Data valuation

UNIVERSIDAD COMPLUTENSE DE MADRID
UNIVERSIDAD POLITÉCNICA DE MADRID

TRABAJO FINAL DE MÁSTER

22 de agosto de 2023

Autor: Francisco AGUILAR MARTÍNEZ
Tutores: Carlos GREGORIO RODRÍGUEZ
Miguel DE BENITO DELGADO

UNIVERSIDAD
COMPLUTENSE
MADRID

Data valuation

UNIVERSIDAD COMPLUTENSE DE MADRID
UNIVERSIDAD POLITÉCNICA DE MADRID

TRABAJO FINAL DE MÁSTER

22 de agosto de 2023

Autor: Francisco AGUILAR MARTÍNEZ
Tutores: Carlos GREGORIO RODRÍGUEZ
Miguel DE BENITO DELGADO

Índice general

CAPÍTULO 1

Estado del arte

Desde la aparición del valor de Shapley como un método de reparto justo de recompensas en juegos cooperativos [1], este concepto se ha utilizado en diversos campos como la economía [2], estudio de sistemas multiagente [3] e incluso en áreas en las que su aplicación puede resultar menos evidente como la genética [4]. Esta versatilidad se debe, en parte, a su sólida base matemática y a sus intuitivas interpretaciónes, entre las que podemos resaltar:

- Pago justo: El valor de Shapley de un jugador es la cantidad que este debería recibir si las recompensas se distribuyesen de manera que los jugadores fueran recompensados en función de su contribución a la recompensa total.
- Poder de negociación: El valor de Shapley puede interpretarse como una medida del poder de negociación de un jugador. Un jugador tiene más poder de negociación si su ausencia causa una mayor disminución en la recompensa total que se puede obtener.

El valor de Shapley se basa en una serie de axiomas fundamentales, que garantizan propiedades como su equidad, eficiencia y simetría. En economía relajar estos axiomas, con el objetivo de dar lugar a nuevas formas de reparto de recomenpensa, ha sido uno de los principales temas de estudio. Ejemplos de esto serían el concepto de semivalor, el cual se obtiene al eliminar el axioma de eficiencia [5, 6]. Este axioma asegura que la suma de los valores de todos los jugadores sea igual a la recompensa total disponible. Por tanto, al suprimirlo, los semivalores permiten cierta flexibilidad en este aspecto, lo que puede ser útil en situaciones en las que no todos los beneficios se pueden distribuir. Del mismo modo, eliminar el axioma de simetría, que establece que dos jugadores con igual contribución deben recibir igual recompensa, lleva al valor de Banzhaf [7], que proporciona una medida de poder de un jugador basada en cuánto puede cambiar el resultado de un juego al unirse o abandonar una coalición. Cabe destacar que tanto el valor de Shapley como el valor de Banzhaf pueden obtenerse como particularizaciones del concepto de semivalor.

Debido a la naturaleza combinatoria del valor de Shapley, el cálculo de este es altamente costoso a nivel computacional y resulta en una tarea cuya complejidad crece exponencialmente al aumentar el número de jugadores. Es por esto que surgen métodos de estimación del mismo, la mayoría de estos métodos se basan en técnias de Montecarlo. En 1960, Irwin Mann y el propio Shapley mencionan las estimaciones basada en muestreo de permutaciones [8]. Pero no es hasta 2015 que se lleva a cabo un análisis de la complejidad a la hora de muestrear usando dicha técnica [9]. Tras esto, en 2019, Covert propone un nuevo método de estimación basado en la técnica de muestreo por importancia que mejora los actuales [10]. Cabe destacar el *ApproShapley* propuesto en [11] desarrollado por los profesores J.Castro, D.Gómez y J.Tejada de la UCM.

Una de las primeras apariciones del valor de Shapley en el campo del aprendizaje computacional data del año 2005 como un método de selección de variables [12]. Más tarde, en 2017, se utiliza en el diseño del marco SHAP [13], enfocado en la evaluación de la importancia de variables en modelos de predicción. Sin embargo, no es asta el año 2019 en el que se introduce como una alternativa a los métodos de valoración de datos del momento [14], acuñándose así el concepto *Data Shapley*.

Al igual que el cálculo del valor de Shapley, el cálculo de *Data Shapley* es altamente costoso a nivel computacional, por lo que surgen también varios métodos de aproximación, entre los que podemos destacar *Group Testing* [15], métodos de aproximación y cálculo exacto para problemas en los que se aplican métodos como KNN o derivados de este [16] y diversas técnicas basadas en métodos de Montecarlo como las vistas en [14].

