Interpretabilidad y Explicabilidad (XAI) en Sistemas Inteligentes

Estrategias computacionales para mitigar la caja negra

Prof. D.Sc. BARSEKH-ONJI Aboud
Facultad de Ingeniería
Universidad Anáhuac México
https://orcid.org/0009-0004-5440-8092
aboud.barsekh@anahuac.mx

7 de mayo de 2025

Agenda

1. Introducción: ¿Por qué 'entender' a la IA?
2. El Espectro de la Opacidad
3. Panorama General de Estrategias XAI
4. Lógica Difusa: Iluminando la Caja Negra
5. Modelos de Fuzzy Logic en XAI
6. Mirando Hacia Adelante: Retos y Futuro de XAI
7. Conclusiones

El Dilema Moderno: Potencia vs. Opacidad

La Era Dorada de la IA:

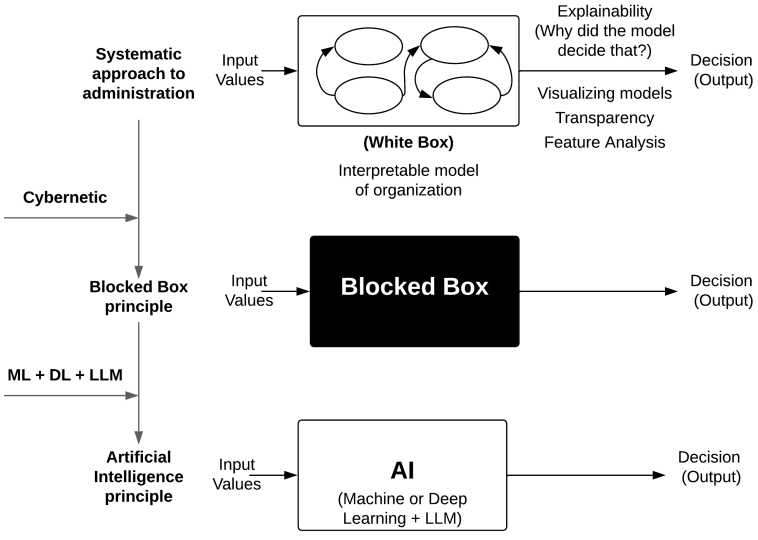
- Capacidades asombrosas (visión, lenguaje, juegos).
- Modelos cada vez más potentes (Deep Learning).

El Lado Oscuro: La Caja Negra

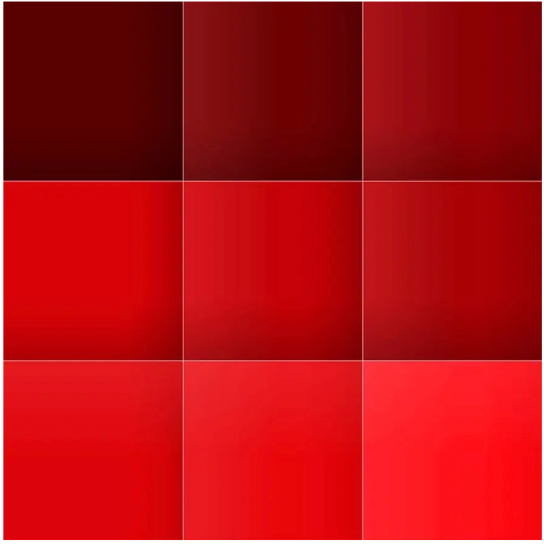
- Modelos complejos (DNNs, ensambles) cuyo funcionamiento interno es difícil de comprender.
- Reciben entradas $\rightarrow$ Producen salidas. ¿Pero cómo?

¿Podemos confiar en lo que no entendemos?

El Dilema Moderno: Potencia vs. Opacidad



Conceptos Clave: Interpretabilidad vs. Explicabilidad



Conceptos Clave: Interpretabilidad vs. Explicabilidad

Interpretabilidad

Habilidad de entender la mecánica de causa-efecto dentro de un sistema de IA.

- ¿Cómo mapea el modelo entradas a salidas?
- ¿Qué lógica (matemática, estructural) sigue?
- Enfoque en el **mecanismo** del modelo.
- Propiedad intrínseca del modelo.

