Centro de Investigación en Computación
Instituto Politécnico Nacional

Profundizando en el conocimiento: métodos para explicar los modelos de lenguaje – Parte II

EVCCPLN Segunda Edición

Tallerista: Sergio Arturo Damián Sandoval

Julio 2024

Agenda

1. **Introducción a LIME**
   - Definición
   - Ejemplo
   - Fórmula matemática
   - Demo
   - Limitaciones LIME
2. **SHAP vs LIME**
   - Comparación
3. **Otros métodos de explicabilidad**
4. **Resumen**
5. **Sesión de Preguntas**

Método 2: LIME

1. Introducción a LIME
(Local Interpretable Model-agnostic Explanations)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Mas de 18 mil citas

Código abierto

https://lime.data-imaginist.com/

https://lime-ml.readthedocs.io/en/latest/index.html

https://github.com/marcotcr/lime

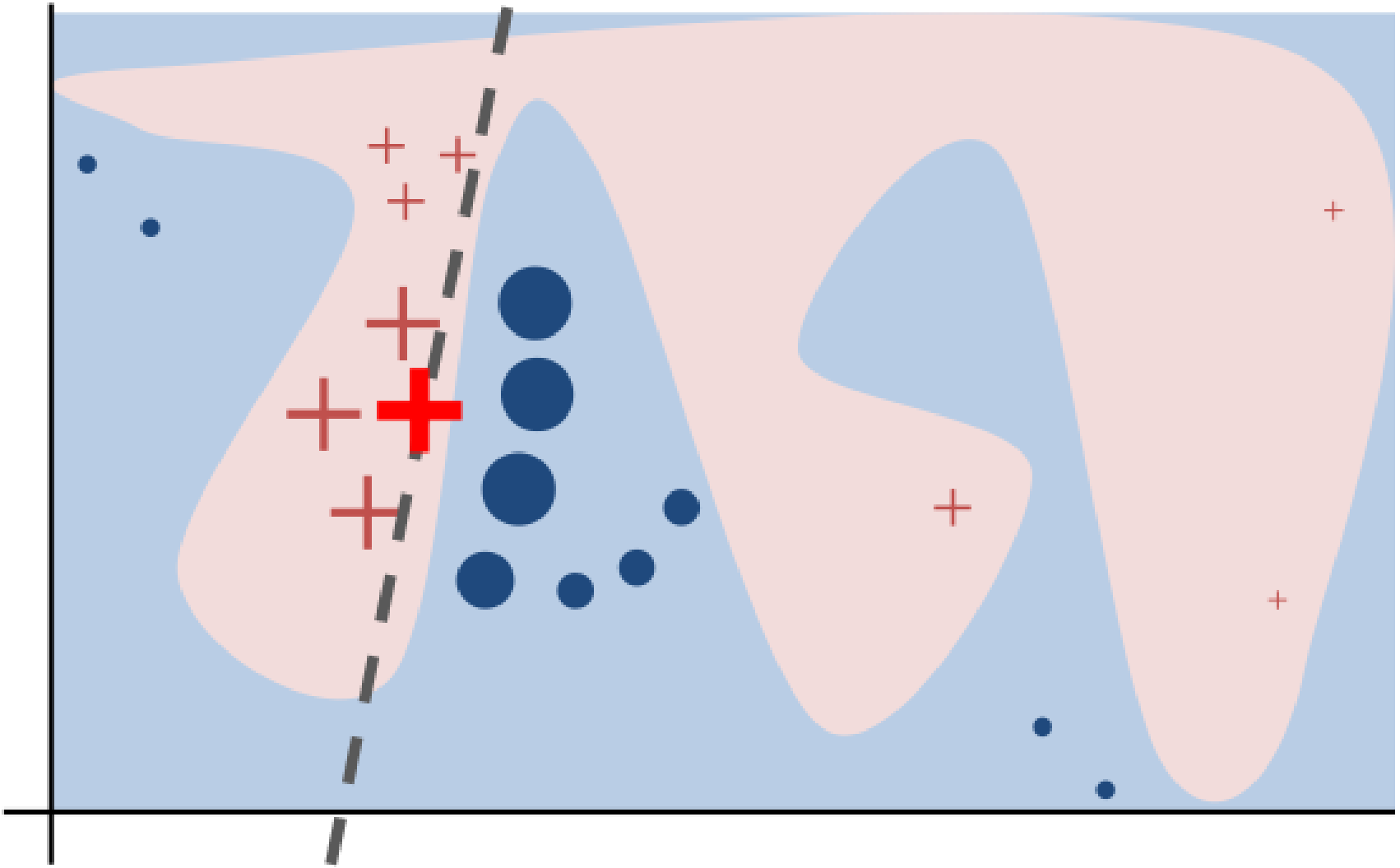
ArXiv: 1602.04938

lime

> There once was a package called lime,
> Whose models were simply sublime,
> It gave explanations for their variations,
> one observation at a time.

