Índice general

Capítulo 1

Introducción

En luz de las nuevas y populares tendencias en el mundo de la estadística computacional, a veces llamada *Aprendizaje de Maquina o Machine Learning (ML)*, 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: tanto consideraciones teóricas e históricas hasta diferentes paradigmas de aprendizaje; así como su implementación y validación en una computadora. Todo esto, para dar un contexto sobre la, también llamda, *Inteligencia Artificial*, lo cual no es más que estadística computacional llevada al límite.

Este tipo de modelos, han resultado ser de enorme efectividad en ámbitos que van, desde la medicina hasta las finanzas. Sin embargo, en ocaciones, los métodos de ML son tratadas como *cajas negras*; se tienen datos que se alimentan a un modelo complejo y este arroja resultados. Aunque estos métodos son 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. Además, la interpretación, validación y análisis de los resultados, deben ser realizados por alguien que conozca, al menos de manera general, que está haciendo la computadora.

En particular, el modelo que se presenta a continuación, busca la predicción 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 ira esclareciendo poco a poco conforme se construye el modelo. En el fondo, se busca clasificar cada observación i como: *éxito o fracaso*, *positivo o negativo*, *hombre o mujer* o cualquier otra respuesta binaria y_i, a través de información adicional $\mathbf{x}_i$ conocida como covariable. El problema radica en que es que está información adicional, puede contener un patrón complejo que es difícil de identificar a través métodos tradicionales. En la Figura 1.1, se tiene un ejemplo gráfico de este

tipo de clasificadores.

> [...], 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.[1]

Usando esta cita como concepto rector del trabajo, además de una extensa discusión teórica, se hace énfasis en el desarrollo de un paquete en el software estadístico `R` para su implementación. Esto, en respuesta a que, pasar de la teoría a un código funcional, resultó ser más difícil de lo que se esperaba.

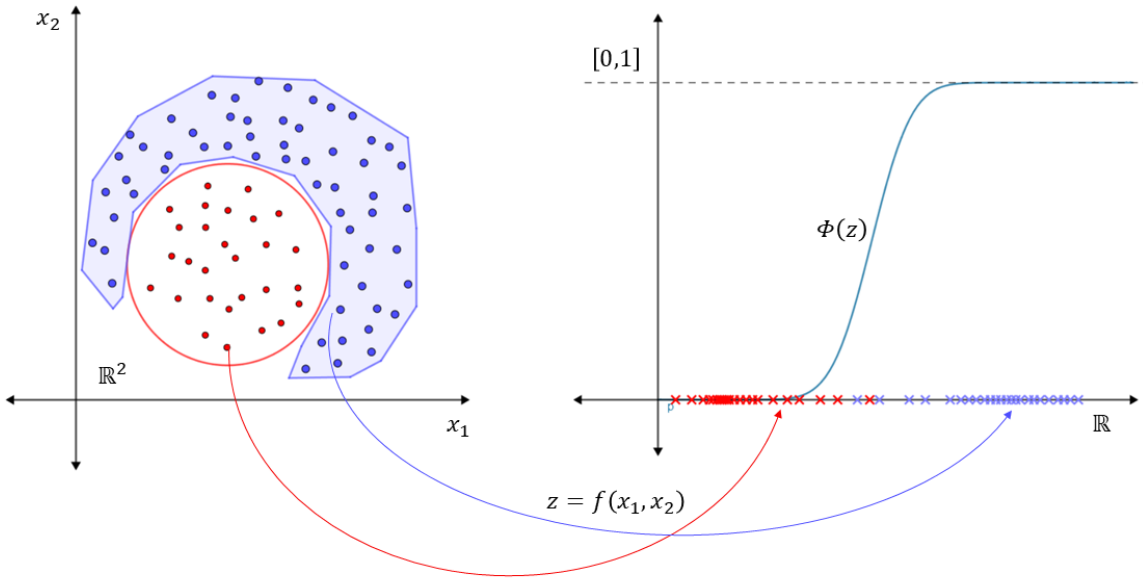

Figura 1.1: **Diagrama explicativo del modelo**. Se tienen observaciones del grupo azul y del grupo rojo con una clara separación no lineal en las covariables x_1 y x_2. El modelo busca *entrenar* una función f que logre separar lo mejor posible este espacio. Posteriormente, esta separación, induce una clasificación (0 y 1 correspondiendo a rojo y azul respectivamente) a través de la función de acumulación normal Φ, de ahí a que el modelo sea *probit*.

