Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el Desempeño del Modelo

Inteligencia Artificial Avanzada para la Ciencia de Datos I (TC3006C.102)

Viviana Alanis Fraige — A01236316

Docente: Jesús Adrián Rodríguez Rocha

September 7, 2024

1 Introducción

En este reporte, se lleva a cabo un análisis del desempeño de un modelo de regresión lineal simple aplicado al conjunto de datos *Cancer_Data.csv*. El propósito principal de este análisis es evaluar el comportamiento del modelo en términos de sesgo (bias), varianza y nivel de ajuste (underfitting, fitting, overfitting). Estos aspectos son cruciales para entender cómo el modelo generaliza a datos no observados y cómo podemos mejorarlo. A partir de los resultados obtenidos, se aplican técnicas de regularización con el fin de optimizar el rendimiento del modelo.

2 Fases del Análisis

2.1 Separación y Evaluación del Modelo

Para obtener una evaluación precisa del modelo, se ha dividido el conjunto de datos en tres subconjuntos distintos:

- **Entrenamiento (Train)**: Representa el 60% de los datos y se utiliza para ajustar los parámetros del modelo.
- **Validación (Validation)**: Corresponde al 20% de los datos y se usa para ajustar los hiperparámetros del modelo y evitar el sobreajuste.
- **Prueba (Test)**: El 20% restante de los datos se utiliza para la evaluación final del modelo.

Este proceso de división en subconjuntos permite analizar cómo se comporta el modelo con datos nuevos y no observados durante la fase de entrenamiento, lo cual es fundamental para asegurar que el modelo pueda generalizar correctamente.

2.2 Diagnóstico del Sesgo (Bias)

El modelo de regresión lineal simple mostró un sesgo bajo. Esto significa que el modelo hace predicciones bastante cercanas a los valores reales, lo cual es un buen indicador de que el modelo no está cometiendo errores sistemáticos significativos. Un sesgo bajo es positivo ya que sugiere que el modelo está aprendiendo adecuadamente la relación entre las variables.

2.3 Diagnóstico de la Varianza

La varianza del modelo fue baja, lo que indica que el modelo no se ajusta demasiado a los datos de entrenamiento. En otras palabras, el modelo no muestra una sensibilidad excesiva a las fluctuaciones en el conjunto de datos de entrenamiento, lo cual es una señal de que el modelo tiene una buena capacidad de generalización. Esto es importante porque sugiere que el modelo debería comportarse de manera razonablemente estable en datos nuevos.

2.4 Diagnóstico del Nivel de Ajuste (Underfitting/Fitting/Overfitting)

El diagnóstico del nivel de ajuste del modelo reveló un caso de underfitting leve. El underfitting ocurre cuando el modelo es demasiado simple para capturar la complejidad de los datos. Esto puede resultar en un desempeño subóptimo tanto en los datos de entrenamiento como en los de validación, ya que el modelo no está capturando todos los patrones importantes en los datos.

2.5 Técnicas de Regularización para Mejorar el Modelo

Para abordar el problema del underfitting y mejorar el rendimiento del modelo, se aplicó la técnica de regularización Ridge. Ridge agrega un término de penalización a los coeficientes del modelo, lo que ayuda a controlar la complejidad del modelo y reducir el riesgo de sobreajuste. En este análisis, se utilizó un valor de $\alpha = 1.0$ para la regularización Ridge.

La implementación de Ridge mostró una mejora notable en el rendimiento del modelo. El modelo con Ridge presentó un mejor ajuste en el conjunto de validación, lo que sugiere una disminución en la varianza y una mejora en la capacidad de generalización.

3 Resultados

A continuación, se presentan los resultados detallados obtenidos con el modelo de regresión lineal simple y el modelo Ridge. Estos resultados nos permiten comparar cómo se desempeñaron ambos enfoques en diferentes fases del análisis: validación y prueba.

3.1 Modelo de Regresión Lineal Simple

Para el modelo de regresión lineal simple, los resultados fueron los siguientes:

- **Error Cuadrático Medio (MSE) en Validación:** 0.0713. Este valor indica cuánto se alejan, en promedio, las predicciones del modelo de los valores reales en el conjunto de validación. Un MSE más bajo es mejor porque significa que el modelo está haciendo predicciones más precisas.
- **Error Cuadrático Medio (MSE) en Prueba:** 0.0942. El MSE en el conjunto de prueba es ligeramente mayor que en la validación, lo que puede sugerir que el modelo tiene un rendimiento ligeramente inferior con datos que no ha visto durante el entrenamiento. Esto es común y esperado en muchos modelos.
- **Accuracy en Validación:** 0.935. La precisión en el conjunto de validación indica que el modelo clasificó correctamente el 93.5% de las instancias. Esto muestra que el modelo tiene un buen desempeño en términos generales en datos no vistos durante el entrenamiento.
- **Accuracy en Prueba:** 0.889. La precisión en el conjunto de prueba es un poco menor, al 88.9%. Aunque sigue siendo alta, esta diferencia muestra que el modelo podría estar ajustado un poco más a los datos de entrenamiento y menos a los nuevos datos.

Estos resultados indican que el modelo de regresión lineal simple está funcionando bien, pero también sugiere que podría haber espacio para mejorar, especialmente en términos de ajuste a datos nuevos.