Cuando se prueba la eficacia de *Data Shapley* en problemas de aprendizaje computacional como la detección de outliers o datos corruptos [14], la investigación sigue el mismo camino que años antes en teoría de juegos y se empiezan a reciclar conceptos como los semivalores o el valor de Banzhaf. Es en la línea de los semivalores que en 2022 surge *Beta Shapley* [17], una generalización de *data Shapley* que supera los resultados de los métodos más actuales de valoración de datos en varias tareas como son detección de muestras mal etiquetadas y selección de puntos problemáticos a la hora de entrenar un modelo.

En 2023, aparece *data Banzhaf* [18], un nuevo método de valoración de datos derivado del valor de Banzhaf. Este nuevo método surge como una solución a la falta de robustez de las herramientas de valoración de datos existentes, falta de robustez causada en parte por factores difíciles de controlar como la aleatoriedad del método del descenso del gradiente estocástico, el cual es ampliamente usado hoy en día. Para solventar esta falta de robustez se apoya en el concepto de *Safety Margin*, y demuestra que el valor de Banzhaf es el semivalor con mayor *Safety Margin*. *Data Banzhaf* supera a los existentes métodos de valoración basados en semivalores en varias tareas de aprendizaje automático.

Aunque en este trabajo nos centramos en métodos de valoración de datos que se derivan directamente de conceptos de la teoría de juegos, existen otros métodos que, aún utilizando teoría de juegos, siguen enfoques relativamente distintos. Podemos destacar algunas obras como [19] en la que sugieren un método de valoración de datos para modelos generativos que utiliza la discrepancia media máxima (MMD) entre la fuente de datos y la distribución real de datos. En [20] proponen una medida de diversidad, llamada volumen robusto (RV), para valorar las fuentes de datos. La robustez de RV se discute en términos de la estabilidad frente a la replicación de datos. Finalmente en [21] utilizan diferencias estadísticas entre los datos de origen y un conjunto de datos de referencia como la métrica de valoración. Estas diferencias estadísticas se miden mediante el uso de los conceptos de diversidad y relevancia de los datos previamente comentados.

Como algo casi anecdótico hay literatura en la que se lleva a cabo valoración de datos mediante aprendizaje por refuerzo [22]. En dicho trabajo, se utiliza una red neuronal profunda para obtener un estimador de la probabiliadad de cada dato de ser usado en el entrenamiendo del modelo de predicción. Este estimador se obtiene mediante aprendizaje por refuerzo.

CAPÍTULO 2

Marco teórico

Introducción

aaa

2.1. Teoría de juegos

Definición y conceptos básicos

La *teoría de juegos* puede ser entendida como la rama de las matemáticas que analiza situaciones en las que el resultado para cada jugador o participante depende no solo de sus propias decisiones, sino también de las tomadas por otros jugadores. Estas situaciones se conocen como *juegos*.

Definición 2.1.1. Un *juego* es una interacción entre jugadores racionales[1], mutuamente conscientes, en la que las decisiones de un jugador impactan en las ganancias de otros. Un juego se define por:

- Los jugadores que intervienen. Cada *jugador* es un agente que tiene a su disposición diversas estrategias basadas en las posibles recompensas que podría recibir.
- Las estrategias disponibles para cada jugador. Una *estrategia* es un plan de acción que un jugador puede adoptar dentro de un juego. Esta estrategia dicta las acciones que tomará en cada situación posible que se presente en el juego.
- Las ganancias de cada jugador en función de los resultados. Las *ganancias* son representaciones de las motivaciones de los jugadores, pudiendo representar beneficios, cantidades, o simplemente reflejar la conveniencia de los diferentes desenlaces.

[1] Entendemos la racionalidad como el hecho de que cada jugador intenta maximizar su propio beneficio.

En el contexto de teoría de juegos aplicada a la valoración de datos, nos centraremos en juegos cooperativos.

Definición 2.1.2. Un *juego cooperativo* es aquel en el que los jugadores pueden comunicarse y negociar con el fin de establecer acuerdos vinculantes.

Los acuerdos mencionados se denominan *coaliciones*. Se trata de grupos de jugadores que eligen actuar juntos para lograr un objetivo común. Las coaliciones pueden variar desde la compuesta por todos los jugadores hasta la que incluye a un único jugador.

Un juego cooperativo queda completamente definido por el conjunto de sus jugadores N y por su función característica. Esta función es el nexo entre la formación de coaliciones y los beneficios obtenidos por los jugadores, ya que asigna a cada coalición posible un beneficio que puede lograr.

Definición 2.1.3. Una *función característica* v es una función

$$\begin{aligned} v : 2^N &\longrightarrow \mathbb{R} \\ S &\longmapsto v(S). \end{aligned}$$

Asignando a cada posible coalición la máxima utilidad que los jugadores de S pueden obtener, independientemente de lo que haga el resto de jugadores.

