En luz de las nuevas y populares tendencias en el mundo de la estadística computacional, llamada en ocasiones aprendizaje estadístico o aprendizaje de máquina;[1] este trabajo plantea como objetivo: estudiar, explicar e implementar un modelo de clasificación supervisada con base en la extensión del modelo *probit* al que se le añade un componente no lineal de bases aditivas en covariables. Asimismo, se desarrolla un algoritmo asociado de aprendizaje para la inferencia del modelo y generación de predicciones con base en el paradigma bayesiano.[2]

El modelo hace inferencia sobre un conjunto de datos y *aprende* acerca de los patrones subyacentes que estos mismos puedan contener para posteriormente, predecir el resultado de las variables de respuesta. Este tipo de modelos, han resultado ser de enorme efectividad en ámbitos tan diversos, como lo son la medicina y las finanzas. Bajo esta óptica, se busca que el modelo sea práctico y útil, sin perder de vista el componente teórico que lo sustenta. Por lo tanto, se busca explicar a detalle cada componente del modelo y del algoritmo para que éste no sea tratado como una caja negra computarizada.

Los modelos probit son un tipo de regresiones generalizadas, que buscan explicar la clasificación de variables de respuesta y_i binarias (éxito o fracaso, positivo o negativo, etcétera) con base en un conjunto de covariables $\mathbf{x}_i$ que contienen información para cada una de las observaciones $i = 1, \ldots, n$.[3] Sin embargo, la relación entre y_i con $\mathbf{x}_i$

1. *Machine Learning (ML) en la mayoría de la literatura.*

2. Es común, hacer una distinción entre el aprendizaje estadístico y el aprendizaje de máquina pues, mientras que los modelos son los mismos, difieren en perspectiva. El aprendizaje estadístico presta mayor atención al aspecto inferencial e interpretación, cuando el aprendizaje de máquina coloca mayor énfasis en la implementación computacional y los resultados.

3. Es usual en la literatura, hablar de *clasificadores* cuando las respuestas son categorías (codi-

puede ser compleja y no necesariamente lineal; esto lleva a que la predicción de las respuestas con base en las covariables sea difícil. Para sobrepasar esto, al modelo se le agrega un componente no lineal en covariables que permite discernir entre estos patrones. Como se verá en el trabajo, el modelo induce fronteras no lineales de clasificación en el espacio donde $\mathbf{x}_i$ tome valores. En la figura 1, se tiene un ejemplo gráfico de tipo de clasificación que lleva a cabo el 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 proceso de aprendizaje busca *entrenar*, bajo el paradigma bayesiano, a una función η que logre separar este espacio de la mejor forma posible. Esta separación, induce una clasificación binaria (0 y 1 correspondiendo a rojo y azul respectivamente) a través de la función de distribución normal Φ. Con un modelo cuya frontera fuera lineal en covariables llevar a cabo esta clasificación sería imposible.

Se comienza con una discusión teórica en el capítulo **??** donde se presentan los conceptos que irán constuyendo al modelo. Primeramente se estudian los modelos lineales generalizados (GLM), específicamente los modelos probit. Los GLM como su nombre lo indica, generalizan las regresiones tradicionales donde la respuesta y_i es escalar ($y_i \in \mathbb{R}$) a regresiones donde la respuesta puede ser discreta o restringida a cierto dominio (**maccullagh1989generalized**). No obstante, los GLM siguen siendo lineales en covariables pero se pueden flexibilizar usando diversas técnicas; entre ellas, los modelos aditivos generalizados (GAM) presentados en **hastie1986generalized**. En estos modelos, la flexibilización se logra al transformar las covariables $\mathbf{x}_i$ mediante una función de predicción η, usando métodos no paramétricos con base en

ficadas en variables discretas) y *regresiones* cuando las variables de respuestas son continuas.

