Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Segundo Semestre de 2024

IIC 2433 – Minería de datos
Tarea 2 - Boosting

1. Enunciado

La meta de esta tarea es intentar subir el 0.48 de accuracy que nos dio el método de árboles de decision para el caso de los helados. Como ya vimos que con ensamblajes ya no funcionó, vamos a usar boosting.

1.1. Información importante

Esta tarea es individual, se entrega como un notebook con todas las celdas corridas.
Fechas. La fecha de entrega de la tarea es el Viernes 13 de Septiembre, a las 20:00 hrs.

1.2. Primer boosting

1. Primero, del dataset de helados genera un train/test split con un 30 % de datos de test y un 70 % en el entrenamiento.

2. Entrena un clasificador de árbol de decisión sobre los datos de entrenamiento, como lo vimos en clases, llamémoslo **clf**. Comprueba que el accuracy para los datos de test sigue siendo 0.48 o menor.

3. Recuerda que podemos usar `clf.predict_proba` para obtener las probabilidades de clasificar cada instancia. Define entonces el error e como $e[i] = y[i] - \text{proba}[i]$, donde y es el vector que contiene las clases correctas de los datos de entrenamiento, y proba$[i]$ es la segunda componente del i-esimo elemento del vector que resulta al llamar a `clf.predict_proba` con los datos de entrenamiento.

4. Entrena un árbol de regresión usando la clase `DecisionTreeRegressor` de `sklearn.tree`, de forma que pueda predecir el vector e a partir de los datos de entrenamiento. Llamemoslo **clf_boost**

5. Define un nuevo clasificador que clasifique a una entrada como un 1 si el promedio entre el resultado de las probabilidades de **clf** (usando predict_proba) y clf_boost es mayor o igual a 0,5, y como 0 en otro caso.

6. Prueba el accuracy de este clasificador al usar el set de test. ¿Es mejor?

1.3. Segundo boosting

Vuelve a entrenar ahora un tercer clasificador, `DecisionTreeRegressor`, ahora sobre el error definido como la diferencia entre e y el resultado de **clf_boost**, tal y como lo hiciste en la sección anterior. Verifica si logras subir o no el accuracy más allá de lo logrado en la sección anterior para alguna combinación lineal entre el resultado de clf (probabilidades), clf boost y este tercer clasificador.

1.4. ¡Cuidado con las métricas!

Mira ahora el score f1 de lo que has hecho. Examina un poco mejor para ver por qué (en este caso) el boosting resultó en un mal clasificador.