Tarea E8

Alexander Zapata Galindo – Código 201425426

E8 - Ensemble Trees Overview

Write at least 300 words explaining why ensemble is a successful strategy in machine learning.You can send your answer through this form: https://forms.gle/p2xbPohphBGXo56F9. Due date: 15/07/2021 5:59 pm.

El ensamble de modelos o modelos combinados (ensamblaje) es una técnica que se utiliza en “machine learning” que busca mejorar el rendimiento de los modelos, así como su precisión y su poder predictivo.

En este sentido, en lugar de hacer un único modelo y esperar a que este sea el mejor que podamos encontrar, la idea básica es construir estratégicamente una cierta cantidad de modelos y luego promediar o combinar los resultados de estos para producir un modelo final que permita resolver un problema en particular.

Los modelos a combinar pueden ser disímiles entre sí: a) pueden haber utilizado muestras de diferentes poblaciones de datos, b) pueden haber sido modelados con técnicas distintas, y c) pueden estar intentando probar hipótesis diversas.

Generalmente, el ensamblaje funciona combinando muchos predictores que son sub-óptimos, de forma tal que el método de combinación corrija las razones del “under-performance”.

Si bien existe varios tipos de métodos de ensamblaje, los más utilizados en la industria son i) Stacking, ii) Agregación Bootstrap o Bagging, y iii) Boosting.

Staking es la idea básica de combinar predictores buenos. En general, este ensamble es una combinación lineal de los predictores originales, donde el método busca los pesos o ponderaciones óptimas.

Bagging es una técnica que toma un conjunto de datos del cual se extraen varias muestras seleccionadas de manera aleatoria. Posteriormente, se entrenan modelos de manera separada para cada una de los subconjuntos anteriormente definidos. El modelo final es producto de la combinación de las proyecciones de todos los sub-modelos. Esta técnica funciona mejor cuando cada modelo tiene mucha varianza y poco sesgo.

Boosting funciona entrenando un modelo con todo el conjunto de entrenamiento, y los modelos posteriores se construyen ajustando los valores de los errores residuales del modelo inicial. Una vez que se crea la secuencia de modelos, las predicciones hechas por los modelos son ponderadas por sus puntuaciones de precisión. Las estimaciones del modelo final se construyen en un proceso de ponderación, dando mayor peso a aquellas observaciones que el modelo anterior no estimó muy bien. Esta técnica funciona mejor en modelos que tiene mucha varianza y poco sesgo.