Una vez formadas las coaliciones y determinadas las ganancias, surge la pregunta: ¿cómo se dividen las ganancias entre los miembros de la coalición? Varias respuestas a esta pregunta han surgido en la literatura, pero el valor de Shapley es uno de los conceptos más prominentes.

2.1.1. El valor de Shapley

El *Valor de Shapley* es un concepto de teoría de juegos que asigna de manera equitativa las ganancias entre los miembros de una coalición. Propuesto por Lloyd Shapley en 1952 [1],y se fundamenta en los siguientes axiomas:

- Simetría. Si dos jugadores son simétricos, es decir, si su contribución a cualquier coalición es la misma, entonces presentan el mismo valor.

$$\text{Si } \forall S \subseteq N \setminus \{i,j\},\ v(S \cup \{i\}) = v(S \cup \{j\}) \implies \phi(i;v) = \phi(j;v).$$

- Eficiencia. El valor total producido por la coalición conformada por todos los jugadores se distribuye entre los jugadores. Es decir $v(N) = \sum_{i \in N} \phi(i;v)$.
- Linealidad. Dados dos juegos v y w, el valor del juego $v + w$ es la suma de los valores de cada juego. Es decir, $\phi(i; v + w) = \phi(i; v) + \phi(i; w)$.
- Jugador nulo. Los jugadores que no aportan a ninguna coalición tendrán valor nulo. Es decir,

$$\text{Si } \forall S \subseteq N \setminus \{i\},\ v(S \cup \{i\}) = v(S) \implies \phi(i;v) = 0.$$

El siguiente teorema, extraído de [1] prueba que se trata del único método de valoración de datos satisfaciendo estos axiomas.

Teorema 2.1.4. *Existe una única función ϕ satisfaciendo los axiomas de simetría, eficiencia, linealidad y jugador nulo, y viene dada por la fórmula:*

$$\phi(i;v) = \sum_{S \subseteq N} \frac{|S| - 1!(|N| - |S|)!}{|N|!}(v(S) - v(S \setminus \{i\}))$$

Demostración. La demostración detallada se encuentra en [1]. □

Proposición 2.1.5. *El valor de Shapley puede ser expresado como:*

$$\phi(i;v) = \frac{1}{n}\sum_{k=1}^{n}\binom{n-1}{k-1}^{-1} \sum_{S \subseteq N \setminus \{i\}, ||S||=k-1} (v(S \cup \{i\}) - v(S))$$

Demostración. Pendiente. □

Como observación final, es importante señalar que se toman en cuenta todas las coaliciones que no incluyen al jugador en cuestión. Para cada una de esas coaliciones, se determina la *contribución marginal* del jugador, la cual representa la diferencia entre la ganancia de la coalición con y sin el jugador. El Valor de Shapley, en última instancia, se calcula como el promedio de todas estas contribuciones marginales.

2.1.2. Semivalores

Los semivalores son una generalización del valor de Shapley, que surgen al relajar el axioma de eficiencia. Mientras que el Valor de Shapley garantiza que la suma de los valores individuales de los jugadores sea igual a la utilidad total generada por la coalición total, los semivalores no imponen esta restricción. Así, permiten una variedad más amplia de métodos de valoración en el estudio de juegos cooperativos, siendo especialmente útiles en escenarios donde no es necesario o deseado que la totalidad de la recompensa sea distribuida.

La clave de los semivalores es que asignan pesos a las coaliciones basándose en el tamaño de estas. Estos pesos son utilizados para calcular la contribución marginal promedio de cada jugador.

A diferencia del valor de Shapley, que es único dada su definición basada en axiomas [1], hay múltiples semivalores posibles, dependiendo de cómo se determinen los pesos. Esta variabilidad en los semivalores queda formalizada en el siguiente teorema:

Teorema 2.1.6. *Representación de semivalores [6].*

Una función ϕ es un semivalor, si y solo si, existe una función peso $w : [n] \to \mathbb{R}$ tal que

$$\sum_{k=1}^{n}\binom{n-1}{k-1}w(k) = n$$

, que permite representar ϕ mediante la expresión:

$$\phi(i;v) = \sum_{k=1}^{n} \frac{w(k)}{n} \sum_{S \subseteq N \setminus \{i\}, ||S||=k-1} (v(S \cup \{i\} - v(S)))$$

Es relevante notar que el valor de Shapley es un caso particular de semivalor, donde los pesos son determinados por la fórmula: ($w_{Shapley} = \binom{n-1}{k-1}^{-1}$).

2.1.3. El valor de Banzhaf

El *Valor de Banzhaf*, también denominado índice de poder de Banzhaf[7], es una métrica en la teoría de juegos cooperativos que busca cuantificar el poder e influencia de un jugador dentro de una coalición. Este índice fue introducido por John F. Banzhaf III en 1965, con la finalidad de ofrecer una herramienta analítica que pudiese determinar el poder de influencia de un jugador, especialmente en escenarios de votación ponderada.

