Interpretabilidad del Deep Learning

Coeficiente de Explicabilidad-Rendimiento

Christian Oliva Moya

Contenido del curso

- Primera semana:
  - Introducción de conceptos
  - Explicabilidad genérica
    - SHAP
    - LIME
  - **Coeficiente de Explicabilidad-Rendimiento**

Motivación (I)

- Distintos algoritmos de explicabilidad dan distintos resultados

¿Cómo sabemos cuál es mejor?

- Hay que medir el equilibrio entre el rendimiento del modelo y la calidad de su explicabilidad: **Coeficiente de Explicabilidad Rendimiento** [1]

- Necesitamos definir una métrica R para definir atributos:
  - Positivamente relevantes
  - Negativamente relevantes

[1] C. Oliva, and L. F. Lago-Fernández. “The Explainability-Performance Coefficient: A New Metric for Model Transparency, ICANN, 2025.

Motivación (II)

- Un atributo es positivamente relevante cuando, si se aumenta su magnitud,

  el modelo aumenta la predicción

- Un atributo es negativamente relevante cuando, si se aumenta su magnitud,

  el modelo disminuye la predicción

Motivación (III)

En otras palabras:

- Un atributo es positivamente relevante cuando, si se anula, **el modelo disminuye la predicción**
- Un atributo es negativamente relevante cuando, si se anula, **el modelo aumenta la predicción**

Requisitos (I)

- Hay que diferenciar varias situaciones diferentes:

    1. La explicabilidad es global (permutación, oclusión global)
    2. La explicabilidad es local (oclusión local, SHAP, LIME)
        a. La relevancia indica el comportamiento respecto a un punto base (oclusión local, SHAP)
        b. La relevancia indica la dirección de cambio en el modelo (LIME)

Requisitos (II)

- **Cuando la explicabilidad es global** (permutación, oclusión global)
    - Rx ya define los atributos positiva y negativamente relevantes

Requisitos (III)

- **Cuando la explicabilidad es local** y la relevancia indica el comportamiento respecto a un punto base (oclusión local, SHAP):
    - Rx ya define los atributos positiva y negativamente relevantes

Requisitos (IV)

- **Cuando la explicabilidad es local** y la relevancia indica la dirección de cambio (LIME):
  - Transformamos la relevancia para saber el comportamiento según un punto base: $Rx = Rx \cdot sign(x)$
    - Si $Rx > 0 \rightarrow$ El atributo es positivamente relevante
    - Si $Rx < 0 \rightarrow$ El atributo es negativamente relevante

Coeficiente de Explicabilidad-Rendimiento (I)

- Queremos definir un método que mida la **calidad** de la explicación
- Recordemos qué es la explicabilidad: Identificación de los atributos del conjunto de datos que tienen una **mayor influencia en las predicciones** de un modelo

¿Qué tiene que pasar si identificamos correctamente los atributos con mayor influencia y los separamos del resto?

Coeficiente de Explicabilidad-Rendimiento (II)

¿Qué tiene que pasar si identificamos correctamente los atributos con mayor influencia y los separamos del resto?

- Si identificamos los atributos con mayor relevancia:
  - Si elimino los que no son relevantes, **al modelo debería darle igual** y seguir funcionando bien
  - Si elimino los que sí son relevantes, **el modelo debería fallar**

Coeficiente de Explicabilidad-Rendimiento (III)

- En otras palabras:
- Si Rx diferencia atributos positiva y negativamente relevantes:
  - Si elimino los negativamente relevantes, **el modelo debería reforzar la clase real**
  - Si elimino los positivamente relevantes, **el modelo debería castigar la clase real**

Coeficiente de Explicabilidad-Rendimiento (IV)

- Para un modelo M y una métrica m (accuracy, R2, etc), si un umbral U divide Rx en relevantes (Rx>U) y no relevantes (Rx<U):
- **Prueba 1**: Si se anulan todos los atributos que quedan por debajo del umbral según Rx < U, se puede calcular m+ sobre los supervivientes
- **Prueba 2**: Si se anulan todos los atributos que quedan por encima del umbral según Rx > U, se puede calcular m- sobre los supervivientes

Coeficiente de Explicabilidad-Rendimiento (V)

- **Hipótesis**:

Si se selecciona el conjunto de atributos más relevantes ar *del total de atributos* at *y el resto se anulan, el rendimiento del modelo* m+ *no debe verse afectado de manera significativa y, en el caso contrario, el rendimiento del modelo* m- *debe reducirse considerablemente*

Coeficiente de Explicabilidad-Rendimiento (VI)

- **Definición**:

$$EPC(M|R,U) = \frac{at - ar}{at} \times \frac{m^{+} - m^{-}}{m}$$

- El primer término define el porcentaje de atributos anulados
- El segundo término mide el efecto de anular dichos atributos en el modelo

Coeficiente de Explicabilidad-Rendimiento (VII)

$$EPC(M|R,U) = \frac{at - ar}{at} \times \frac{m^{+} - m^{-}}{m}$$

Maximizar el EPC consistirá en **eliminar el mayor número de atributos** posibles (aumentando el primer término) **sin disminuir el rendimiento** del modelo (manteniendo el segundo término)

Si una métrica de explicabilidad es buena, definirá correctamente los atributos más relevantes. Por tanto, **se podrá maximizar el EPC**

Coeficiente de Explicabilidad-Rendimiento (VIII)

- Vamos a implementarlo a mano. Vamos al notebook!
    - Notebook *2.2 Coeficiente de Explicabilidad Rendimiento*