Integrantes:

Sebastián Romero

Juan Calderón

Kristian Mendoza

Fecha: 12/03/24

Data Mining

Proyecto 5

Accuracy

$$Accuracy = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$$

Es el estándar en medidas globales de desempeño de un modelo. Se mide como la relación entre el total de los valores predichos correctamente con el total de elementos evaluados, donde 0 y 1 son los casos de un modelo que acierta o falla al predecir todos los elementos del conjunto de test, y 0.5 el caso de un modelo que predice al azar.

Esta métrica es poco confiable ante datasets con un desbalance de clases, ya que el desempeño del modelo al predecir la clase con menor presencia en el dataset será opacado por el desempeño de las demás clases.

Además, al ser una métrica global, el accuracy no proporciona información detallada sobre el desempeño del modelo en cuanto a falsos positivos y falsos negativos. En esta métrica, ambos tipos de errores se consideran igualmente, lo que significa que no se hace distinción entre ellos en términos de predicciones incorrectas.

Recall

$$Recall = \frac{VerdaderosPositivos(TP)}{VerdaderosPositivos(TP) + FalsosNegativos(FN)}$$

Es la relación entre los true positives y el total de true values del conjunto de test, donde 1 y 0 son los casos de un modelo que acierta o falla al predecir todos los positivos del conjunto de test.

A diferencia del accuracy, no ofrece información sobre el desempeño del modelo con los casos negativos. Sin embargo, resulta útil en contextos donde la detección de la mayoría de los positivos en el conjunto de datos es el objetivo principal del modelo. Además, a diferencia del accuracy, esta métrica no se ve afectada por el desbalance de clases en el dataset.

La métrica que utiliza los true negatives con el total de false values es llamado inverse recall.

Precision

$$\text{Pr}\, e\, cisión = \frac{VerdaderosPositivos(TP)}{VerdaderosPositivos(TP) + FalsosPositivos(FP)}$$

Muestra la relación entre los trues positives y los trues predichos por el modelo, donde 1 y 0 son los casos de un modelo que acierta o falla en todos los positivos predichos.

Esto hace que esta métrica no de ninguna información sobre como el modelo se desempeña con los casos negativos, sin embargo, es útil cuando se busca que los positives predichos por el modelo sean en realidad positives, es decir cuando fallar al predecir correctamente un positive es menos deseado que dejar pasar un positive. Además, esta métrica no se ve afectada por un desbalance de clases en el dataset

La métrica que utiliza los true negatives con el total de false predichos es llamado inverse precisión.

AUC and ROC

El valor de AUC (Area Under the Curve) se calcula en base a una curva aproximada en el espacio ROC (Receiver Operating Characteristic), que se forma usando los valores de true positive rate y false positive rate.

La curva en este espacio se forma al usar la propiedad de los clasificadores con redes neuronales de entregar un valor como probabilidad de pertenecer a dicha clase o a otra. De esta forma, al valorarse mediante el conjunto de test se podrá asociar una posición en el espacio ROC de acuerdo con cuál es la label del elemento evaluado y la probabilidad entregada por el clasificador. El conjunto de estos puntos formara la curva ROC.

Una de las propiedades de esta curva es que mientras esta se encuentre cerca del eje true positives de la gráfica ROC mejor será el modelo, ya que los true positives serán más predominantes que los false positives en dicho modelo. Siendo el peor modelo uno que se acerque a una relación lineal entre true positive y false positive ya que esta corresponde a una clasificación al azar de las labels.

De esta forma el AUC permite calificar un modelo, ya que el área bajo la curva en el espacio ROC será mayor a medida que la curva se acerque al eje de true positives, y menor a medida que este se acerque a una relación lineal de los ejes. Esta métrica de desempeño considera todas las clases que puede predecir el modelo y no se ve muy afectada por datasets desbalanceados.

Es de interés conocer que en el caso de poseer una curva que se acerque al eje de false positives podría indicar un problema en la configuración de la red neuronal o de estructura del dataset.(Fawcett, 2016)

Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)}(TP + FN)(TN + FP)(TN + FN)}$$

Este coeficiente es un caso especial del coeficiente phi y fue diseñado inicialmente para la comparación de estructuras químicas y luego fue propuesto como una métrica de desempeño en machine learning, el cual podía ser extendido a multiclases.

Esta métrica no se ve afectada por el desbalance de clases en el dataset como si lo hace accuracy y recall. Y es la única métrica que produce un score alto cuando se predice correctamente la mayoría de positivos y negativos. Esta métrica varía entre –1 y 1 siendo estos extremos los casos todos los positives y negatives son predichos incorrecta o correctamente respectivamente, y 0 el caso donde son predichos aleatoriamente.

Esta métrica funciona bien cuando hay desbalance de clases y provee información global del desempeño del modelo. Sin embargo, se ve en dificultades cuando en la matriz de confusión existe columnas o filas en las que todos sus elementos son 0. Pero existen correcciones de la métrica que toman en cuenta este problema.

F1-score

$$F1 - Score = 2 \times \frac{\mathrm{Pr}\, e\, cisión \times Recall}{\mathrm{Pr}\, e\, cisión + Recall}$$

Conocido como el promedio armónico entre la precision y el recall, es independiente de los true negatives y no distingue entre falsos positivos y falsos negativos al igual que accuracy, su valor varía entre 0 y 1 siendo estos donde 1 indica una precision y recall perfecta y 0 indica que existen alta cantidad de falsos negativos, falsos positivos o ambos.

Al ser una métrica que relación precision y recall es útil en contextos donde se busca balancear una buena relación de positives correctamente identificados dentro del dataset y que la cantidad de false positives no sea alta

La definición de eta metica aplicada en la matriz de confusión le permite estar definida en los casos donde TP=0, FP>0, y FN>0