Definición 2.1.7. El *valor de Banzhaf* de un jugador i en un juego cooperativo v viene dado por la expresión:

$$\phi_{Banzhaf}(i;v) = \frac{1}{2^{n-1}} \sum_{S \subseteq N \setminus \{i\}} [v(S \cup \{i\}) - v(S)] \tag{2.1}$$

Para comprender mejor la esencia y la utilidad del valor de Banzhaf, es fundamental familiarizarse con ciertos conceptos relacionados:

- *Sistema de votación ponderado*: Es un sistema de votación en el que cada jugador tiene un peso o poder de voto particular. Para que una propuesta sea aprobada, la suma de los pesos de los jugadores que votan a favor debe superar un umbral o cuota establecida.
- *Jugador pivote*: Se considera que un jugador actúa como pivote si, al modificar su voto de negativo a positivo, la propuesta es aprobada. Sin embargo, si se abstuviera o mantuviera su voto en contra, la propuesta sería rechazada.
- *Índice de poder de Banzhaf* Este índice mide la frecuencia con la que un jugador se convierte en pivote. Es importante destacar que el poder de un jugador no siempre es directamente proporcional a su peso en la votación.

A pesar de sus similitudes con el valor de Shapley, el índice de Banzhaf se diferencia en su enfoque y en la forma de asignar poder a los jugadores. Mientras que el valor de Shapley se basa en contribuciones marginales promediadas, el índice de Banzhaf se enfoca en la capacidad de un jugador de influir en el resultado final de una votación. Específicamente, es un semivalor con un peso asociado dado por $w_{Banzhaf} = \frac{1}{2^{n-1}}$.

2.2. Valoración de datos

Hoy en día, el dato representa uno de los recursos más valiosos en el mundo para negocios, gobiernos y particulares. La toma de decisiones basada en datos está presente en casi todos los ámbitos de la sociedad, desde la medicina predictiva hasta la publicidad personalizada. Debido a esto, la habilidad para determinar el valor de un dato se ha vuelto indispensable. Es aquí donde entra en juego la valoración de datos.

Cuando nos referimos al valor de un dato, es esencial comprender que dicho valor no es unidimensional. Un dato puede ser valorado desde diferentes perspectivas y categorizado basado en diversas cualidades:

1. **Valor Intrínsico vs. Extrínseco**:
   - **Valor intrínseco**: Se refiere al valor inherente al propio dato, basado en su precisión y calidad. Este valor es independiente del uso que se le dé al dato.
   - **Valor extrínseco**: Se corresponde al valor que se le atribuye al dato en función de su uso. Depende, pues, del contexto en el que se use el dato.
2. **Valor Directo vs. Indirecto**:
   - **Valor directo**: Alude al beneficio inmediato que se obtiene de un dato, como podría ser al venderlo.
   - **Valor indirecto**: Se refiere al beneficio derivado del uso estratégico del dato.

En este trabajo nos centraremos en estudiar el valor extrínseco e indirecto de los datos. Esta investigación nos permitirá, posteriormente, discernir el valor directo de los datos y detectar posibles problemas en su valor intrínseco.

A pesar de que la valoración de datos es un concepto multifacético que depende de varios factores, nos ajustaremos al enfoque propuesto por diversas fuentes[14, 17], el cual se compone de tres elementos esenciales:

1. Denominaremos N al **conjunto prefijado de datos de entrenamiento**, siendo $N = \{(x_i, y_i)\}_1^n$. Aquí, x_i hace referencia a las características del dato i-ésimo, e y_i a su categoría en problemas de clasificación o su valor en problemas de regresión.
2. El **algoritmo de aprendizaje** $\mathcal{A}$, será tratado como una caja negra que toma un conjunto de entrenamiento N y genera un predictor f.
3. La **función de utilidad** v es una aplicación que asigna a cada subconjunto de N un valor, reflejando la utilidad de ese subconjunto. Para problemas de clasificación, la opción común para v es la precisión del modelo entrenado con el subconjunto dado, es decir $v(S) = acc(\mathcal{A}(S))$. Sin pérdida de generalidad asumiremos a lo largo del documento que $v(S) \in [0, 1]$ para cualquier $S \subseteq N$.

Por lo tanto, podemos concebir la valoración de datos como el proceso de asignar un valor a cada dato del conjunto N, reflejando su contribución en el entrenamiento del modelo. Cada uno de estos valores estará determinado por $N, \mathcal{A}$ y v, pero por simplicidad, lo expresaremos como $\phi(i; v)$. A estas puntuaciones se les denomina *data values*.

Pasamos a tratar las nociones principales en cuanto a valoración de datos.