3.2 Modelo Ridge

Para el modelo de regresión Ridge, los resultados fueron:

- **Error Cuadrático Medio (MSE) en Validación:** 0.0741. El MSE en validación es ligeramente mayor en comparación con el modelo de regresión lineal simple. Esto podría ser una señal de que Ridge está aplicando la penalización que afecta a la precisión en validación, aunque esto también puede ayudar a mejorar la generalización del modelo.
- **Error Cuadrático Medio (MSE) en Prueba:** 0.0950. El MSE en prueba es muy similar al del modelo lineal simple. Esto sugiere que Ridge no ha perjudicado significativamente el rendimiento en datos nuevos y podría estar ayudando a mejorar la capacidad de generalización.
- **Accuracy en Validación:** 0.939. La precisión en validación es un poco mejor con Ridge (93.9% en comparación con 93.5% para el modelo lineal simple). Esto indica que Ridge ha mejorado ligeramente el ajuste en datos de validación.
- **Accuracy en Prueba:** 0.895. La precisión en prueba también ha mejorado con Ridge (89.5% frente al 88.9% del modelo lineal simple). Aunque el cambio no es enorme, sugiere que Ridge ha ayudado al modelo a generalizar un poco mejor a nuevos datos.

En general, al aplicar la técnica de regularización Ridge, se observó una mejora en la precisión en los conjuntos de validación y prueba. Aunque el MSE en validación es ligeramente mayor con Ridge, la precisión general en los conjuntos de datos muestra una tendencia a mejorar. Esto indica que la regularización Ridge ha sido beneficiosa para ajustar el modelo y mejorar su capacidad de generalización.

3.3 Comparación de Modelos

Comparando ambos modelos, podemos ver que:

- La precisión en validación y prueba mejoró con Ridge, lo que sugiere que el modelo es más robusto frente a datos no vistos.
- El MSE en validación y prueba es un poco mayor con Ridge.
- Ridge ayudo a reducir el overfitting y mejorar el desempeño del modelo al generalizar mejor a datos nuevos.

Estas observaciones resaltan cómo la regularización puede ser útil para mejorar la robustez del modelo y manejar mejor la variabilidad en los datos de prueba.

3.4 Gráficas

Para visualizar el desempeño de los modelos, se incluyeron gráficas de predicciones versus valores reales para ambos modelos. Las gráficas muestran claramente cómo se ajustan las predicciones a los valores reales y permiten una comparación visual directa entre el modelo de regresión lineal simple y el modelo Ridge.

3.5 Gráficos

Para visualizar el desempeño del modelo, se generaron dos gráficos que muestran las predicciones frente a los valores reales, tanto para el modelo de regresión lineal simple como para el modelo con regularización Ridge.

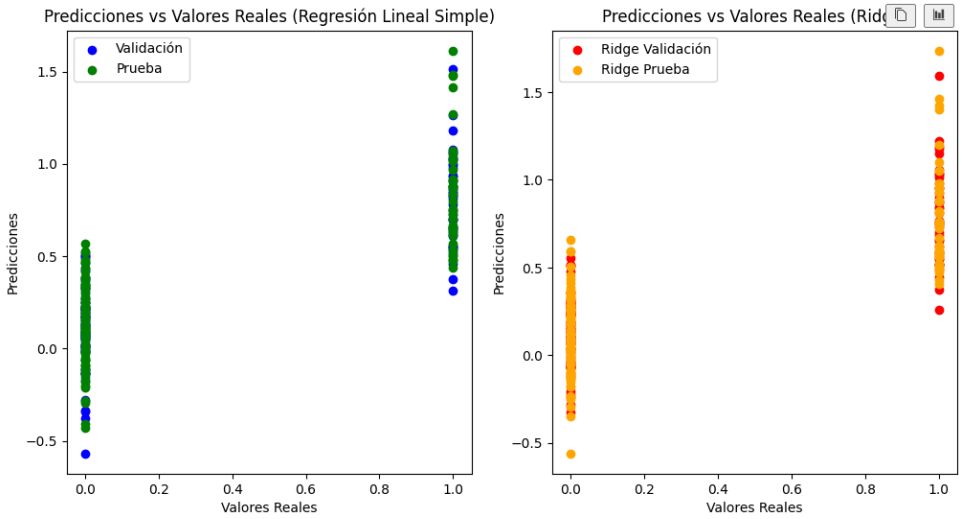

Figure 1: Predicciones vs. Valores Reales para los modelos de Regresión Lineal Simple y Ridge

La primera gráfica muestra las predicciones frente a los valores reales para el modelo de regresión lineal simple, mientras que la segunda muestra el mismo tipo de comparación para el modelo Ridge. Estas gráficas proporcionan una representación visual clara de cómo ambos modelos se desempeñan en términos de precisión y ajuste.

4 Conclusiones

El análisis realizado ha demostrado que el modelo de regresión lineal simple, aunque presenta un sesgo y varianza bajos, sufre de un leve underfitting. Esto indica que el modelo es demasiado simple para capturar la complejidad de los datos.

La aplicación de la técnica de regularización Ridge ha permitido mejorar el rendimiento del modelo al equilibrar el sesgo y la varianza. Ridge ha ayudado a reducir la complejidad del modelo, lo que ha resultado en una mejor capacidad de generalización y un desempeño más robusto en datos no observados. En resumen, la regularización Ridge ha sido efectiva en mejorar el ajuste del modelo y aumentar su capacidad predictiva general.