Búsqueda de hiperparámetros

Inicialmente se realiza una división de los datos con el objetivo de particionar la base en train-test y validación, con un porcentaje de 70/30. La metodología para implementar es realizar una validación cruzada que permita encontrar un error preliminar de prueba que se pueda comparar con el error de validación y tener una idea acerca del error de forma generalizada sobre el conjunto completo de los datos.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Conociendo la naturaleza del problema, como una clasificación binaria, se decide implementar una regresión logística para determinar la probabilidad de un cliente pertenecer a una clase u otra, considerando un umbral de decisión (por defecto del 0.5).

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ahora, la problemática que se presenta al momento de entrenar este modelo es ¿Cómo encontrar el conjunto de hiperparámetros que optimiza los resultados?, si bien existen muchas formas de encontrar la mejor combinación de hiperparámetros (desde “fuerza bruta” a optimizaciones bayesianas) en este ejercicio se presenta la implementación de una búsqueda a través de una malla de hiperparámetros que varia o “muta” a partir de sus mejores individuos (conjuntos de hiperparámetros) teniendo como métrica de referencia, en este caso, la precisión (la cual observamos que tiene un desbalanceo de clases en una proporción del 90%-10%). A continuación se procede a explicar el funcionamiento del concepto de algoritmo evolutivo o genético sobre la tarea de aprendizaje automático implementada en este problema.

Texto

Descripción generada automáticamente

Iniciamos preliminarmente con la definición del modelo a implementar y un espacio de valores en los cuales nuestros parámetros se encontrarán, particularmente la regresión logística entre varios parámetros ajustables cuenta con los parámetros de tol, solver y max\_iter, los cuales nos permiten tener una tolerancia como criterio de parada, varias opciones de algoritmos de resolución de problema o el número máximo de iteraciones en las que el algoritmo realiza una ejecución. Este espacio de parámetros tiene sus propias particularidades y deben ser definidas a través de su naturaleza, es decir, continuas, discretas o categóricas. A adicionalmente tenemos la definición del numero de folds a utilizar (5) para realizar la validación cruzada (la cual nos permite tener un poco de independencia y tranquilidad con respecto a la dependencia del modelo respecto al conjunto de train-test y validación).