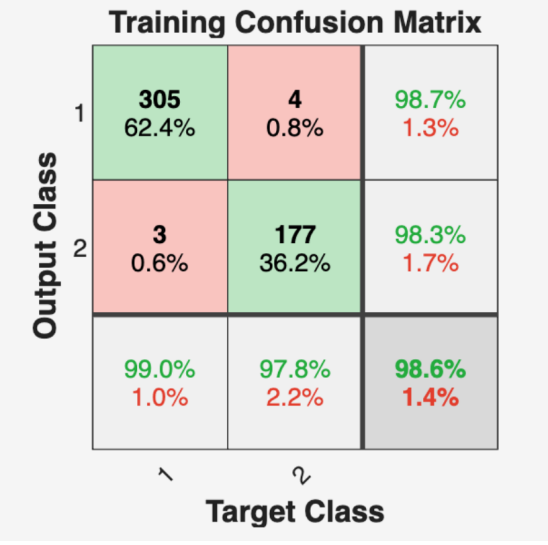
Conceptos Clave: Interpretabilidad vs. Explicabilidad

Explicabilidad (XAI - Explainable AI)

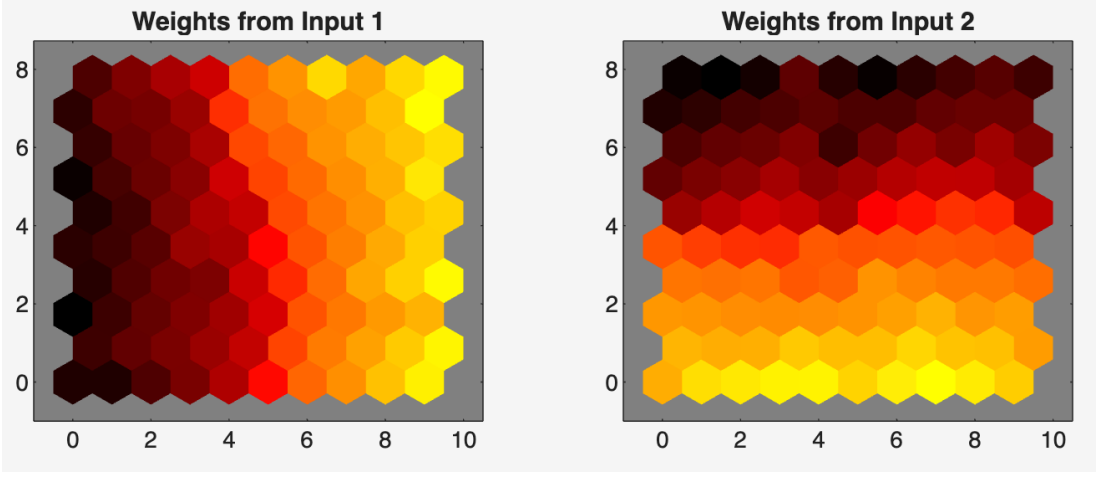
Habilidad de obtener una justificación comprensible para una decisión específica del modelo, usualmente en lenguaje humano.

- ¿Por qué esta predicción/decisión en particular?
- Implica una **interfaz/traducción** para el humano.
- Objetivo final de cara al usuario/regulador.

Conceptos Clave: Interpretabilidad vs. Explicabilidad



Conceptos Clave: Interpretabilidad vs. Explicabilidad



¿Por qué es Crucial XAI?

- **Confianza:** Fundamental para la adopción por usuarios y expertos (médicos, ingenieros...).
- **Ética y Justicia (Fairness):** Detectar y mitigar sesgos indeseados (género, raza...).
- **Robustez y Seguridad:** Entender puntos débiles y posibles ataques adversariales.
- **Depuración y Mejora:** Identificar errores y áreas de mejora del modelo.
- **Cumplimiento Regulatorio:** Normativas como GDPR (derecho a explicación), AI Act (UE).
- **Descubrimiento Científico:** Usar IA para generar hipótesis comprensibles.

XAI no es un lujo, es una necesidad para una IA responsable.

¿Cuándo y Por Qué surgen las Cajas Negras?

Características de los Modelos Opacos:

- **Alta Complejidad:** Millones/miles de millones de parámetros (DNNs).
- **No Linealidad Extrema:** Interacciones complejas y no intuitivas entre variables.
- **Representaciones Internas Distribuidas:** Conocimiento codificado de forma no localizada (capas ocultas).
- **Modelos de Ensamble:** Lógica agregada difícil de seguir (Random Forests, Gradient Boosting).

Consecuencias Reales:

- Decisiones injustas, falta de accountability, dificultad de diagnóstico, obstáculos regulatorios, fragilidad ante ataques, fricción en la adopción.

El Objetivo: Mitigación y Trade-offs

- No siempre es posible (ni necesaria) una transparencia total.
- Existe un compromiso (trade-off) entre **Rendimiento Predictivo** y **Interpretabilidad**.
- El desafío: Encontrar el balance adecuado o desarrollar métodos que mejoren la interpretabilidad sin sacrificar (demasiado) el rendimiento.

