LTIC-0316: Sistemas basados en conocimiento Unidad 7

Tarea 4: Error de sesgo y varianza

Entrega: 6 de mayo, 2025

El entregable de esta tarea es un cuaderno electrónico que contenga el análisis correspondiente a los ejercicios propuestos.

Datos

Para esta tarea vamos a utilizar el conjunto de datos de pacientes hepáticos de la India (*Indian Liver Patient Dataset*), tomado del repositorio de UCI[1]. Este conjunto contiene información médica relevante para el diagnóstico de enfermedades hepáticas, lo que nos permitirá aplicar diferentes modelos de clasificación y analizar el *trade-off* entre sesgo (bias) y varianza (variance). El objetivo será desarrollar un modelo que pueda predecir con precisión si un paciente tiene una enfermedad hepática basándose en sus parámetros clínicos, mientras exploramos cómo diferentes niveles de complejidad en los modelos afectan este equilibrio fundamental en aprendizaje automático.

El conjunto de datos contiene 416 registros de pacientes con problemas hepáticos y 167 registros de personas sin problemas hepáticos. Los datos fueron recolectados de muestras de prueba en el noreste de Andhra Pradesh, India. El atributo `is_patient` corresponde a la clase (paciente hepático o no). El conjunto de datos contiene 441 registros de pacientes masculinos y 142 registros de pacientes femeninos. Cualquier paciente cuya edad excediera los 89 años está registrado como de edad `90`.

Los atributos del conjunto de datos son:

1. *age*: Edad del paciente
2. *gender*: Género del paciente
3. *tot_bilirubin*: Bilirrubina Total
4. *direct_bilirubin*: Bilirrubina Directa
5. *alkphos*: Fosfatasa Alcalina
6. *sgpt*: Alanina Aminotransferasa
7. *sgot*: Aspartato Aminotransferasa
8. *tot_proteins*: Proteínas Totales
9. *albumin*: Albúmina
10. *ag_ratio*: Relación Albúmina y Globulina
11. *is_patient*: Clase (datos etiquetados por expertxs)

[1] `https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset`

Análisis

Mi recomendación para esta tarea es que intentes ponerte en el papel de alguien que de verdad se dedica al análisis de datos. Intenta incluso disfrutar el ejercicio (¡estás practicando para analizar datos médicos! Puedes decir, con lo que sabes de matemáticas y programación, qué personas podrían estar enfermas del hígado!).

1. **Preprocesamiento.** Primero, explora los datos para determinar ajustes que sean necesarios: imputar o eliminar instancias con datos faltantes, normalizar clases, discretizar atributos. En tu cuaderno de trabajo, redacta las decisiones que hayas tomado y su justificación.

2. Divide el conjunto de datos en dos subconjuntos:

   - Train (80 %). Esto lo vamos a usar para entrenamiento y para ajuste de hiperparámetros (usando validación).
   - Test (20 %). Esto lo vamos a reservar para un único uso: la evaluación final solo de los modelos ya seleccionados.

3. En esta tarea, vamos a comparar tres enfoques para hacer la clasificación: regresión logística, k-vecinos más cercanos, y árboles de decisión. Para cada uno de estos métodos, utiliza `sklearn`, a través de comandos como `validation_curve` y `ShuffleSplit` para establecer los parámetros apropiados. En este ejercicio vamos a considerar profundidad del árbol (en árboles de decisión), canidad de vecinos (en kNN), y el inverso de la fuerza de regularización (C, en regresión logística).

   - ¿Qué métrica tiene sentido para este conjunto de datos? Justifica tu respuesta.
   - Utiliza 10 grupos para hacer tus pruebas (en `ShuffleSplit`) y, en cada experimento, 80 % de los datos para entrenamiento y el resto para validación (que serían el *test* de cada prueba).
   - Determina rangos apropiados para cada parámetro que estás investigando (profundidad, k, C)
   - Analiza los resultados (lo más recomendable es a través de una gráfica)

4. Decide cuál es la configuración apropiada para cada modelo. Este es uno de los pasos más importantes de la tarea y el que más me importa. Para cada modelo, dime qué valor de parámetro te parece apropiado y justifica tu respuesta (redacta tu análisis!!!).

5. Ahora la prueba final. Compara los tres modelos (usando la configuración ganadora de cada caso) en el conjunto de prueba (que reservamos al inicio). Además de indicar qué modelo tuvo mejor desempeño, compara la métrica obtenida en el conjunto de entrenamiento y en el de prueba, para que puedas conjeturar qué tipo de error predomina en cada caso. Redacta tus conclusiones.