2.2.1. Métricas de valoración de datos

LOO Error

El método más sencillo para valorar datos consiste en medir la contribución de un punto individual al desempeño global del conjunto de entrenamiento:

$$\phi_{loo}(i; v) = v(N) - v(N \setminus i).$$

Este método es conocido como *leave-one-out*(LOO). Para calcular el valor exacto de los valores LOO para un conjunto de entrenamiento de tamaño N, sería necesario reentrenar el modelo N veces. Este procedimiento resulta poco práctica cuando el tamaño del conjunto de datos es considerablemente grande, como se puede observar en estudios tales como [15]

Data Shapley

En la sección citarsecciónde lvalorDeShapley se introdujo
En [1] definen los métodos de valuación equitativa como sigue.

Definición 2.2.1. Un método de evaluación será equitativo si cumple las siguientes condiciones:

1. Linealidad: Dadas métricas de error V y W, constantes
2. Jugador nulo: Si $\forall S \subseteq N \setminus \{i\}$, $V(S) = V(S \cup \{i\})$ entonces $\phi_i = 0$.
3. Simetría: Fijados $i, j \in N$, si para todo $S \subseteq N \setminus \{i, j\}$, se tiene que $V(S \cup \{i\}) = V(S \cup j)$ entonces $\phi_i = \phi_j$.
4. Esta hay que escribirla bien.

La siguiente proposición es la que da nombre a ...

Proposición 2.2.2. *Cualquier $\phi(N, \mathcal{A}, V)$ que satisfaga las condiciones anteriores será de la forma*

$$\phi_i = C \sum_{S \subseteq N \setminus \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{n-1}{|S|}}.$$

Dónde el sumatorio contempla todos los subconjuntos de N que no contienen a i y C es una constante arbitraria. Llamaremos a ϕ_i el Valor de Shapley asociado al dato i.

Demostración. Se puede consultar en [1]. □

Beta Shapley

En [17] proponen un semivalor concreto... En este artículo se propone utilizar la función β con un par de parámetros positivos (α, β) para definir un semivalor, que vendrá dado por la siguiente función de peso

jeje

We propose to use ψ semi(z*; U, D, w(n) α, β) and call it Beta(α, β)-Shapley value. The pair of hyperparameters (α, β) decides the weight distribution on [n]. For instance, when (α, β) = (1, 1), the normalized weight w(n) α, β (j) : = (n-1 j-1)w(n) α, β (j) = 1 for all j in [n], giving the uniform weight on marginal contributions , i.e., Beta(1,1)-Shapley ψsemi(z*; U, D, w(n) 1,1) is exactly the original data Shapley. Figure 3 shows various weight distributions for different pairs of (α, β). For simplicity, we fix one of the hyperparameter to be one. When $\alpha \geq \beta = 1$, the normalized weight assigns large weights on the small cardinality and remove noise from the large cardinality. Conversely, Beta(1,β) puts more weights on large cardinality and it approaches to the LOO as β increases.

Banzhaf

Este concepto será el central a lo largo de nuestro trabajo. Se define igual que su precursor, el valor de Banzhaf descrito en la sección...

El interés particular de este concepto es que será, de todos los semivalores el que alcanza un mayor *safety margin*, concepto que introduciremos en la sección...

2.3. Midiendo la robustez

En diversas situaciones, como la selección de datos, el orden de los *data values* es lo que aporta valor[17]. Un ejemplo podría ser el filtrado de datos de baja calidad. El escenario ideal, sería aquel en el que incluso estando perturbada la función de utilidad, se preserva el mismo orden de *data values*.

En este contexto, la robustez alude a la resistencia de los métodos de valoración de datos ante perturbaciones o ruido. Del mismo modo que un modelo de aprendizaje robusto debería resistir entradas ruidosas, un método de valoración de datos robusto debería conservar el orden de los *data values* a pesar del ruido intrínseco de los algoritmos de aprendizaje automático.

Ahora, vamos a establecer los conceptos requeridos para formalizar y medir la robustez. Recordemos que un semivalor está determinado por su función de peso w. Así, definimos la diferencia escalada como:

Definición 2.3.1. Sean i y j puntos de N. La diferencia entre los semivalores $\phi(i;v)$ y $\phi(j;v)$ se define como:

$$\begin{aligned} D_{i,j}(v,w) &:= n(\phi_w(i;v) - \phi_w(j;v)) \\ &= \sum_{k=1}^{n-1} (w(k) + w(k+1)) \binom{n-2}{k-1} \Delta_{i,j}^k(v). \end{aligned}$$

Donde $\Delta_{i,j}^k(v) := \binom{n-2}{k-1}^{-1} \sum_{|S|=k-1, S\subseteq N\setminus\{i,j\}} (v(S\cup i) - v(S\cup j))$, representa la distinguibilidad promedio entre i y j en subconjuntos de tamaño k usando una función de utilidad sin ruido v.