El Objetivo: Mitigación y Trade-offs

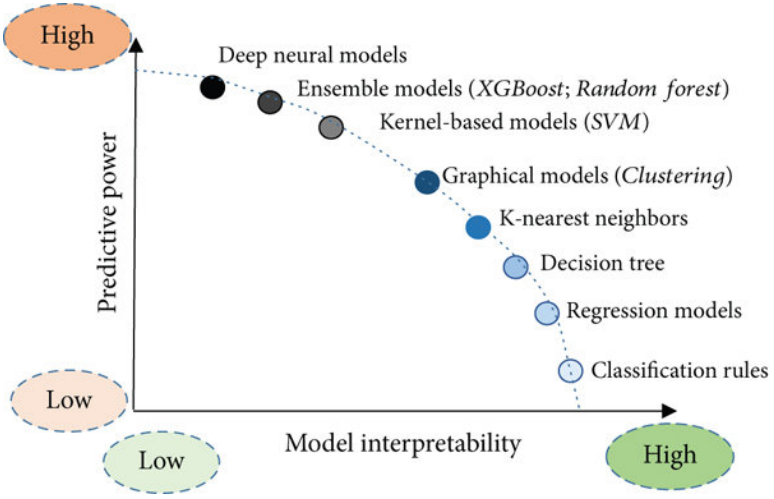


Figura: Trade-off entre Rendimiento e Interpretabilidad (Conceptual)

Dos Grandes Enfoques Metodológicos

A) Modelos Intrínsecamente Interpretables (Ante-Hoc / By Design)

Modelos cuya estructura interna es inherentemente comprensible. **Ejemplos:**

- Regresión Lineal/Logística ($y = \beta_0 + \sum \beta_i x_i$)
- Árboles de Decisión (simples)
- Sistemas Basados en Reglas
- GAMs ($g(E[Y]) = \beta_0 + \sum f_i(x_i)$)
- Lógica Difusa (Fuzzy Logic) $\rightarrow$ *Nuestro foco*

B) Métodos de Explicación Post-Hoc

Técnicas aplicadas *después* de entrenar un modelo (incluso caja negra). **Ejemplos:**

- *Model Agnostic:*
  - LIME (Local)
  - SHAP (Local/Global)
  - Feature Importance
- *Model Specific (DNNs):*
  - Grad-CAM, Saliency Maps...

Pros: Flexibilidad. **Cons:** Aproximaciones, inestabilidad, coste, riesgo de explicaciones engañosas.

Dos Grandes Enfoques Metodológicos

A) Modelos Intrínsecamente Interpretables (Ante-Hoc / By Design)

Modelos cuya estructura interna es inherentemente comprensible. **Ejemplos:**

- Regresión Lineal/Logística ($y = \beta_0 + \sum \beta_i x_i$)
- Árboles de Decisión (simples)
- Sistemas Basados en Reglas
- GAMs ($g(E[Y]) = \beta_0 + \sum f_i(x_i)$)
- Lógica Difusa (Fuzzy Logic) $\rightarrow$ *Nuestro foco*

B) Métodos de Explicación Post-Hoc

Técnicas aplicadas *después* de entrenar un modelo (incluso caja negra). **Ejemplos:**

- *Model Agnostic:*
  - LIME (Local)
  - SHAP (Local/Global)
  - Feature Importance
- *Model Specific (DNNs):*
  - Grad-CAM, Saliency Maps...

Pros: Flexibilidad. **Cons:** Aproximaciones, inestabilidad, coste, riesgo de explicaciones engañosas.

Principios Fundamentales (1/7): Motivación y Conjuntos Difusos

Motivación (Lotfi Zadeh, 1965)

Superar la rigidez de la lógica clásica (verdadero/falso, 0/1) para modelar conceptos vagos e imprecisos del lenguaje natural y razonamiento humano.

- 'Temperatura *un poco alta*', 'Cliente *bastante satisfecho*'
- Permite **grados de verdad**.

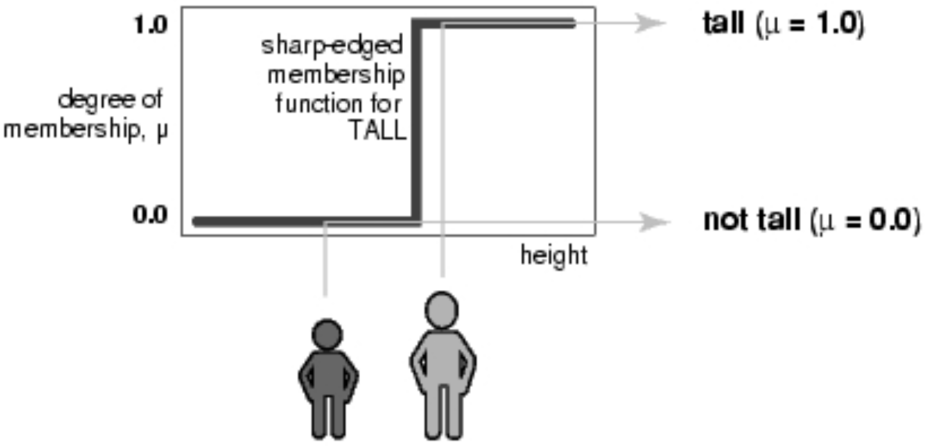
Principios Fundamentales (2/7): Motivación y Conjuntos Difusos