lime-rick by Mara Averick

Definición



ArXiv: 1602.04938

La regresión lineal es interpretable

La fórmula de la regresión lineal es sencilla: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

Cada coeficiente β_i representa la influencia de la característica x_i en la predicción

No hay operaciones complejas como en otros modelos

Es fácil de identificar el impacto positivo o negativo de cada característica y en qué medida

Modelos de caja blanca:

- Regresión lineal
- Regresión logística
- Árboles de decisión
- Regresión Ridge y Lasso

Fórmula Matemática

Explicación generada por LIME para la instancia x

Función de pérdida que mide que tan bien el modelo g aproxima las predicciones de f en el vecindario π_x

$$\xi(x) = \arg min_{g \epsilon G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Buscamos el modelo interpretable g dentro de la familia G que minimice la función de pérdida ponderada

Penalización en la complejidad del modelo interpretable g, para asegurar que la explicación sea simple y fácil de entender

$$\pi_x(z) = \exp(-\frac{D(x,z)^2}{\sigma^2})$$

Ejemplo

$x_0 =$ La calidad del producto es excelente y el servicio fue increíble

Perturbaciones:

$x_1 =$ La calidad del producto es _____ y el servicio fue increíble
$x_2 =$ La calidad del producto es excelente y el servicio fue _____
$x_3 =$ La calidad del producto es buena y el servicio fue increíble
$x_4 =$ La calidad del producto es excelente y el servicio fue malo

$$y = \beta_0 + \beta_1 La + \beta_2 calidad + \beta_3 del + \cdots + \beta_{11} increible$$

Explicación de LIME:

“excelente”: +0.4
“increíble”: +0.3
“malo”: -0.3
“buena”: +0.1

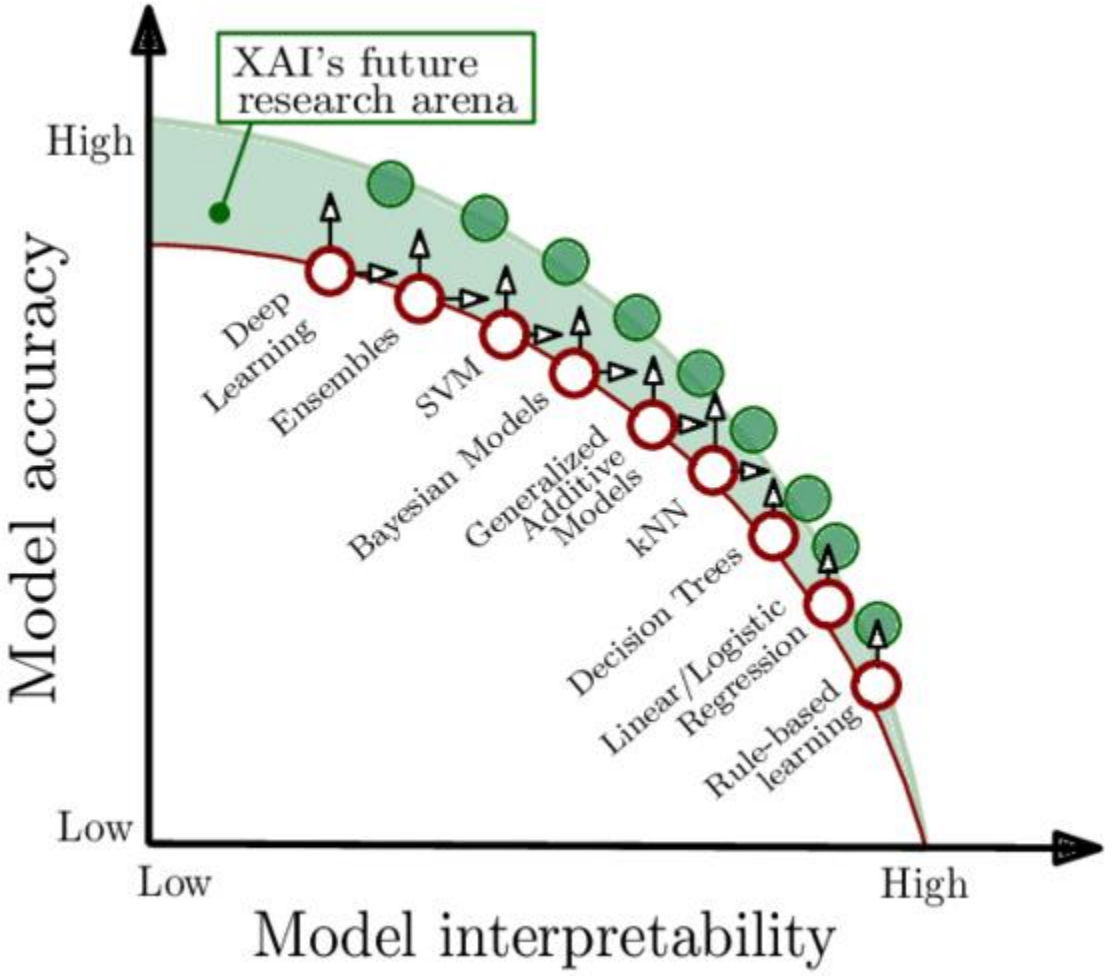
Predicciones del modelo:

$x_0 = 0.9$ (positiva)
$x_1 = 0.8$ (positiva)
$x_2 = 0.7$ (positiva)
$x_3 = 0.75$ (positiva)
$x_4 = 0.4$ (negativa)


Demo

Limitaciones LIME

- La calidad de la explicación depende de la calidad del modelo local
- Puede no capturar adecuadamente interacciones complejas entre palabras o frases
- Las explicaciones pueden ser sensibles a las perturbaciones y variaciones en los datos
- Puede presentar poca consistencia y generar diferentes explicaciones debido a su enfoque en perturbaciones locales
- Eficiencia vs Precisión



2. SHAP vs LIME

SHAP

La interpretación de los valores Shapley puede ser compleja para usuarios no técnicos.

Puede ser computacionalmente costoso.

Teóricamente fundamentado, es una garantía sólida sobre la distribución justa de las contribuciones.

Puede proporcionar explicaciones globales y locales

En teoría no es sensible a perturbaciones (mayor robustez)

LIME

Fácil de implementar y entender. Los resultados son intuitivos y accesibles para usuarios no técnicos.

Las explicaciones están basadas en una aproximación local, por lo que puede no generalizar bien.

Proporciona explicaciones locales.

Sensible a perturbaciones (menor robustez).

Tiene mayor flexibilidad al poder usar diferentes tipos de modelos interpretables.