Es fundamental entender a fondo cada pedazo del modelo, por ello, en el Capítulo 2 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. Los GLM dan el salto

1. (Barber 2012)

de una regresión donde la respuesta y_i es real, a regresiones donde la respuesta, puede ser discreta o restringida a cierto dominio (MacCullagh y Nelder 1989). 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 (Hastie y Tibshirani 1986). 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 (Denison, Mallick y Smith 1998), en el que se llevan un paso más allá la transformación para hacerla *tan flexible como sea necesaria*; todo bajo un paradigma de aprendizaje bayesiano. 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 2, 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 3, 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 la implementación algorítmica del modelo recae en una técnica fundamental de esta diciplina, el *Gibbs Sampler*, usando las ideas de (Albert y Chib 1993). Con esta poderosa herramienta, se presenta los detalles y lógica del algoritmo. 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 B, y corresponde a que, no solo se simplifica la implementación, sino que se está fomentando el fácil uso del modelo y su validación a terceras personas que se puedan 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.[2]

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 4, 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 ejemplifican muy bien el modelo y muestran la flexibilidad de los polinomios por partes logrando encontrar fronteras de clasificación complejas y evidentemente no lineales. Uno de los ejemplos replica de forma fiel, la Figura 1.1.

Finalmente, en el Capítulo 5, 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

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

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.

Capítulo 2

Modelo en su forma matemática

Como base fundamental de este trabajo, a continuación, se expondrá a detalle el modelo. El objetivo, es construir un clasificador binario flexible con buena fuerza predictiva. La notación se irá explicando conforme aparece pero existe un compendio en el Apéndice C. En general, se trata de respetar la notación que usan en los libros (Hastie, Tibshirani y Friedman 2008) y (James y col. 2013)

Se supone la siguiente estructura en los datos:

- $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ con n el tamaño de la muestra.
- $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ variables de respuesta binarias o *output*.
- $\mathbf{x}_i \in \mathcal{X}^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ covariables, regresores o *input*.
- $d \in \mathbb{N}$ dimensionalidad de mis covariables.

El modelo en si, se presenta a continuación de forma general para cualquier pareja de datos $(y, \mathbf{x})$:

$$y \,|\, z \sim \mathrm{Be}(y \,|\Phi(z)) \tag{2.1}$$

$$z \,|\, x \sim \mathrm{N}(z \,|f(\mathbf{x}), 1) \tag{2.2}$$

$$f(\mathbf{x}) \approx \sum_{j=0}^{d} \beta_j f_j(x_j) \tag{2.3}$$

$$f_j(x_j) \approx \sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_j, \mathcal{P}_j) \quad \forall j = 0, 1, \ldots, d \tag{2.4}$$

En las expresiones (2.1) y (2.3), dejando de un lado ecuación (2.2), se tiene una versión ligeramente modificada de un GLM (Sec. 2.1). Esto, pues la variable de respuesta y es binaria modelada con una

distribución Bernoulli. Además, (2.3) es una función de proyección lineal como las que se usan el los modelos tradicionales. En el contexto de un modelo probit, esta función f, busca separar el espacio d-dimensional de covariables $\mathcal{X}^d$ en regiones identificables en una sola dimensión $\mathbb{R}$. Esta función de proyección, asume que la dependencia entre mis covariables se puede modelar como la suma ponderada de los componentes f_j (Sec. 2.2). Para poder hacer la liga entre ambas ecuaciones, se requiere de la incorporación de una variable latente z, vista en la ecuación (2.2), esta variable es meramente estructural y será modelada a través de una distribución normal, lo cual lleva a tener un modelo probit. Finalmente (2.4) hace una transformación no lineal de cada dimensión j y trata de encontrar las tendencias individuales de cada una de las covariables. Esto se logra, haciendo un suavizamiento por medio de polinomios por partes que dependen de 3 objetos: una partición del intervalo $\mathcal{P}_j$, un vector de pesos w_j y parámetros que captura la N^* especificando la forma y grado de los polinomios. La forma funcional de Ψ es compleja y relativamente arbitraria dependiendo de la selección de la base, por lo tanto, no se especifican aún y se deja para la Sección 2.3. Se hace notar que el componente bayesiano se explora hasta el Capítulo 3 pues va estrechamente ligado con su implementación. En la Figura 2.1 se hace una representación visual del modelo para su mejor comprensión.