Consideremos $\hat{v}$ un estimador de v. Sabemos que $\hat{v}$ y v generarán diferentes *data values* para un par de puntos i y j si, y solo si, $D_{i,j}(v,w)D_{i,j}(\hat{v},w) \leq 0$.

Se podría pensar inicialmente en definir la robustez de un semivalor como la menor cantidad de ruido $||\hat{v} - v||$ que alteraría el orden de los *data values*. Sin embargo, una definición así dependería de la función de utilidad original v. Si la función original v no es capaz de diferenciar dos puntos i y j ($\Lambda_{i,j}^{(k)}(v) \simeq 0$, para todo $k = 1,\ldots,n-1$), entonces $D_{i,j}(v,w)$ será casi 0, y cualquier mínima perturbación podría modificar el orden entre $\phi(i;v)$ y $\phi(j;v)$.

Por ello, para definir de formar razonable la robustez de un semivalor, debemos considerar solo las funciones de utilidad que sean capaces de distinguir entre i y j.

Definición 2.3.2. Distinguibilidad Diremos que un par de puntos (i,j) son τ-distinguibles por la función de utilidad v si, y solo si, $\Lambda_{i,j}^{(k)}(v) \geq \tau$ para todo $k \in \{1,\ldots,n-1\}$.

Sea ahora $\mathcal{V}_{i,j}^{(k)}$ el conjunto de todas las funciones de utilidad v que son capaces de τ-distinguir (i,j). Usando la definción anterior, podemos caracterizar la robustez de un semivalor mediante su *safety margin*, que representa la menor cantidad

de ruido $||\hat{v} - v||$ que, al añadirse, invertiría el orden de los *data values* de al menos un par de puntos (i, j), para al menos una función de utilidad $v \in \mathcal{V}_{i,j}^{(k)}$.

Definición 2.3.3. Safety margin Dado $\tau > 0$, definimos el *safety margin* de un semivalor para un par de puntos (i, j) como:

$$\text{Safe}_{i,j}^{(k)}(\tau; w) := \min_{v \in \mathcal{V}_{i,j}^{(\tau)}} \min_{\hat{v} \in \{\hat{v}: D_{i,j}(v;w) D_{i,j}(\hat{v};w) \leq 0\}} ||\hat{v} - v||.$$

El *safety margin* de un semivalor es:

$$\text{Safe}(\tau; w) := \min_{i,j \in N, i \neq j} \text{Safe}_{i,j}(\tau; w).$$

La intuición detrás del *safety margin* es que muestra la máxima cantidad de ruido que puede ser añadida a un semivalor sin que se altere el orden de los *data values* de ningún par de puntos que fuera distinguible por la función de utilidad original.

2.3.1. Robustez del valor de Banzhaf

Los resultados aquí mostrados pertenecen a la sección 4 de [18].

Teorema 2.3.4. *Para cualquier $\tau > 0$, el valor de Banzhaf alcanza el mayor safety margin* ,

$$\textit{Safe}(\tau; w_{\textit{Banzhaf}}) = \frac{\tau}{2^{\frac{n}{2}-1}}.$$

De entre todos los semivalores.

Demostración. Consultar... □

Intuitivamente, este resultado se debe a cómo los semivalores asignan diferentes pesos en función del tamaño de los subconjuntos evaluados. Así, es posible construir una perturbación de la función de utilidad que maximice la influencia sobre el semivalor correspondiente, introduciendo ruido en los subconjuntos con mayor peso asignado. De ahí que la estrategia óptima para robustecer sea asignar pesos uniformes a todos los subconjuntos, tal como lo hace el valor de Banzhaf.

Además, se puede demostrar que el valor de Banzhaf es el semivalor más robusto en el sentido de que el ruido de la utilidad afecta mínimamente a los cambios en los *data values*. En concreto, el valor de Banzhaf alcanza la menor constante de Lipschitz L tal que $||\phi(v) - \phi(\hat{v})|| \leq L||v - \hat{v}||$ para todos los posibles pares de funciones de utilidad v y $\hat{v}$.

Teorema 2.3.5. *El valor de Banzhaf, con $w(k) = \frac{n}{2^{n-1}}$, logra la menor constante de Lipschitz, $L = \frac{1}{2^{\frac{n}{2}-1}}$ de entre todos los semivalores.*

Demostración. Consultar apéndice C.4 de [18]. □

2.3.2. Estimación eficiente

Dado que es prácticamente imposible calcular los *data values* exactos en métodos de valoración de datos basados en semivalores, debido a la necesidad de un número exponencial de evaluaciones de la función de utilidad, se debe recurrir a métodos de aproximación.

