Análisis de Grandes Volúmenes de Datos
Carlos Espinoza Peraza - B92786
Reporte 24 de mayo

Resumen de los temas estudiados

Evaluación de Modelos

Hay diferentes métodos para saber si nuestro modelo es efectivo:

```
- Aproximación apartir de un proceso
- Se crea un ambiente de evaluación
- Generación de reportes con gráficas
```

Conceptos básicos:

- Train set: Es la porción del dataset con el cual se va a entrenar el modelo
- Test set: Es la porción del dataset sobre la cual se van a realizar predicciones

Entrenamiento y Evaluación

No hay garantía sobre la representatividad de los conjuntos, es decir, puede que las muestras del set de entrenamiento sean las justo las que me permiten generar un modelo preciso o por el contrario, por una mala distribución, puede que el modelo no sea capaz de predecir de manera precisa porque en el set de entrenamiento no se vieron casos parecidos a los que se están evaluando

> Resultados altamente sensibles

Particionar el dataset

- **Training**: Conocer el problema y ajustar el modelo
- **Validation**: Evaluar la efectividad del modelo durante su ajuste
- **Test**: Evaluar el rendimiento del modelo

Cross-Validation

- Método estadístico diseñado para analizar el rendimiento de un modelo
- Hace particiones del dataset y entrena diferentes modelos con estos subsets
- Una de sus formas más conocidas es el K-Fold

Tipos:

```
- Stratified K-Fold Cross-validation
- Leave one out Cross-validation LOOCV
- Shuffle split Cross-validation
- Cross-validation with Groups
- Nested Cross Validation
```

Comentarios sobre la materia estudiada

Me parece importante que analicemos más a profundidad los resultados que pueden arrojar los modelos. Como hemos visto en varias ocaciones, a veces fue suerte que el subset que escogimos para entrenamiendo y evaluación arrojaran buenos resultados en cuanto a precisión, pero este no siempre era el caso. Además, la importancia que tiene el método de prueba y error en este campo de estudio. La configuración de hiperparámetros no es una ciencia exacta por lo que métodos como `CV` nos permiten hacer pruebas sobre diferentes combinaciones para lograr encontrar un rendimiento óptimo

Dudas sobre la materia estudiada

> ¿Cuál es la mejor forma de analizar una problema de clasificación teniendo datos desbalanceados?

Posible uso como profesional

Utilizar este tipo de evaluaciones para revisar la eficiencia real de un modelo. Por ejemplo, en los modelos que predicen cuales son las recomendaciones de productos según datos de compras anteriores del cliente o personas similares a este.

Problemas que podría resolver con las técnicas estudiadas

Se puede resolver la incertidumbre sobre si un modelo es apto para atacar un problema. Con esto se pueden tener medidas mucho más precisas sobre el rendimiento real de los modelos. Además, permite analizar las configuraciones de distintos modelos para encontrar cual es el mejor ajuste de los hiperparámetros, obteniendo así el máximo rendimiento.