Examples

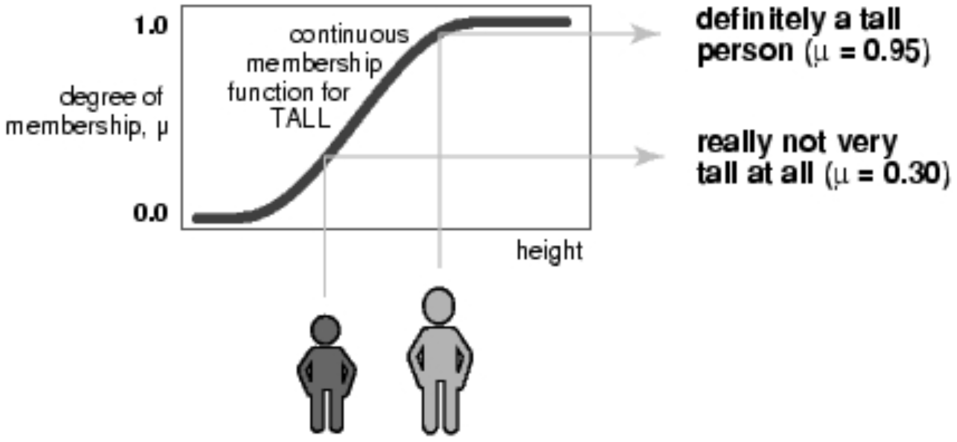
Conjuntos Difusos (Fuzzy Sets) Generalización de conjuntos clásicos. Un elemento pertenece con un **grado de pertenencia** $\mu \in [0, 1]$.

- Definido por una **Función de Pertenencia (MF)** $\mu_A : X \to [0, 1]$.
- X: Universo de discurso (rango de valores).
- $\mu_A(x)$: Grado en que x pertenece al conjunto difuso A.
- *Ejemplo:* $\mu_{\text{Cálida}}(20^\circ C) = 0{,}7$.

Principios Fundamentales (2/7): Motivación y Conjuntos Difusos



Principios Fundamentales (2/7): Motivación y Conjuntos Difusos



Principios Fundamentales (3/7): Variables Lingüísticas y MFs

Variables Lingüísticas

Variables cuyos valores son palabras o frases (términos lingüísticos) que representan conceptos difusos.

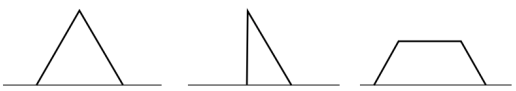
- *Ejemplo:* Variable 'CalidadServicio', Términos = {'Mala', 'Regular', 'Buena', 'Excelente'}.
- Cada término se define mediante una MF sobre un universo numérico (ej. [0, 10]).

Principios Fundamentales (4/7): Variables Lingüísticas y MFs

Examples

Funciones de Pertenencia (MFs) Representación gráfica/matemática de los conjuntos difusos. Mapean valor numérico → grado de pertenencia.

- **Formas Comunes:** Triangular (trimf), Trapezoidal (trapmf), Gaussiana (gaussmf)...
- Parámetros definen la forma (ej. trimf(x; a, b, c)).
- Elección crucial (conocimiento experto / datos).



Principios Fundamentales (5/7): Operadores y Reglas Difusas

Operadores Lógicos Difusos

Extienden AND, OR, NOT a grados de pertenencia $\mu \in [0, 1]$.

- **AND (t-norma):** Comunes: mín(a, b), $a \cdot b$.
- **OR (t-conorma / s-norma):** Comunes: máx(a, b), mín$(a + b, 1)$.
- **NOT:** Común: $1 - a$.

Principios Fundamentales (6/7): Operadores y Reglas Difusas

Reglas Difusas (IF-THEN)

El núcleo del conocimiento. Estructura: `IF <antecedente> THEN <consecuente>`.

