Cálculo de Valores SHAP en Machine Learning Interpretable

1 Introducción

Los valores SHAP (SHapley Additive exPlanations) constituyen un método fundamental en el campo del machine learning interpretable. Estos valores, basados en la teoría de juegos cooperativos, proporcionan una explicación de las predicciones de modelos de machine learning complejos. Este documento presenta una explicación técnica detallada del cálculo de los valores SHAP, incluyendo su base teórica, formulación matemática y ejemplos prácticos de cálculo.

2 Fundamentos Teóricos

2.1 Teoría de Juegos Cooperativos y Valor de Shapley

Los valores SHAP se fundamentan en el concepto del valor de Shapley, introducido por Lloyd Shapley en 1953 en el contexto de la teoría de juegos cooperativos. El valor de Shapley proporciona una manera de distribuir equitativamente tanto las ganancias como los costos entre varios actores trabajando en coalición.

En el contexto del machine learning, se considera cada característica del modelo como un "jugador" en un juego cooperativo, donde la "ganancia" es la predicción del modelo para una instancia específica.

2.2 Propiedades de los Valores SHAP

Los valores SHAP poseen varias propiedades deseables:

1. **Eficiencia**: La suma de los valores SHAP para todas las características es igual a la diferencia entre la predicción del modelo para la instancia actual y la predicción promedio del modelo.

2. **Simetría**: Si dos características contribuyen de manera idéntica a todas las posibles coaliciones, sus valores SHAP serán idénticos.

3. **Dummy**: Una característica que no cambia la predicción del modelo para ninguna coalición tendrá un valor SHAP de cero.

4. **Aditividad**: Para un modelo compuesto por la suma de varios submodelos, el valor SHAP de una característica es la suma de sus valores SHAP en cada submodelo.

3 Formulación Matemática

3.1 Definición Formal

Sea f un modelo de machine learning y x un vector de características de entrada. El valor SHAP ϕ_i para la característica i se define como:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)] \tag{1}$$

Donde:

- N es el conjunto de todas las características
- S es un subconjunto de características
- $f_x(S)$ es el valor esperado del modelo condicionado a las características en S

3.2 Interpretación de la Fórmula

La fórmula anterior calcula el promedio ponderado de las contribuciones marginales de la característica i a todas las posibles coaliciones de características. El peso $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ asegura que todas las coaliciones del mismo tamaño tengan el mismo peso total, y que la suma de todos los pesos sea 1.

4 Métodos de Cálculo

Existen varios métodos para calcular o aproximar los valores SHAP, cada uno con sus propias ventajas y desventajas en términos de precisión y eficiencia computacional.

4.1 Cálculo Exacto

El cálculo exacto de los valores SHAP implica evaluar todas las posibles coaliciones de características, lo que resulta computacionalmente costoso para modelos con un gran número de características. La complejidad temporal es $O(2^{|N|})$, donde $|N|$ es el número de características.

4.2 Aproximación de Kernel SHAP

Kernel SHAP es un método de aproximación que utiliza técnicas de regresión ponderada para estimar los valores SHAP. Este método es más eficiente computacionalmente que el cálculo exacto, especialmente para modelos con un gran número de características.

El algoritmo de Kernel SHAP se puede resumir en los siguientes pasos:

1. Generar un conjunto de coaliciones de características $z' \in \{0,1\}^M$
2. Calcular $\phi = (X^T W X)^{-1} X^T W f(z)$

Donde:

- X es una matriz donde cada fila representa una coalición
- W es una matriz diagonal de pesos
- $f(z)$ son las predicciones del modelo para cada coalición

4.3 TreeSHAP

TreeSHAP es un algoritmo específico para modelos basados en árboles (como árboles de decisión, bosques aleatorios y gradient boosting machines) que permite un cálculo más eficiente de los valores SHAP.

El algoritmo TreeSHAP recorre cada árbol en el modelo una sola vez para calcular los valores SHAP exactos, con una complejidad temporal de $O(TLD^2)$, donde T es el número de árboles, L es el número promedio de hojas por árbol, y D es la profundidad máxima de los árboles.

5 Ejemplo de Cálculo

Consideremos un modelo simple de regresión logística con dos características, x_1 y x_2, para ilustrar el cálculo de los valores SHAP.

Supongamos que el modelo tiene la siguiente forma:

$$f(x) = \sigma(w_1 x_1 + w_2 x_2 + b) \tag{2}$$

Donde σ es la función sigmoide, $w_1 = 1$, $w_2 = 2$, y $b = -1$.

