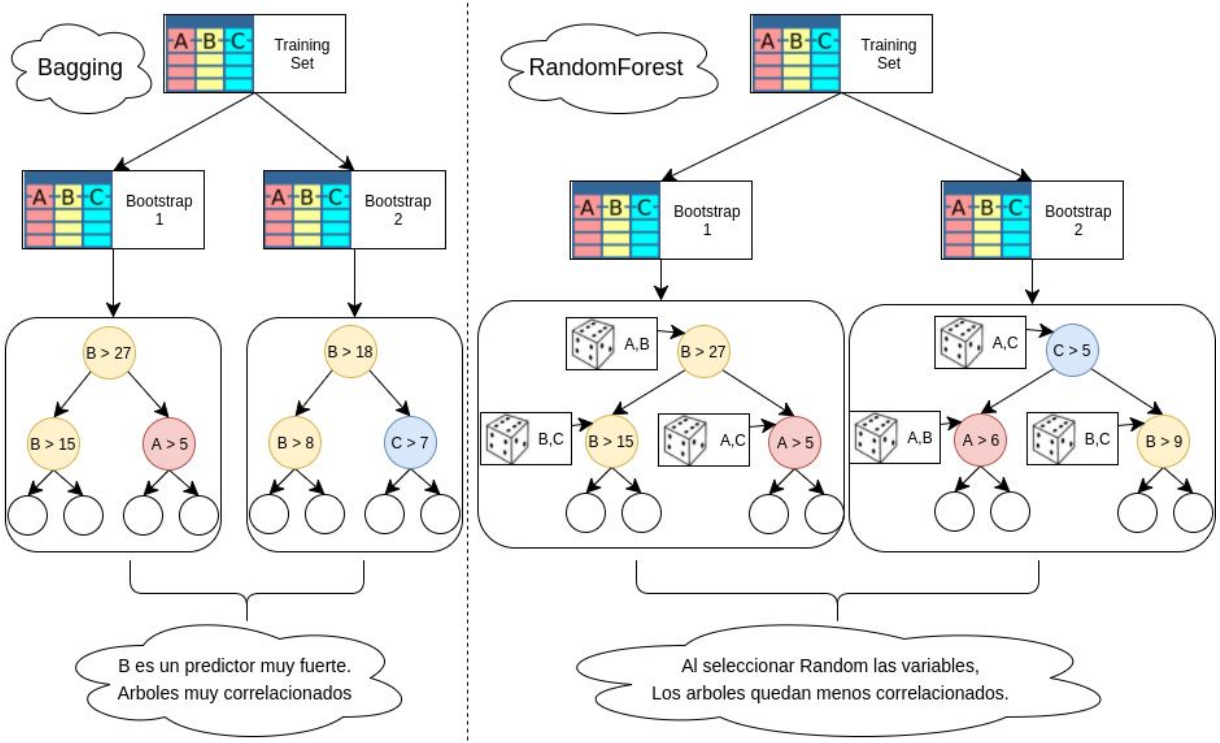
Random Forest

- Para un problema de clasificación con p variables, se suelen utilizar $\sqrt{p}$ de las variables en cada división.

- Para problemas de regresión, recomiendan utilizar p/3.

- Pero también podría considerarse como un hiperparámetro para tunear.

Random Forest



Random Forest

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

Síntesis

- Ensambles: herramientas potentes
- Uso de la aleatoriedad para incrementar la capacidad del modelo
- Bagging = Bootstrap Aggregating
- Random Forest = Bagging + random selection de features
- Extra Randomized Trees: Random Forest + random splits