Revisión del libro de Molnar

Conceptos básicos sobre explicabilidad

Javier París

Table of contents

1 Interpretabilidad

1.1 ¿Qué es la interpretabilidad?

Es un grado, no una propiedad binaria.

- El nivel con el que un humano puede entender las causas de una decisión.
- El nivel con el que un humano puede predecir el resultado de un modelo.

1.2 ¿Qué es la explicación?

Es una respuesta a una pregunta del tipo:

- "why" question
- "what if" question

1.3 Importancia (por qué no confiar en un modelo)

- **Seguridad**: poder confiar en aplicaciones críticas
- **Sesgos**: detectar y corregir sesgos
- **Conocimiento**: el objetivo de la ciencia es el entendimiento
- **Aceptación**: para ser humana/socialmente aceptable

1.4 Objetivo de la interpretabilidad

- Equidad - *Fairness*
- Privacidad - *Privacy*
- Fiablididad - *Reliability*
- Causalidad - *Causality*
- Confianza - *Trust*

1.5 Diferencia entre modelo y método

Modelo

Algoritmo que ajusta una función a los datos de entrenamiento.

- KNN
- SVM
- ANN

Método

Algoritmo para obtener explicaciones de modelos ya entrenados.

- LIME
- SHAP
- Counterfactuals

2 Taxonomía

2.1 Según el modelo

- **Modelos intrínsecamente interpretables**: modelos que son interpretables por sí mismos.
  - Modelos lineales
  - Modelos basados en reglas
- **Post-hoc** métodos sobre modelos ya entrenados
  - LIME
  - SHAP

2.2 Resultados de los métodos de interpretación

- Resumen estadístico de atributos
- Resumen visual de atributos
- Explicaciones internas del modelo
- Explicación mediante datos individuales
- Aproximación mediante modelos interpretables

2.3 Según el método

- Métodos agnósticos de modelo: LIME
- Métodos específicos: ANN

2.4 Según el rango de interpretación

- **Local**: interpretación de una predicción (un dato)
- **Global**: interpretación del modelo completo

2.5 Rango de la interpretabilidad

- **Transparencia**: ¿Cómo el algoritmo crea el modelo?
- **Global**:
  - **Holístico**: ¿Cómo se lleva a cabo la inferencia?
  - **Modular**: ¿Qué parte del modelo afecta a qué en la inferencia?
- **Local**:

  - **Individual**: ¿Por qué un modelo predice una salida concreta para una entrada?
  - **Grupal** ¿Por qué un modelo predice una salida similar para un grupo de entradas?

3 Propiedades de las explicaciones

3.1 Propiedades de los métodos

- **Expresividad**: estructura de la explicación
- **Transparencia**: cuánto necesita saber la explicación del interior del modelo
- **Portabilidad**: rango de modelos a los que se puede aplicar
- **Complejidad**: computacional

3.2 Propiedades de las explicaciones individuales

- **Precisión**: cómo de bien la explicación predice datos desconocidos
- **Fidelidad**: cómo de bien la explicación aproxima la predicción del modelo
- **Consistencia**: si la explicación difiere entre modelos similares
- **Estabilidad**: si la explicación cambia poco al cambiar la entrada
- **Comprensibilidad**: si la explicación es fácil de entender
- **Certeza**: si la explicación es recoge la incertidumbre del modelo
- **Importancia**: si la explicación recoge la importancia de los atributos
- **Novedad**: si el método es capaz de explicar datos alejados del conjunto de entrenamiento
- **Representatividad**: cuántos datos cubre una explicación

4 Medida de la interpretabilidad

4.1 Puntos de análisis

- Entrenamiento
  - Memoria
  - Proceso algorítmico - optimización
  - Aprendizaje gradual del modelo
- Inferencia
  - Proceso algorítmico - toma de decisiones
  - Modelo aprendido - pesos, frontera, etc.

4.2 Medidas subjetivas

- **Transparencia**:
  - **Modularidad**: diferenciación de partes del modelo
  - **Simulabilidad**: capacidad de un humano para simularlo
- **Parsimonia**:
  - **Parámetros**: número de parámetros
  - **Estructura**: modelo estructural

4.3 Medidas objetivas

- **Algorítmicas**:
  - **Complejidad algorítmica**: tiempo y espacio $O(f(n))$
  - **Complejidad ciclomática**: número de caminos independientes
  - **Número de decisiones**: reducción del problema a un sistema basado en reglas
- **Locales**:
  - **Fidelidad**: si la explicación se ajusta al modelo (es capaz de explicar datos contrafácticos)
  - **Consistencia**: si la explicación difiere entre modelos similares
  - **Monotoneidad**: si la explicación cambia de forma monótona con los datos
- **Globales**:
  - **Importancia de atributos**: medir la importancia de cada atributo, o de relaciones entre ellos