A continuación, introducimos el concepto de error asociado a un estimador.

Definición 2.3.6. Un estimador de un semivalor $\hat{\phi}$ es una (ϵ, δ)-aproximación del semivalor ϕ en norma l_p si, y solo si,

$$P_{\hat{\phi}}[||\hat{\phi} - \phi)||_p \leq \epsilon] \geq 1 - \delta.$$

Donde la aleatoriedad se da en la construcción del estimador.

2.3.3. Estimador Simple de Montecarlo

El valor de Banzhaf puede ser reformulado como:

$$\phi_{Banzhaf}(i; v) = \mathcal{E}_{S \sim \text{Unif}(2^{N \setminus i})}[v(S \cup \{i\}) - v(S)]. \tag{2.2}$$

A partir de esto, un método de Montecarlo directo para estimar $\phi_{Banzhaf}(i; v)$ consistiría en generar muestras uniformes de $\mathcal{S}_i \subset 2^{N \setminus \{i\}}$ y calcular:

$$\hat{\phi}_{MC}(i; v) = \frac{1}{|\mathcal{S}_i|} \sum_{S \in \mathcal{S}_i} [v(S_j \cup \{i\}) - v(S_j)]. \tag{2.3}$$

Al repetir este proceso para cada punto $i \in N$, obtendremos el estimador $\hat{\phi}_{MC} = [\hat{\phi}_{MC}(1), \ldots, \hat{\phi}_{MC}(n)]$.

Teorema 2.3.7. *El estimador de Montecarlo simple $\hat{\phi}_{MC}$ es una (ϵ, δ)-aproximación de $\phi_{Banzhaf}$ en norma l_p con $\mathcal{O}(\frac{n^2}{\epsilon^2} \log(\frac{n}{\delta}))$ evaluaciones de v, y $\mathcal{O}(\frac{n}{\epsilon^2} \log \frac{n}{\delta})$ evaluaciones de v en la norma $l_{\text{ínf}}$.*

Demostración. Consultar apéndice C.1.2 de [18]. □

El método anterior podría mejorarse en eficiencia, dado que cada muestra $S \in \mathcal{S}_i$ generada solo contribuye a la estimación de $\hat{\phi}_{Banzhaf}(i; v)$. Esto introduce un factor de n en la complejidad, ya que es necesario generar un mismo número de muestras para cada dato.

Es en este contexto donde surge el concepto del estimador de máxima reutilización (MSR), propuesto en [18]. La idea es explotar la linealidad de la esperanza, de forma que:

$$\phi_{Banzhaf}(i; v) = \mathcal{E}_{S \sim \text{Unif}(2^{N \setminus i})}[v(S \cup \{i\})] - \mathcal{E}_{S \sim \text{Unif}(2^{N \setminus i})}[v(S)]. \tag{2.4}$$

Tomemos como ejemplo un conjunto de m muestras $\mathcal{S} = \{S_1, \ldots, S_m\}$, generado de manera uniforme. Para cada $i \in N$, podemos clasificar las muestras de $\mathcal{S}$ en dos categorías:

- $\mathcal{S}_{\ni i}$: el conjunto de muestras que contienen el dato i, es decir, $\mathcal{S}_{\ni i} = \{S \in \mathcal{S} : i \in S\}$.
- $\mathcal{S}_{\not\ni i}$: el conjunto de muestras queno contienen el dato i, esto es, $\mathcal{S}_{\not\ni i} = \{S \in \mathcal{S} : i \notin S\}$.

Así, para cada jugador, diferenciamos entre las muestras que incluyen a dicho jugador y las que no.

Utilizando esta clasificación, podemos estimar $\phi_{Banzhaf}(i; v)$ de la siguiente manera:

$$\hat{\phi}_{MSR}(i; v) = \frac{1}{|\mathcal{S}_{\ni i}|} \sum_{S \in \mathcal{S}_{\ni i}} v(S) - \frac{1}{|\mathcal{S}_{\not\ni i}|} \sum_{S \in \mathcal{S}_{\not\ni i}} v(S). \tag{2.5}$$

A este método le denominamos *estimador de máxima reutilización* (MSR).

Teorema 2.3.8. *$\hat{\phi}_{MSR}$ es una (ϵ, δ)-aproximación de $\phi_{Banzhaf}$ en norma l_p con $\mathcal{O}(\frac{n}{\epsilon^2}\log(\frac{n}{\delta}))$ evaluaciones de v, y $\mathcal{O}(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$ evaluaciones de v en la norma $l_{\text{ínf}}$.*

Demostración. Consultar apéndice C.1.2 de [18]. □

Uno podría cuestionarse: ¿por qué se opta por el valor de Banzhaf en lugar de otro semivalor? La razón radica en que el valor de Banzhaf es el único semivalor que posibilita la implementación del algoritmo MSR. Como puede verse en el apéndice C.2 de [18].