Para una instancia específica ($x_1 = 1, x_2 = 0.5$), calcularemos los valores SHAP.

5.1 Paso 1: Calcular todas las coaliciones posibles

- $f_x(\{\}) = \sigma(-1) \approx 0.269$
- $f_x(\{1\}) = \sigma(1 \cdot 1 - 1) = \sigma(0) = 0.5$
- $f_x(\{2\}) = \sigma(2 \cdot 0.5 - 1) = \sigma(0) = 0.5$
- $f_x(\{1,2\}) = \sigma(1 \cdot 1 + 2 \cdot 0.5 - 1) = \sigma(1) \approx 0.731$

5.2 Paso 2: Calcular las contribuciones marginales

Para x_1:

- $f_x(\{1\}) - f_x(\{\}) = 0.5 - 0.269 = 0.231$
- $f_x(\{1,2\}) - f_x(\{2\}) = 0.731 - 0.5 = 0.231$

Para x_2:

- $f_x(\{2\}) - f_x(\{\}) = 0.5 - 0.269 = 0.231$
- $f_x(\{1,2\}) - f_x(\{1\}) = 0.731 - 0.5 = 0.231$

5.3 Paso 3: Aplicar la fórmula SHAP

Para x_1:

$$\phi_1 = \frac{1}{2}(0.231) + \frac{1}{2}(0.231) = 0.231 \tag{3}$$

Para x_2:

$$\phi_2 = \frac{1}{2}(0.231) + \frac{1}{2}(0.231) = 0.231 \tag{4}$$

5.4 Paso 4: Verificar la propiedad de eficiencia

La suma de los valores SHAP debe ser igual a la diferencia entre la predicción del modelo para la instancia actual y la predicción promedio:

$$\phi_1 + \phi_2 = 0.231 + 0.231 = 0.462 \tag{5}$$

$$f_x(\{1,2\}) - f_x(\{\}) = 0.731 - 0.269 = 0.462 \tag{6}$$

Como podemos observar, la propiedad de eficiencia se cumple.

6 Interpretación de los Valores SHAP

Los valores SHAP proporcionan una medida de la importancia de cada característica para una predicción específica. Un valor SHAP positivo indica que la característica aumenta la predicción del modelo, mientras que un valor negativo indica que la disminuye.

En nuestro ejemplo, ambas características tienen un valor SHAP positivo e igual, lo que significa que ambas contribuyen de manera igual y positiva a la predicción final del modelo para esta instancia específica.

7 Ventajas y Limitaciones

7.1 Ventajas

- **Consistencia**: Los valores SHAP son consistentes, lo que significa que si un modelo cambia de tal manera que la contribución de una característica aumenta o se mantiene igual para todas las posibles entradas, el valor SHAP de esa característica no disminuirá.
- **Localidad**: Los valores SHAP proporcionan explicaciones locales, es decir, específicas para cada predicción individual.
- **Globalidad**: Al promediar los valores SHAP absolutos sobre múltiples instancias, se puede obtener una medida global de la importancia de las características.
- **Modelo-agnóstico**: Los valores SHAP pueden aplicarse a cualquier tipo de modelo de machine learning.

7.2 Limitaciones

- **Costo computacional**: El cálculo exacto de los valores SHAP puede ser computacionalmente costoso, especialmente para modelos con un gran número de características.
- **Independencia de las características**: Los valores SHAP asumen que las características son independientes, lo cual no siempre es cierto en la práctica.
- **Sensibilidad a la multicolinealidad**: En presencia de multicolinealidad, los valores SHAP pueden ser inestables.
- **Interpretación cuidadosa**: La interpretación de los valores SHAP requiere un entendimiento cuidadoso del contexto del problema y del funcionamiento del modelo.

8 Conclusión

Los valores SHAP representan una herramienta poderosa en el campo del machine learning interpretable, proporcionando una manera de explicar las predicciones de modelos complejos de una manera consistente y teóricamente fundamentada. Su cálculo, aunque puede ser computacionalmente costoso, ofrece insights valiosos sobre la importancia de las características tanto a nivel local como global.

Sin embargo, como con cualquier herramienta de interpretación de modelos, es crucial utilizar los valores SHAP en conjunto con otras técnicas y siempre en el contexto del problema específico que se está abordando. La interpretación cuidadosa y la consideración de las limitaciones de los valores SHAP son esenciales para su aplicación efectiva en la práctica.