- **Antecedente:** Proposiciones difusas ('Var IS Term') conectadas por AND/OR difusos. *Ej: `IF Temperatura IS Alta AND Humedad IS Baja`*
- **Consecuente (Mamdani):** Un conjunto difuso sobre la variable de salida. *Ej: `THEN VelocidadVentilador IS Rapida`*
- **Consecuente (Sugeno/TSK):** Una función (ej. lineal) de las entradas. (Menos interpretable directamente).

Nos centraremos en Mamdani por su interpretabilidad.

Principios Fundamentales (7/7): Sistema de Inferencia Difusa (FIS Mamdani)

Proceso para mapear entradas (crisp) $\rightarrow$ salidas (crisp).

1. **Fuzzificación:** Entradas numéricas $\rightarrow$ Grados de pertenencia (usando MFs). $\mu_{A_i}(x_0)$.
2. **Inferencia (Evaluación de Reglas):**
   - *Antecedente:* Calcular grado de activación (α) de cada regla (usando operadores AND/OR).
   - *Implicación:* Aplicar α al consecuente (ej. **Clipping**: mín$(\alpha, \mu_C(z))$).
3. **Agregación:** Combinar MFs de salida resultantes de todas las reglas activadas (ej. **Máximo**). $\mu_{agg}(z) = \text{máx}_j(\mu_{C'_j}(z))$.
4. **Defuzzificación:** Convertir $\mu_{agg}(z)$ (difusa) $\rightarrow$ valor numérico final z_{final} (crisp).
   - Método común: **Centroide (COG)**: $z_{COG} = \frac{\int z \cdot \mu_{agg}(z) dz}{\int \mu_{agg}(z) dz}$.

Principios Fundamentales (7/7): Sistema de Inferencia Difusa (FIS Mamdani)

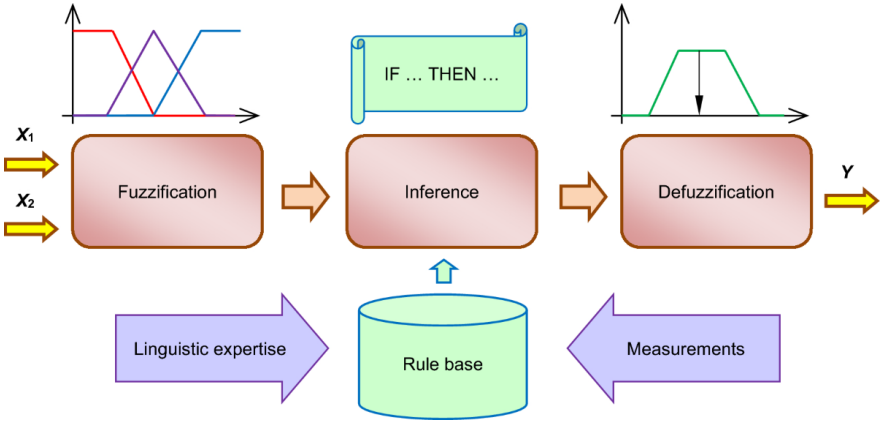


Figura: Diagrama del sistema Mamdani

¿Por qué la Lógica Difusa es Interpretable?

- **Transparencia Estructural:** El modelo *es* la base de reglas y las MFs. Sin cajas negras ocultas. El razonamiento es trazable.
- **Semántica Cercana al Lenguaje Natural:** Variables lingüísticas y reglas IF-THEN reflejan (idealmente) el razonamiento humano. Validable por expertos.
- **Modularidad:** Cada regla captura una faceta del conocimiento. Permite ajustes localizados (con cuidado).
- **Visualización Intuitiva:** Las MFs y la activación/agregación de reglas se pueden visualizar gráficamente.

Ofrece un puente entre la computación y el sentido común.

Ejemplo Práctico: Cálculo de Propina (1/5) - Definición

Problema

Determinar propina (%) basado en calidad del servicio y comida (escala 0-10).

Variables y Términos Lingüísticos:

- **Entrada 1:** 'Servicio' ($s \in [0, 10]$) $\rightarrow$ {Malo, Bueno, Excelente}
- **Entrada 2:** 'Comida' ($c \in [0, 10]$) $\rightarrow$ {Mala, Decente, Deliciosa}
- **Salida:** 'Propina' ($p \in [0, 25]$) $\rightarrow$ {Baja, Media, Alta}

Funciones de Pertenencia (MFs): Usaremos Triangulares (trimf).