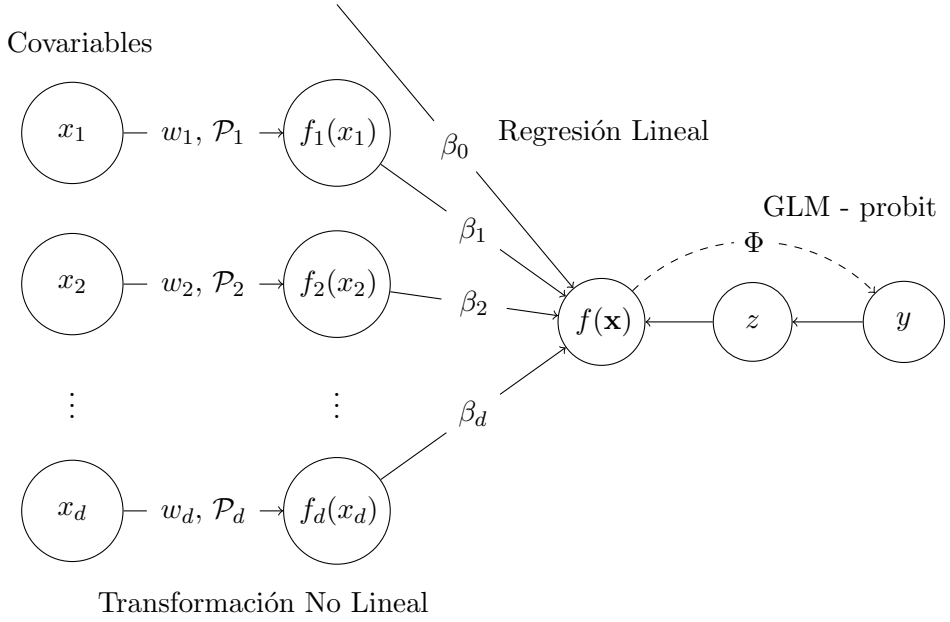

Figura 2.1: **Diagrama del modelo.** Se hace una transformación no lineal de las covariables x_j a través de los parámetros w_j y P_j. Con los datos transformados f_j, se lleva a cabo un modelo probit con función liga Ψ para lograr la clasificación binaria en y.

Antes de continuar, vale la pena recordar que:

All models are wrong but some are useful[1]

1. (Box 1979)

Escoger un modelo que explique perfectamente los datos o que logre predecir todo sería una tarea inútil. Sin embargo, no significa que no se pueda intentar discernir un patrón y es justamente lo que se busca con la construcción de este modelo. Además de entender a profundidad un modelo que sirve como base para modelos que se están usando en el mundo de la inteligencia artificial. En particular, este modelo tiene la ventaja que es flexible y, al menos en teoría, debería de servir para representar una gran cantidad de datos.

2.1. Modelos Lineales Generalizados (GLM)

Los modelos lineales generalizados, (Sundberg 2016) y (MacCullagh y Nelder 1989), surgen como una generalización del modelo lineal ordinario $y = \beta^t x + \epsilon$ donde $y \in \mathbb{R}$. En esta generalización, se busca darle otros rangos a y pues tenemos casos donde está restringida a un subconjunto de los reales como lo es el caso binario. Sin embargo, este cambio vuelve el modelo más complejo y lleva a técnicas diferentes en la estimación de los parámetros β. Además, se pierde algo de la interpretabiliad del modelo[2]. Sin embargo, han resultado ser realmente útiles.

Los GLM se especifican (de manera muy general) de la siguiente manera:

$$
\begin{aligned}
y &\sim F(\theta(x)) \\
z &= \beta^t x \\
\theta &= g^{-1}(z)
\end{aligned}
$$

con los siguientes tres elementos:

1. F: **Tipo de distribución** de la familia exponencial que describa el dominio de las respuestas y. Por ejemplo: Bernoulli si y es binaria, Poisson si $y \in \mathbb{Z}^+$ o una distribución Gamma si $y \in \mathbb{R}^+$

2. z: **Proyector lineal** que explique (linealmente) la variabilidad sistemática de tus datos. En el modelo tradicional $\dim(\beta) = d < n$.

3. g: **Función liga** que una la media (o los parámetros canónicos) θ de mi distribución con el proyector lineal. Es decir: $\theta(x) = \mathbb{E}[y|x] = g^{-1}(\beta^t x)$.