Aunque el MSR destaca por ser superior al método de Montecarlo simple, surge una pregunta válida: ¿es verdaderamente eficiente el MSR?

La respuesta la encontramos en el siguiente teorema:

Teorema 2.3.9. *Todo estimador aleatorio del valor de Banzhaf que sea una (ϵ, δ)-aproximación en norma $l_{\text{ínf}}$ con $\delta \in (0, \frac{1}{2})$ requiere al menos $\Omega(\frac{1}{\epsilon})$.*

Como hemos visto anteriormente el algoritmo MSR presenta una complejidad de $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{n}{\delta}))$ en la norma $l_{\text{ínf}}$. Lo que quiere decir que se aleja de la optimalidad en un factor de $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{n}{\delta}))$.

Bibliografía.

[1] Lloyd S Shapley «A Value for n-Person Games», 1952 ed. por Harold W. Kuhn y Albert W. Tucker, págs. 307-317.

[2] Alvin E Roth «*The Shapley value: essays in honor of Lloyd S. Shapley*» Cambridge University Press, 1988.

[3] Shaheen S. Fatima, Michael Wooldridge y Nicholas R. Jennings «A linear approximation method for the Shapley value» *Artificial Intelligence* 172.14, 2008, págs. 1673-1699.

[4] Stefano Moretti, Vito Fragnelli, Fioravante Patrone y Stefano Bonassi «Using coalitional games on biological networks to measure centrality and power of genes» *Bioinformatics (Oxford, England)* 26, 2010, págs. 2721-30.

[5] Pradeep Dubey y Robert J. Weber *Probabilistic Values for Games* Cowles Foundation Discussion Papers 471 Cowles Foundation for Research in Economics, Yale University, 1977.

[6] Pradeep Dubey, Abraham Neyman y Robert James Weber «Value Theory without Efficiency» *Mathematics of Operations Research* 6.1, 1981, págs. 122-128.

[7] J.F. Banzhaf «Weighted voting doesn't work: A mathematical analysis» *Rutgers Law Review* 19.2, 1965, págs. 317-343.

[8] Irwin Mann y Lloyd S. Shapley «Values of Large Games, IV: Evaluating the Electoral College by Montecarlo Techniques», 1960.

[9] Sasan Maleki «Addressing the computational issues of the Shapley value with applications in the smart grid» Tesis doct. University of Southampton, ago. de 2015.

[10] Ian C Covert, Scott M Lundberg y Su-In Lee «Shapley feature utility» *IEEE Transactions on Information Theory*, 2019.

[11] Javier Castro, Daniel Gómez y Juan Tejada «Polynomial calculation of the Shapley value based on sampling» *Computers & Operations Research* 36.5, 2009 Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X), págs. 1726-1730.

[12] Shay Cohen, Eytan Ruppin y Gideon Dror «Feature selection based on the shapley value» *other words* 1.98Eqr, 2005, pág. 155.

[13] Scott Lundberg y Su-In Lee «A Unified Approach to Interpreting Model Predictions», 2017 eprint: arXiv:1705.07874.

[14] Amirata Ghorbani y James Zou «Data Shapley: Equitable Valuation of Data for Machine Learning», 2019 eprint: arXiv:1904.02868.

[15] Ruoxi Jia et al. «Scalability vs. Utility: Do We Have to Sacrifice One for the Other in Data Importance Quantification?», 2019.

[16] Ruoxi Jia et al. «Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms» *Proc. VLDB Endow.* 12.11, 2019, págs. 1610-1623.

[17] Yongchan Kwon y James Zou «Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning», 2021 eprint: arXiv:2110.14049.

[18] Jiachen T. Wang y Ruoxi Jia «Data Banzhaf: A Robust Data Valuation Framework for Machine Learning», 2022 eprint: arXiv:2205.15466.

[19] Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo y Bryan Kian Hsiang Low «Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards», 2021 arXiv: 2112.09327 [cs.LG].

[20] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo y Bryan Kian Hsiang Low «Validation Free and Replication Robust Volume-based Data Valuation» 34, 2021 ed. por M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang y J. Wortman Vaughan, págs. 10837-10848.

[21] Mohammad Mohammadi Amiri, Frederic Berdoz y Ramesh Raskar «Fundamentals of Task-Agnostic Data Valuation», 2022 arXiv: 2208.12354 [cs.LG].

[22] Jinsung Yoon, Sercan Arik y Tomas Pfister «Data Valuation using Reinforcement Learning» Proceedings of Machine Learning Research 119, 2020 ed. por Hal Daumé III y Aarti Singh, págs. 10842-10851.