- $\mu_{S,Malo}(s) = \text{trimf}(s; [0, 0, 5])$
- $\mu_{S,Bueno}(s) = \text{trimf}(s; [0, 5, 10])$
- ... (etc. para todas las variables y términos)

Ejemplo Práctico: Cálculo de Propina (2/5) - MFs

Base de Reglas:

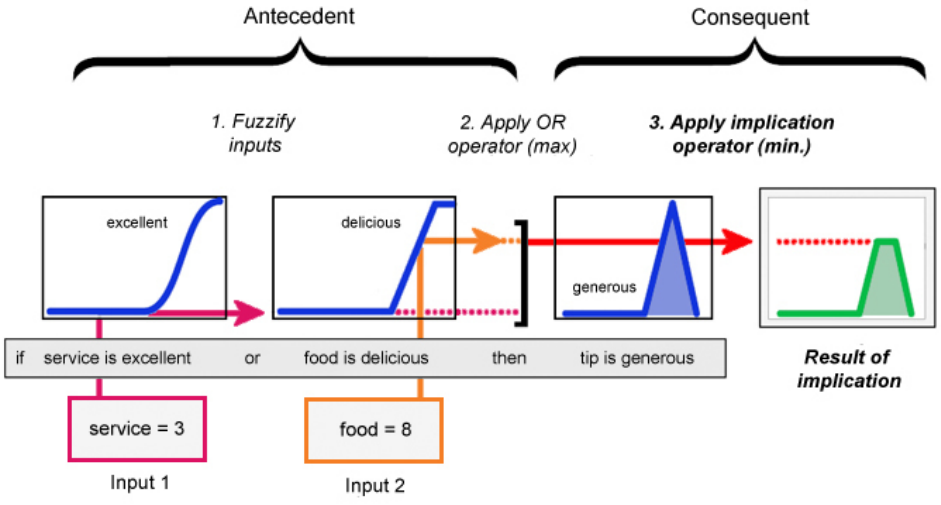
1. `IF Servicio IS Malo OR Comida IS Mala THEN Propina IS Baja`
2. `IF Servicio IS Bueno THEN Propina IS Media`
3. `IF Servicio IS Excelente AND Comida IS Deliciosa THEN Propina IS Alta`

Ejemplo Práctico: Cálculo de Propina (3/5) - Inferencia

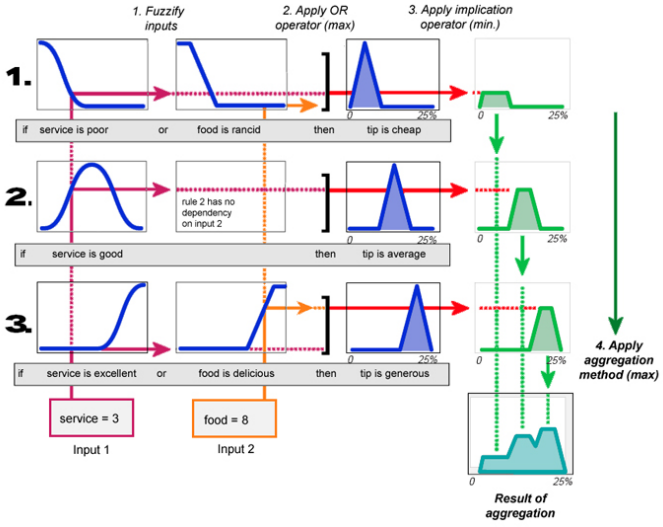
Entradas: 'Servicio = 7.5', 'Comida = 8.0'

1. **Fuzzificación:**
   - $\mu_{S,Bueno}(7{,}5) = 0{,}5$, $\mu_{S,Excelente}(7{,}5) = 0{,}5$
   - $\mu_{C,Decente}(8{,}0) = 0{,}4$, $\mu_{C,Deliciosa}(8{,}0) = 0{,}6$
   - (Otras MFs dan 0)
2. **Inferencia (Evaluación de Reglas):** (Usando máx para OR, mín para AND)
   - R1: $\alpha_1 = \text{máx}(\mu_{S,Malo}, \mu_{C,Mala}) = \text{máx}(0,0) = 0 \rightarrow$ No activa.
   - R2: $\alpha_2 = \mu_{S,Bueno}(7{,}5) = 0{,}5 \rightarrow$ Activa Propina Media con fuerza 0.5.
   - R3: $\alpha_3 = \text{mín}(\mu_{S,Excelente}, \mu_{C,Deliciosa}) = \text{mín}(0{,}5, 0{,}6) = 0{,}5 \rightarrow$ Activa Propina Alta con fuerza 0.5.
3. **Implicación (Clipping):** Cortar MFs de salida $\mu_{P,Media}$ y $\mu_{P,Alta}$ a altura 0.5.
4. **Agregación (Máximo):** Combinar las dos MFs cortadas.
   $\mu_{agg}(p) = \text{máx}(\text{mín}(0{,}5, \mu_{P,Media}(p)), \text{mín}(0{,}5, \mu_{P,Alta}(p)))$