2. Dependiendo de la especificación, su interpretación puede ser complicada. Por ejemplo, cuando se tiene un modelo logit tradicional, se logra expresar el logaritmo de la proporción de probabilidades (*Log-Odds-Ratio*) como una combinación lineal de las covariables. $\ln(\pi_i/\pi_0) = \beta^t x$

Como ejemplos clásicos se tiene la función $\text{logit}(p) = \ln(p/(1-p))$ o la $\text{probit(p)} = \Phi^{-1}(p)$, donde $p = \mathbb{E}[y|x]$ y $\Phi(\cdot)$ es la función de acumulación de una distribución normal estándar. En la Figura 2.2 podemos ver una representación gráfica para su mejor comprensión.

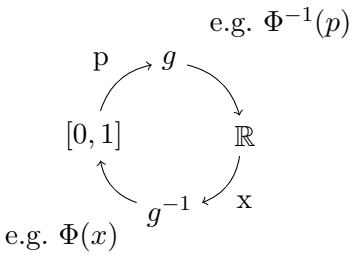

Figura 2.2: **Esquema de función liga** g

Para este trabajo, se busca construir un clasificador binario por lo que $y \in \{0, 1\}$, por lo cual, es natural modelar y como una distribución Bernoulli. Notese que si $Y \sim \text{Be}(y|p)$ se tienen las siguientes propiedades:

$$\mathbb{E}[Y] = p = P(y = 1)$$
$$\mathbb{V}[Y] = p(1-p)$$

Por lo que solo se tiene un parámetro p y la varianza queda determinada automáticamente. Además, esta especificación deja como opción para las función liga, a las inversas de las funciones *sigmoidales* $s(x)$. Las funciones sigmoidales, son funciones $s : \mathbb{R} \to (0, 1)$, estrictamente monótonas y por ende, biyectivas. Algunos ejemplos son las ya mencionadas logit, probit y la curva de Gompertz[3]. Estas funciones cumplen un papel de activación, es decir, una vez que se rebase cierto umbral, crecen rápidamente y toman valores más cercanos a uno, *activando* asi la probabilidad de que y sea un éxito. [4]. Esto las hace perfectas herramientas para ligar el proyector lineal $z \in \mathbb{R}$ con una probabilidad $p \in [0, 1]$.

2.1.1. Uso de la Variable Latente

Ahora, para entender el papel que juega z, se necesita entender que es posible estructurar estos modelos como *modelos de variable latente* (Albert y Chib 1993). Bajo esta formulación, se asume que la relación

3. Para no caer en redundancia de notación para este trabajo se tiene a partir de ahora: $s(x) = g^{-1}(x) = \Phi(x)$

4. En un contexto de redes neuronales, lo que se activa es la neurona y recientemente, se usa la función *ReLu* $(x) := \text{máx}\{0, x\}$

entre y y x no es directa, sin embargo, existe una variable no observada z estructural que nos ayuda a discernir un vínculo entre ellas. En la Figura 2.3 tenemos esa representación del modelo.

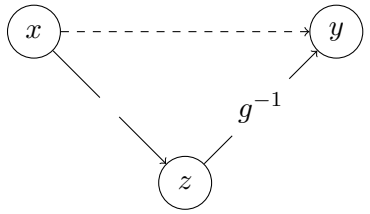

Figura 2.3: **Modelo de variable latente**

Tradicionalmente, la normalidad en z es derivada de asumir normalidad en los errores; es decir, dada la regresión lineal $z = \beta^t x + e$ se asume (y se debe verificar) que $e \sim N(0, \sigma^2)$. Lo cual lleva a $z \sim N(\beta^t x, \sigma^2)$. Además este supuesto facilita la estructura de los modelos y el algoritmo de ajuste. Bajo un paradigma frequentista, la estimación de los parámetros β se reduce a encontrar los estimadores de mínimos cuadrados. Sin embargo, bajo el paradigma bayesiano, dentro de un modelo probit como el de este trabajo, se adopta la normalidad en z pues en (Albert y Chib 1993), se sugiere un algoritmo *Gibbs Sampler* con distribuciones truncadas de la normal para encontrar β.