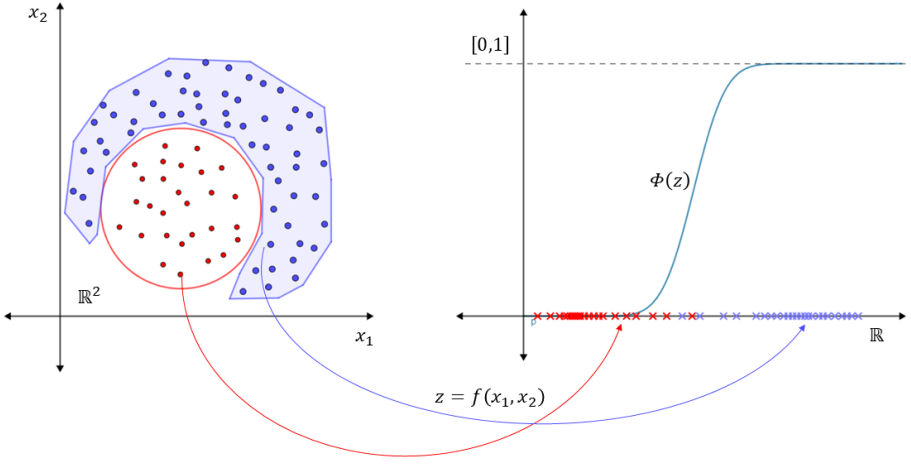

Figura 1: Diagrama explicativo de un modelo de clasificación probit no lineal

suavizadores. Para este trabajo, se toman esos conceptos y se mezclan con los de **mallik1998automatic** los cuales optan por darle una forma funcional concreta a η, correspondiente a una expansión de bases funcionales, particularmente, en polinomios por partes de continuidad y grado arbitrarios. Asimismo, a lo largo del capítulo se verá que los conceptos mostrados abren las posibilidades en cuanto a modelos y datos sobre los que se puede hacer inferencia. Para finalizar el capítulo, se define una versión preliminar del modelo.

Para complementar la formulación funcional previa, el capítulo **??** introduce de forma breve las ideas de la escuela bayesiana de la estadística, en particular el apredizaje bayesiano bajo un contexto de regresión. Este paradigma, responde a

que bajo el trabajo de **albert1993bayesian**, el modelo se puede plantear de tal forma que el algoritmo de aprendizaje se induce de forma natural. Para ello, se debe estudiar el muestreo de Gibbs que recae sobre otros conceptos fundamentales de la estadística bayesiana. Asimismo, el paradigma bayesiano resuena con las ideas del autor en cuanto a lo que implica la probabilidad como una forma de *medir la incertidumbre* y la actualización del conocimiento.

Motivado por los conceptos del aprendizaje bayesiano, el capítulo **??** especifica en su forma final el modelo el cual se titula *bpwpm* por las siglas en inglés de *bayesian piecewise polinomial model*. Asimismo, el capítulo presenta en su forma más completa el algoritmo de aprendizaje usado para la estimación de los parámetros. Esta implementación se realiza a través de un paquete del mismo nombre desarrollado en el lenguaje abierto de programación estadística `R`.[4]

En el capítulo **??** el modelo se prueba y se valida haciendo inferencia sobre seis bases de datos. No obstante, primero se hace una breve discusión sobre cómo evaluar la efectividad y precisión de un modelo como el presentado en este trabajo. Posteriormente, se estiman los parámetros del modelo en cinco bases de datos simuladas, todas con dos covariables ($\mathbf{x}_i \in \mathbb{R}^2$). Estas pruebas preliminares sirven para demostrar las capacidades predictivas del modelo y sobre todo, para hacer más concretas las matemáticas subyacentes, además de poder visualizar las diferentes fronteras flexibles obtenidas por el modelo. Asimismo, en este capítulo se discute la convergencia

4. El desarrollo y explicación del paquete de cómputo se detalla en el apéndice **??**. El paquete se puede descargar libremente de: `https://github.com/PaoloLuciano/bpwpm`. Asimismo, en la página **??** se presenta un ejemplo mínimo funcional para que el lector pueda poner el modelo a prueba.

de las cadenas obtenidas por el muestreador de Gibbs. Para cerrar el capítulo, se replica un escenario real de análisis y modelado usando una base de datos médicos de cáncer.

Para terminar el trabajo, el capítulo **??** abre la discusión sobre las limitantes del modelo y se presentan las consideraciones finales. Sin embargo, también discuten las múltiples posibles extensiones para mejorarlo. Posteriormente, se da un rápido vistazo a modelos relativamente más modernos los cuales han sido capaces de proezas computacionales que se creían imposibles hace algunas décadas. No obstante, se verá que muchos de estos modelos más complejos son generalizaciones de modelos clásicos y extensiones análogas del trabajo presentado.