3. Otros métodos de explicabilidad

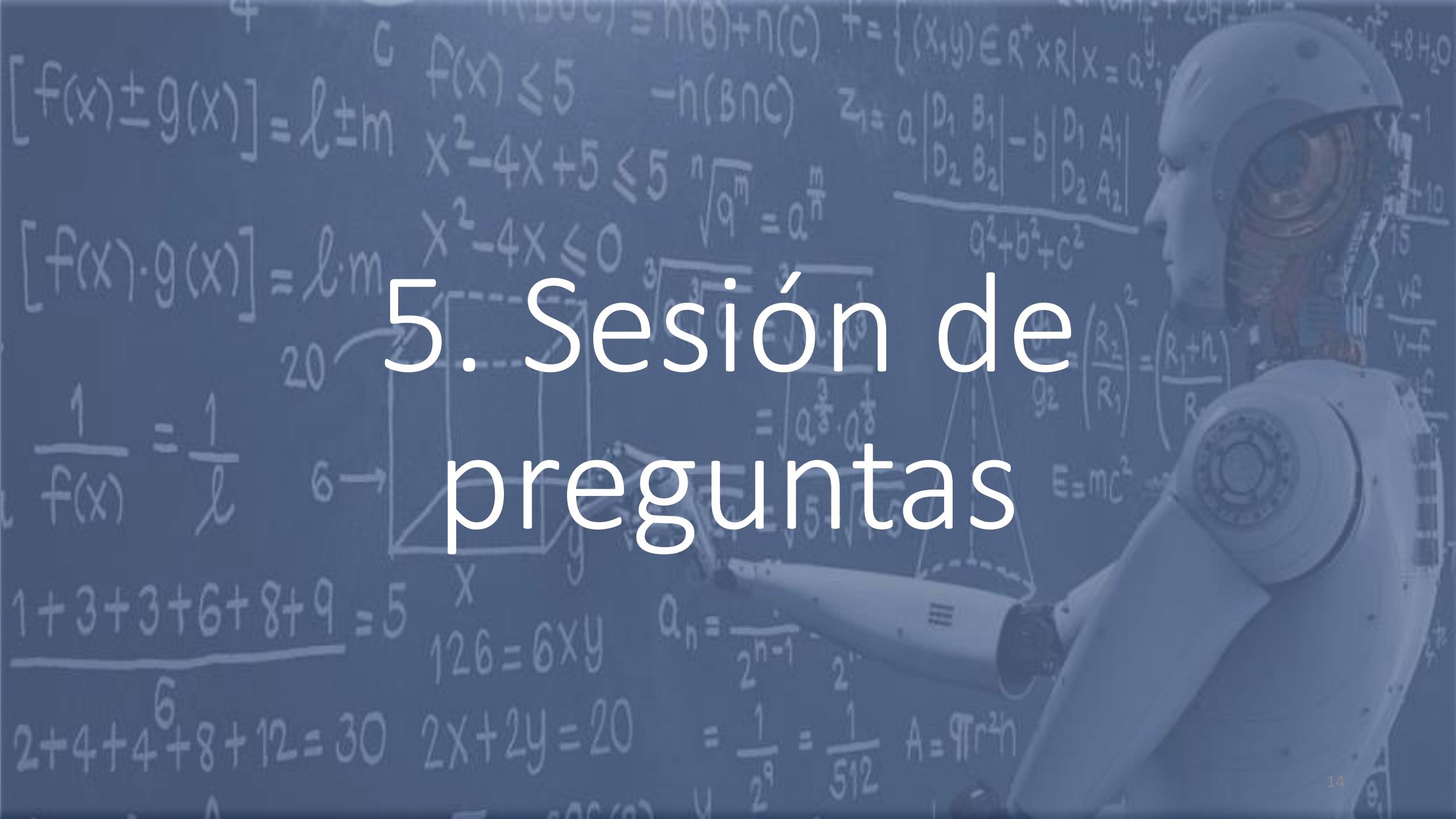
Tipo	Método	Descripción	Ejemplos
Local	Feature Importance	Se asigna un valor a cada característica para determinar su importancia en la predicción del modelo	SHAP, LIME
Local	Rule Based	Traducen el comportamiento del modelo en una serie de reglas simples y comprensibles	Anchors, LORE
Local	Saliency Maps	Muestran partes de una representación visual como las más importantes (CV)	LRP, Integrated Gradients
Local	Prototypes Based	Un objeto que representa el comportamiento del modelo mediante ejemplos representativos o prototipos	Prototype Selection, TracIn
Local	Counterfactuals	Nos muestra una relación de cómo pudo haber cambiado una predición si la entrada se hubiera modificado de cierta manera	DiCE, FACE
Global	Collection of Local Explanations	Agregación de k explicaciones locales	
Global	Representation Based	En lugar de explicar la salida del modelo directamente, se centran en entender cómo las características se representan internamente y cómo estas representaciones afectan las decisiones	Network Dissection, Compositional Explanation
Global	Model Distillation	Transfieren el conocimiento del modelo de caja negra a un modelo más sencillo y interpretable	LGAE, Decision Trees as Global Explanations
Global	Summaries of Conterfactuals	Construcción de una explicación global o resumen mediante el método de Counterfactuals	AReS

https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability

4. Resumen final

- Definición de inteligencia artificial explicable XAI

- En el contexto de la clasificación de textos con modelos de lenguaje:
  - Uso de SHAP como método de explicabilidad
  - Uso de LIME como método de explicabilidad
  - Comparación de SHAP y LIME
  - Ventajas y desventajas de ambos métodos
  - Descripción de otros métodos de explicabilidad



5. Sesión de preguntas

Material adicional