La función liga probit se escoge como consecuencia de la normalidad en z, o viceversa, dependiendo de como se quiera ver. Esta función $\Phi^{-1}(p)$ es la inversa de la función de acumulación normal estándar:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

la cual no tiene forma cerrada. Sin embargo, tiene justamente las propiedades que se necesitan pues es claramente sigmoide. Esta función, cumple el propósito de modelar y cuantificar la incertidumbre pues está transformando la variable real z en una probabilidad p con su característica forma de "s". Se hace notar que se podría haber usado una función más flexible o que incluso se podría dejar la función como una parte del modelo a estimar. Sin embargo, al adoptar el algoritmo antes citado, se requiere esta especificación.[5]. La parte flexible de este modelo se encuentra en el proyector lineal.

Habiendo definido la función liga, distinguir entre si $y = 1$ ó 0, éxito o fracaso respectivamente, se reduce a distinguir en que área del espacio de covariables $\mathcal{X}^d$ nos encontramos. Esto se debe a que $y = 1$ cuando $\Phi(z) > 1/2$ que sucede *si y solo si* $z > 0$ lo cual, depende en gran media de su media; en este caso la función de proyección $f(\mathbf{x})$. Si esta función es muy positiva en alguna región, implicará que el modelo

5. En modelos multinomiales bayesianos, tomar esta decisión estructural lleva a que inclusive, se puede asumir una estructura de interdependencia en los errores aleatorios. **e** $N_k(0, \Sigma)$ con Σ una matriz de correlaciones

tiene mucha evidencia para confiar que, al menos en esa área, la respuesta y es un éxito. El razonamiento, funciona de forma análoga para los casos donde $y = 0$, claramente, para esas regiones, buscamos que $f(\mathbf{x})$ sea negativa. Por lo tanto, es fundamental para el modelo que se realice una correcta estimación de los parámetros de la función de proyección. Nótese además, que z le agrega cierta *estocasticidad* al modelo. Supongamos que existe una pareja $(y_i, \mathbf{x}_i)$ tal que $f(\mathbf{x}_i) = 0$; alrededor de una vecindad de este punto, no se tendría evidencia para clasificar a y_i como un éxito o como un fracaso; sería mejor un volado.

Otro factor importante a considerar, es que el modelo asume que la varianza de z es constante, específicamente $\sigma^2 = 1$. Dado que la escala de z es completamente arbitraria pues es una variable auxiliar, se puede *restringir* z al rango que se desee. El método de simulación para z usando una distribución normal truncada se simplifica ligeramente usando esta varianza unitaria. Se verá en los resultados, sin embargo, que dada la naturaleza global de los polinomios que se usan, la escala de z, o al menos la estimación de su media $\hat{f}(\mathbf{x})$, puede variar mucho dependiendo de los datos, mas esto no representa un problema. Pues, en la practica, al usar el algoritmo de Albert y Chibb z sirve mucho más, para hacer la ligadura de y hacia f y no viceversa. En z se codifica, mediante una normal truncada, los casos de éxito y de fracasos de y; posteriormente, se estima el vector β para la función f. Por ello, en la Figura (2.1), se representan las flechas de y a f por medio de z y solidas, en contraposición con la flecha punteada que va, directamente de f a y y pasa por la función Φ. Los detalles y su justificación probabilista, se tocan en detalle en el Capítulo **??**.

Es importante mencionar, que el *corte* que se hace en $z = 0$ para la clasificación, es resultado de hacer una clasificación binaria. En modelos multinomiales también se debe tomar en cuenta los intervalos en $\mathbb{R}$ para los que la observación se clasificaría en alguna de las posibles clases y por ende, estimar los umbrales o usar una función diferente a las sigmoides. Este hecho, lleva a la realización de que z y su media f son *ajenas más no independientes*. Esto quiere decir que la parametrización de z como una normal $N(\mu, 1)$ es equivalente a la parametrización $N(0, 1)$. Este hecho se hará más claro cuando hablemos del papel de la β_0 en la función de proyección.

Finalmente, si se quisiera ver la relación de x en y directamente, se puede lograr usando el teorema de la probabilidad total. Se puede calcular (al menos de forma teórica), la distribución marginal de y dado $\mathbf{x}$ sumando sobre z:

$$
\begin{aligned}
P(y|x) &= \int_{-\infty}^{\infty} P(y|z)P(z|x)\,dz \\
&= \int_{-\infty}^{\infty} p(y;\Phi(z))p(z;f(\mathbf{x}),\sigma^2)\,dz \\
&= \int_{-\infty}^{\infty} \Phi(z)^y(1-\Phi(z))^{1-y}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-f(\mathbf{x}))^2}\,dz
\end{aligned}
$$

Sin embargo, está claro que esta derivación no lleva a ningún resultado analítico cerrado pues la relación es bastante más compleja como para resultar en una distribución tradicional; si lo hiciera, el propósito de la z se perdería.