Ejemplo Práctico: Cálculo de Propina (3/5) - Inferencia



Ejemplo Práctico: Cálculo de Propina (3/5) - Inferencia



Ejemplo Práctico: Cálculo de Propina (4/5) - Defuzzificación e Interpretación

4. **Defuzzificación (Centroide - COG):** Calcular centro de gravedad del área $\mu_{agg}(p)$.

$$p_{COG} \approx \frac{\sum_i p_i \cdot \mu_{agg}(p_i)}{\sum_i \mu_{agg}(p_i)} \approx 16{,}7\,\%$$

Ejemplo Práctico: Cálculo de Propina (5/5) - Defuzzificación e Interpretación

Interpretación de la Salida

'Para un servicio de 7.5 ('Bueno' y 'Excelente' con grado 0.5) y comida de 8.0 ('Decente' 0.4, 'Deliciosa' 0.6):

- La regla 'Si Servicio Bueno $\rightarrow$ Propina Media' se activó con fuerza 0.5.
- La regla 'Si Servicio Excelente Y Comida Deliciosa $\rightarrow$ Propina Alta' se activó con fuerza $\min(0.5, 0.6) = 0.5$.
- Combinando estas sugerencias (Media y Alta, ambas con peso 0.5), el sistema recomienda una propina final de **16.7 %**.'

Explicación trazable, basada en reglas lingüísticas y grados de activación.

Ventajas y Desventajas de la Lógica Difusa

Ventajas (Pros)

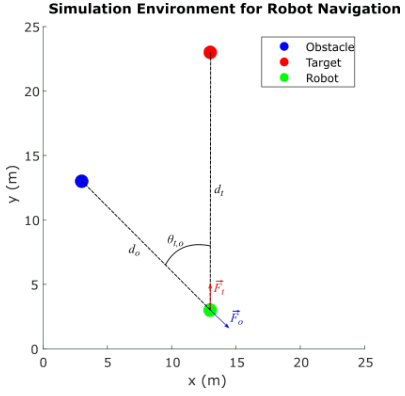
- **Alta Interpretabilidad** Inherente
- Manejo Natural de **Vaguedad** e Incertidumbre
- Incorporación de **Conocimiento Experto** (Reglas)
- **Robustez** ante Ruido (en MFs)
- Razonamiento Aproximado

Desventajas (Contras)

- Diseño de MFs y Reglas puede ser **Subjetivo/Laborioso**
- Rendimiento Predictivo puede ser **inferior a Cajas Negras** en tareas muy complejas
- **'Maldición de la Dimensionalidad'** (muchas reglas si hay muchas entradas)
- **Defuzzificación** introduce cierta pérdida de información / puede oscurecer

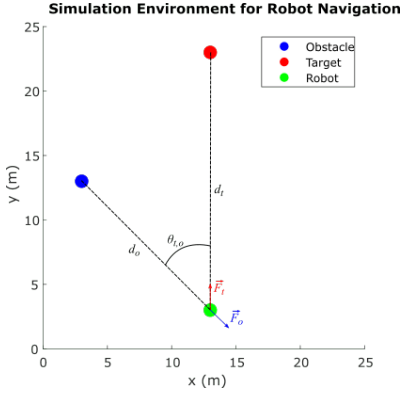
Examples

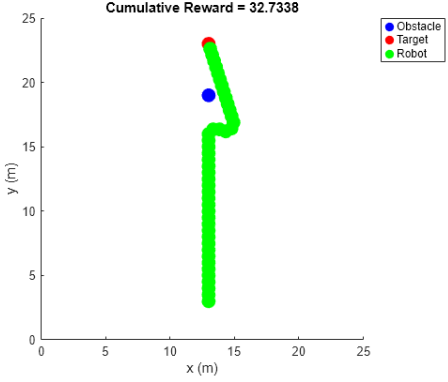
Este ejemplo es para mostrar como un modelo basado en lógica difusa podría 'interpretar' las decisiones tomadas por un modelo tipo 'black box'.



