**Módulo: Machine Learning**

**Actividad: Optimizacion de modelos**

**Nombre: Alan Giovanni Venegas Turrubiartes**

**Docente: Jesús Manuel Vázquez Nicolá**

**Fecha: 15/09/2024**

Introducción

En este reporte se presenta el análisis detallado de la optimización de un modelo de clasificación utilizando Support Vector Machine (SVM). Se empleó GridSearchCV para ajustar los hiperparámetros del modelo y maximizar su rendimiento. El objetivo fue encontrar la mejor configuración de los parámetros C, gamma y kernel para mejorar la precisión y la capacidad del modelo para identificar correctamente las clases en el conjunto de datos.

Metodología

Modelo Inicial: Primero se entrenó un modelo de SVM utilizando valores predeterminados de los hiperparámetros. Esto nos proporcionó una referencia inicial del rendimiento del modelo.

Optimización de Hiperparámetros:

Parámetros evaluados:

C (Coeficiente de regularización): Valores de [0.1, 1, 10, 100]. Este parámetro controla la complejidad del modelo; valores más altos permiten una mayor complejidad, mientras que valores bajos simplifican el modelo.

gamma (Parámetro del núcleo RBF): Valores de [1, 0.1, 0.01, 0.001]. Este parámetro controla el alcance del impacto de un solo punto de datos; valores más bajos significan un alcance mayor.

kernel (Tipo de núcleo): ['rbf'], el núcleo Radial Basis Function (RBF) es usado para manejar la no linealidad en los datos.

Técnica: Se utilizó GridSearchCV con validación cruzada de 5 pliegues para evaluar la combinación de hiperparámetros. Esto significa que el conjunto de datos se dividió en 5 partes, y el modelo fue entrenado y evaluado en cada una de ellas para asegurar que los resultados no dependieran de una sola división de datos.

Número total de combinaciones: 16 combinaciones posibles de hiperparámetros (4 valores de C x 4 valores de gamma x 1 valor de kernel).

Resultados del Modelo Inicial

Antes de la optimización, el modelo de SVM mostró las siguientes métricas:

Accuracy (Exactitud): 0.9561. Esto significa que el modelo clasificó correctamente el 95.61% de las muestras.

Precision (Precisión): 0.9459. Indica que el 94.59% de las veces que el modelo predijo una clase positiva, lo hizo correctamente.

Recall (Recuperación): 0.9859. Muestra que el modelo identificó correctamente el 98.59% de las muestras positivas reales.

F1 Score: 0.9655. Esta métrica combina precisión y recuperación en un solo valor, indicando que el equilibrio entre ambos es alto.

Resultados de la Optimización

Después de realizar la búsqueda de hiperparámetros, la mejor configuración encontrada fue:

C: 1

gamma: 0.001

kernel: 'rbf'

Las métricas de rendimiento con esta configuración óptima fueron:

Accuracy (Exactitud): 0.9298. La exactitud del modelo disminuyó ligeramente a 92.98%. Esto significa que el modelo clasifica correctamente el 92.98% de las muestras. La ligera disminución respecto al modelo inicial puede sugerir que la optimización ha ajustado el modelo para balancear mejor entre precisión y recuperación.

Precision (Precisión): 0.9565. La precisión mejoró ligeramente a 95.65%, lo que indica que el modelo es muy eficiente en identificar correctamente las clases positivas sin generar muchos falsos positivos.

Recall (Recuperación): 0.9296. La recuperación disminuyó a 92.96%, lo que significa que el modelo identificó el 92.96% de las muestras positivas reales. La ligera disminución en comparación con el modelo inicial sugiere que el modelo es un poco menos sensible a las muestras positivas, pero aún así mantiene un buen rendimiento.

F1 Score: 0.9429. La puntuación F1 disminuyó a 94.29%, reflejando una ligera reducción en el equilibrio entre precisión y recuperación. A pesar de esta reducción, la puntuación sigue siendo alta, indicando un buen rendimiento general.

Análisis de Resultados

Accuracy (Exactitud): La exactitud ligeramente reducida sugiere que el modelo optimizado podría estar mejor equilibrado para manejar datos nuevos, evitando el sobreajuste a los datos de entrenamiento.

Precision (Precisión): La mejora en la precisión indica que el modelo optimizado es más efectivo en evitar falsos positivos, lo cual es útil en aplicaciones donde es crítico no etiquetar erróneamente un caso positivo.

Recall (Recuperación): La ligera disminución en la recuperación puede significar que el modelo está siendo más selectivo en cuanto a qué casos etiquetar como positivos. Esto puede ser una compensación por una mayor precisión.

F1 Score: La ligera caída en la puntuación F1 refleja el cambio en el equilibrio entre precisión y recuperación, pero el valor sigue siendo alto, mostrando que el modelo mantiene un buen rendimiento general.

Conclusión

La optimización del modelo SVM utilizando GridSearchCV ha permitido encontrar una configuración de hiperparámetros que ofrece un balance mejorado entre precisión y recuperación. Aunque hay una ligera disminución en la exactitud y la recuperación en comparación con el modelo inicial, la mejora en la precisión y la alta puntuación F1 indican que el modelo optimizado está bien ajustado y es robusto. Esta configuración sugiere un modelo que es estable y eficaz para la clasificación en datos nuevos.

Para futuras optimizaciones, se podría considerar explorar más técnicas de preprocesamiento de datos, o ajustar otros hiperparámetros para lograr un equilibrio aún mejor, dependiendo de los requisitos específicos de la aplicación.

BIBLIOGRAFIAS