Recapitulando, mediante la función liga Φ se une la media p, la probabilidad de éxito o fracaso, de la respuesta y con los datos $\mathbf{x}$. Esto se logra, a través de una variable auxiliar z cuya media $f(\mathbf{x})$ es una función de proyección lineal.

$$
P(y=1) = p(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g^{-1}(f(\mathbf{x})) = \Phi(f(\mathbf{x})) \tag{2.5}
$$

2.2. Función de proyección f

Tradicionalmente, se asumía que z es una combinación lineal de los parámetros β y las covariables x, pero, como se explica en la pagina 6 de (James y col. 2013), conforme avanzaron los métodos y el poder computacional disponible se fueron desarrollando técnicas cada vez más poderosas que permitían romper la linealidad. En 1986, Hastie y Tibshirani introducen los Modelos Aditivos Generalizados (GAM), una clase de modelos donde se lleva a romper la linealidad en las covariables, esto permite flexibilizar aún el modelo. Como se vio anteriormente, los GLM siguen la forma especificada en la Sección: 2.1, sin embargo, al tratar el modelo con la variable latente $z|x \sim N(f(\mathbf{x}),\sigma^2)$ con función liga probit, se tiene la ecuación (2.5).

Esta nueva *capa* no hace otra cosa más que ir colapsando dimensiones, ie: $\mathcal{X}^d \to \mathbb{R} \to [0,1]$. Es por esto que se llama función de proyección, pues *proyecta* el espacio $\mathcal{X}^d$ en $\mathbb{R}$. Sin embargo, la forma en la que lo hace debe de ser muy sutil pues el modelo recae en que este colapso detecte los patrones correctos en las covariables que llevan a la correcta identificación de y. Por lo tanto, f como función de proyección es el corazón del modelo, por lo que su correcto entrenamiento es fundamental. La idea, recapitulando, es que f separe el espacio de covariables para que sea positiva en las regiones donde tengamos éxitos y negativa

donde tengamos fracasos; para ello, es fundamental entender los GAM.

2.2.1. Modelos Aditivos Generalizados (GAM)

Un GAM como lo introducen Hastie y Tibshirani en (Hastie y Tibshirani 1986) remplaza la forma lineal $\sum_1^d \beta_i x_i = \beta^t x$ con una suma de funcione suaves $\sum_i^d f_i(x_i)$. Estas funciones no tienen una forma cerrada y son no especificadas, es decir, no hay tienen una forma funcional concreta y representable algebraicamente. Donde recae la fuerza del modelo, es que se estiman usando técnicas de suavizamiento no paramétricas[6] como lo sería un Suavizamiento Loess. En estos modelos, se asume que por más grande que sea $\mathcal{X}^d$, la relación que existe entre cada una de las dimensiones se puede explicar de manera aditiva. Esta especificación fue revolucionaria pues no solo regresa interpretabilidad al modelo, sino que simplifica la estimación usando técnicas prácticamente automáticas con el algoritmo de *backfitting*. Además, Los GAM, demuestran que son efectivos en descubrir efectos no lineales en las covariables.

Para este trabajo, se escoge la ecuación (2.3):

$$f(\mathbf{x}) \approx \sum_{i=0}^{d} \beta_i f_i(x_i)$$

Esta función se escoge respondiendo a que este modelo, se planea usar para aplicaciones en econometría, donde las dimensiones i's son independientes entre si y corresponden a diferentes *características* o variables con las que se planea modelar la variable de respuesta. Por ejemplo, se puede pensar que cada $x_i \quad \forall i = 1, \ldots, d$ como d series de tiempo con las que se planea predecir cuando vender o comprar cierta acción.