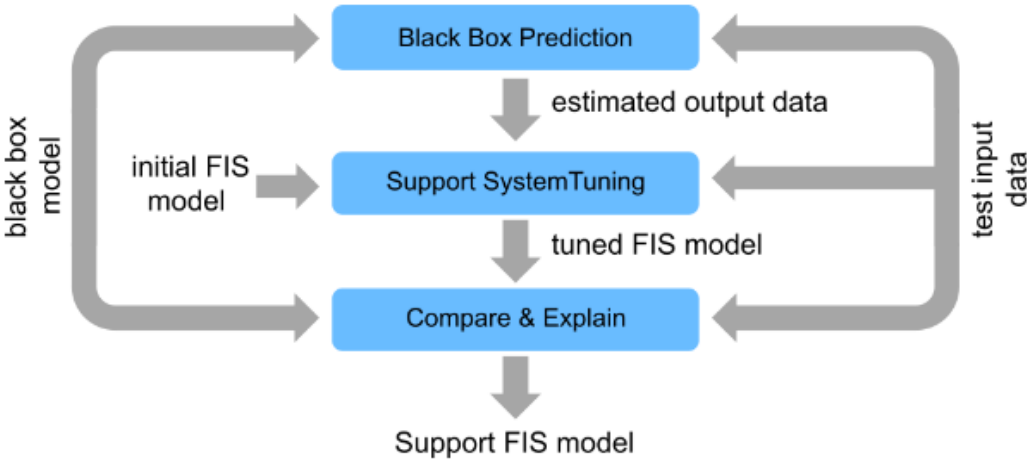
Examples

Este ejemplo es para mostrar como un modelo basado en lógica difusa podría 'interpretar' las decisiones tomadas por un modelo tipo 'black box'.

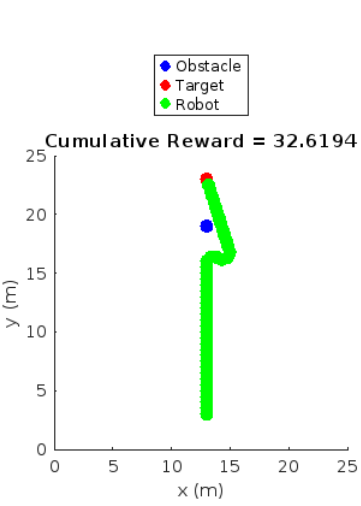


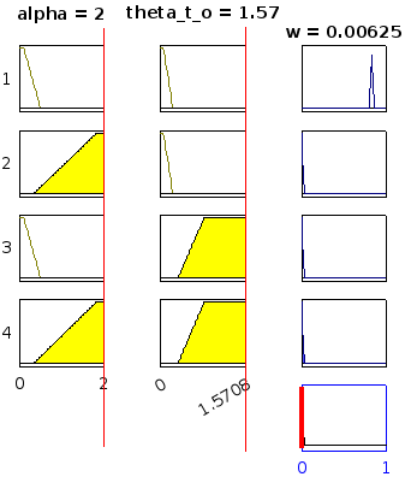


Fuzzy logic models in XAI



Video Ejemplo: Fuzzy XAI system





Fuzzy logic models in XAI

Examples

Este ejemplo es para mostrar como un modelo basado en lógica difusa podría 'interpretar' las decisiones tomadas por un modelo tipo 'black box'.

Retos y Futuro de XAI

- **Integración de Paradigmas:**
  - Sistemas Neuro-Difusos (ej. ANFIS): Aprendizaje + Interpretabilidad.
  - Combinar XAI post-hoc con modelos ante-hoc.
- **Automatización y Aprendizaje:**
  - Aprender reglas difusas y MFs automáticamente desde datos.
- **Calidad y Fidelidad de las Explicaciones:**
  - ¿Cómo medir si una explicación es correcta (fidelidad) y útil? Métricas XAI.
  - Evitar explicaciones engañosas. Robustez de explicaciones.
- **XAI Centrada en el Humano:**
  - Personalización de explicaciones (tipo de usuario, contexto).
  - Interfaces interactivas (what-if, contrafactuales).
  - Estudios de usuario (cómo entienden y usan las explicaciones).

Conclusiones Clave

- La IA moderna presenta un desafío crítico de 'caja negra', limitando confianza, justicia y robustez.
- La **Interpretabilidad y Explicabilidad (XAI)** son cruciales para una IA responsable.
- Existen diversas estrategias:
  - Modelos **intrínsecamente interpretables (ante-hoc)**: Regresión, Árboles simples, Lógica Difusa...
  - Métodos de **explicación post-hoc**: LIME, SHAP...
- La **Lógica Difusa** ofrece alta interpretabilidad semántica (reglas lingüísticas IF-THEN), modelando incertidumbre y conocimiento experto de forma transparente.

Conclusiones Clave

Mensaje Final (Takeaway)

La interpretabilidad debe ser una **consideración central** en el diseño de IA. Explorar y dominar técnicas XAI es fundamental para construir una IA **comprensible, confiable y alineada con nuestros valores**.

¿Preguntas?

¡Gracias por su atención!

Prof. D.Sc. BARSEKH-ONJI Aboud
aboud.barsekh@anahuac.mx