En luz de las nuevas y populares tendencias en el mundo de la estadística computacional, llamada en ocasiones aprendizaje estadístico u aprendizaje de máquina,[1] este trabajo, busca desarrollar y entender desde sus cimientos, un modelo aplicable a esta categoría. Este modelo, buscará hacer inferencia sobre una base de datos y *aprender* sobre los patrones subyacentes que estos puedan contener. Se busca revisar todos los aspectos de su construcción: desde consideraciones teóricas hasta diferentes paradigmas de aprendizaje, así como su implementación y validación.

Este tipo de modelos, han resultado ser de enorme efectividad en ámbitos que van desde la medicina hasta las finanzas. En ocasiones sin embargo, por su complejidad, los métodos de ML son tratadas como *cajas negras* computacionales; se tienen datos que se alimentan a un modelo complejo y este arroja resultados. Sin dudarlo útiles, el tratamiento de los datos y el modelo en si no se debe dejar de un lado, pues, existen consideraciones técnicas y supuestos que se deben cumplir. Asimismo, la interpretación, validación y análisis de los resultados, deben ser realizados por alguien que conozca, al menos de manera general, lo que está haciendo el algoritmo empleado por la computadora.

En particular, el modelo presentado a continuación, realiza la estimación de variables binarias en un contexto de regresión bayesiana a través de un proyector aditivo con una transformación no lineal de los datos. Esta maraña de términos técnicos, se irá esclareciendo poco a poco conforme se construye el modelo. En el fondo, el modelo busca encontrar patrones de segmentación. Esto lo logra, clasificando

1. *machine learning (ML)*

cada una de las n observaciones como *éxito o fracaso*, *positivo o negativo*, *hombre o mujer* o cualquier otra posible respuesta binaria y_i, dependiendo de información adicional $\mathbf{x}_i$ conocida como covariable donde $i = 1, \ldots, n$. El problema radica en que la información adicional es compleja y puede contener patrones difícil de identificar, lo cual, hace que distinguir entre los posibles resultados de y_i sea difícil. En la Figura 1 se tiene un ejemplo gráfico de este tipo de clasificadores. Con algunos de los modelos tradicionales, por construcción, llevar a cabo esta clasificación sería imposible.

Para llevar a cabo esta construcción, se comienza con una extensa discusión teórica y matemática. No obstante, se hace énfasis en el desarrollo del algoritmo, esto, pues la implementación del modelo es fundamental para su aplicación practica.

> *[...], it is more common in machine learning to view the model as core, and how this is implemented is secondary. From this perspective, understanding how to translate a mathematical model into a piece of computer code is central.*[2]

Es fundamental entender a fondo cada pedazo del modelo, por ello, en el Capítulo **??** se hace una exploración de su forma matemática más rigurosa. Dada su estructura, el modelo se puede estudiar de arriba hacia abajo, es decir, de la parte más general a la parte más profunda. Por lo tanto, primero se estudian los Modelos Lineales Generalizados (GLM), específicamente los modelos probit asociados a la distribución normal. Los GLM dan el salto de una regresión donde la respuesta y_i es real, a regresiones donde la respuesta, puede ser discreta o restringida a

2. **barber2012bayesian**

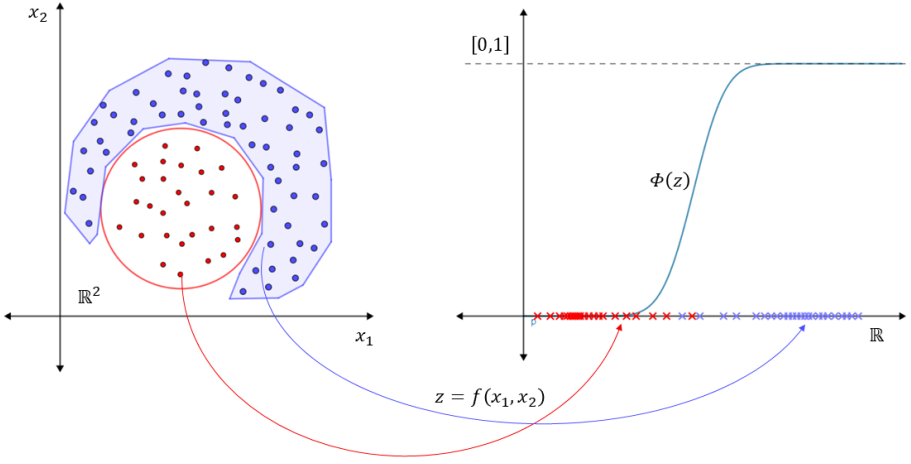

Figura 1: Diagrama explicativo del modelo.

Se tienen observaciones del grupo azul y del grupo rojo con una clara separación no lineal en las covariables x_1 y x_2. El modelo busca *entrenar* una función f que logre separar lo mejor posible este espacio. Posteriormente, esta separación, induce una clasificación (0 y 1 correspondiendo a rojo y azul respectivamente) a través de la función de distribución normal Φ.

cierto dominio (**maccullagh1989generalized**). Los GLM, como su nombre lo indica, siguen siendo lineales, pero este proyector lineal se puede flexibilizar un poco más usando las ideas de los Modelo Aditivo Generalizado (GAM) presentadas en **hastie1986generalized**. Los GAM, buscan transformar a las covariables $\mathbf{x}_i$, previamente a la regresión usando métodos no paramétricos. Este trabajo, toma esas ideas y las combina con las de **mallik1998automatic**, en el que se llevan un paso más allá la transformación para hacerla *tan flexible como sea posible*. Esta transformación, corresponde a una serie de polinomios por partes de continuidad y grado

arbitrarios, sujetos a ciertos nodos, lo cual representa la parte más profunda del modelo. La expansión que se presenta, resulta que conectan muchas disciplinas y ramas de las matemáticas que han sido de mucha utilidad no solo en el campo de la estadística. Al final del Capítulo **??**, se verá que con estos principios se abre un mundo de posibilidades en cuanto a modelos y datos sobre los que se pueden aplicar.

Posteriormente en el Capitulo **??**, se hace una breve introducción a la estadistica bayesiana, en particular al apredizaje bayesiano en el contexto de regresión lineal. Esto se debe a que, usando las ideas de **albert1993bayesian**, el algoritmo asociado al modelo recae en una técnica fundamental de esta diciplina: el *Gibbs sampler*. Con esta poderosa herramienta, se presenta los detalles y lógica detrás de la implementación. Además, se explica a grandes razgos como se hizo el desarrollo y de un paquete computacional de código abierto en el software `R` para el uso del modelo. El desarrollo de un paquete se detalla en el Apéndice **??**, y corresponde a que, no solo se simplifica el proceso de apredizaje del modelo, sino que se fomentando su fácil uso y su validación por terceras personas que se pudieran interesaran en el. Se hace notar, que al paquete se le añadió funcionalidad adicional para la visualización de ciertas partes del modelo bajo algunos supuestos facilitando su interpretación.[3]

Una vez que el modelo fue funcional y fácil de implementar, se probó y se validó contra una serie de bases de datos, tanto simulados como reales para probar su efectividad. En el Capítulo **??**, se puede estudiar el modelo en una forma más pragmática, pues el uso del paquete lo facilita mucho. En particular, las bases de datos simuladas

3. El paquete se puede descargar libremente de: `https://github.com/PaoloLuciano/bpwpm`

ejemplifican muy bien las matemáticas detrás del modelo y muestran la flexibilidad de los polinomios por partes, pues, logran encontrar fronteras de clasificación complejas y evidentemente no lineales. Uno de los ejemplos replica de forma fiel, la Figura 1.

Finalmente, en el Capítulo **??**, se verán las consideraciones finales y limitaciones del modelo. Sin embargo, se abre una discusión a posibles extensiones para mejorarlo. Posteriormente, se da un vistazo a modelos más modernos los cuales han sido capaces de proezas computacionales que se creían imposibles hace algunas décadas. Se verá, sin embargo, que muchos de los modelos más avanzados y usados hoy en día, son generalizaciones de los modelos tradicionales presentados en este trabajo. Si estos modelos, se comienzan a anidar unos dentro de los otros se logra extender el *aprendizaje* más allá de datos binarios y lograr clasificaciones de imágenes, sonidos y datos poco ortodoxos para la estadística.