La f es una versión versión modificada de un GAM tradicional con tres cambios fundamentales. La primera modificación es que estamos ponderando cada f_i por un parámetro β_i, esto es, para suavizar aún más cada dimensión y captar el patrón general y no tanto los componentes individuales de cada x_i. Se puede pensar en cada f_i como una transformación no-lineal de x_i (como lo sería una transformación logarítmica o una transfomación Box-Cox) por lo que se puede dar una interpretación al parámetro β_i como el efecto que tiene la dimensión i en particular para el modelo. Se hace notar que se deja espacio para un término independiente β_0. Por convención $f_0(\cdot) = 1$ por lo que se puede re-expresar la ecuación anterior (2.3) como:

6. Las técnicas no paraméticas están fuera del alcance de este trabajo. Sin embargo, vale la pena una mención especial por su funcionalidad, practicidad y forma intuitiva, además del sinfín de aplicaciones que tienen. Una guía comprensiva de estas se encuentra en el libro *All of Non-parametric statistics*(Wasserman 2007).

$$f(\mathbf{x}) \approx \beta_0 + \sum_{i=1}^{d} \beta_i f_i(x_i) = \boldsymbol{\beta}^t F$$

Donde usando notación vectorial $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ y $F \in \mathbb{R}^{d+1}$. La inclusión de este parámetro es fundamental para la correcta especificación del modelo pues ayuda a dar un *sesgo o nivel* base contra el cual comparar la suma y escalar la f para que sea compatible con el umbral de corte en 0 haciendo ecuivalente la parametrización de z con una normal estándar.

La segunda modificación, es una más sutil, aunque es muy práctico manejar las f_i's como indeterminadas y estimarlas con procedimientos de suavizamiento no parametricos, también se puede optar por la vía en la que se especifica su forma funcional, no por ello quitandoles flexibilidad. Los detalles de esto lo dejaremos para la Sección 2.3 donde se trata de adapart el procedimiento de (Denison, Mallick y Smith 1998). Esta modificación, obedece a que para ciertas aplicaciones, sirve hacer el modelo paramétrico en donde cada f_i se modela en su expansión de bases y se puede hacer el ajuste con un algoritmo de minimos cuadrados. (Vease Capitulo 9.1 y Ejemplo 5.2.2 de (Hastie, Tibshirani y Friedman 2008)).

La tercera modificación, es que estamos trabajando con una aproximación en vez de una igualdad para f. Se hace notar que esta es uno de los supuestos más fuertes del modelo, esto responde a que el *error aleatorio*, sistemático de los datos, está siendo capturado por una aproximación a la f real, dentro de cada una de las f_i's. En la siguiente sección se explora el porqué de esta aproximación usando técnicas de análisis más avanzadas.

Imaginar o peor aún, visualizar f es complicado, pero el objetivo es que (si se tienen datos continuos) f agrega los efectos de cada una de las componentes y separa el espacio en regiones de éxitos y fracasos. Es decir, si tenemos puntos en $\mathbb{R}^2$, f se podrá visualizar en $\mathbb{R}^3$ como una serie de picos y valles donde será positiva en caso de ser éxito y negativa en caso contrario.

Finalmente, notemos que este modelo se podría confundir con un Modelos en Bases de Funciones Lineales como lo presentado en Capitulo 3 en (Bishop 2006) donde se tiene:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{d} \beta_i f_i(\mathbf{x}) \tag{2.6}$$

La diferencia radica en que cada f_i es función de todas mis covariables en lugar de solo la dimensión i. A estas funciones se les conocen como *funciones base* y son lineales en β, pero no-lineales para $\mathbf{x}$. Se hará

una exposición más a detalle en la Sección 2.3 pero, las posibilidades son ilimitadas para estas funciones, algunos ejemplos son:

- **Bases Gaussianas:**
$$f_i(\mathbf{x}) = \exp\left\{-\frac{(\mathbf{x}-\mu_i)^2}{2s^2}\right\}$$
- **Funciones sigmoidales:**
$$f_i(\mathbf{x}) = \sigma\left(\frac{\mathbf{x}-\mu_i}{s}\right)$$

Sin embargo, estos son un grupo de modelos completamente diferente cuyas aplicaciones usualmente son en estimación de curvas y no tanto inferencia como lo busca este trabajo.

2.3. Funciones f_i

Finalmente se trata la parte más profunda del modelo, las funciones f_i que, como se mencionó anteriormente, son transformaciones no lineales de cada componente x_i que buscan suavizar la nube de datos, para posteriormente sumarlas entre si y dar una medida f que resuma la información. Como se menciona en la introducción de (Härdle y col. 2004), el suavizamiento de los datos es central en la estadística inferencial. La idea es extraer la señal entre el ruido y para ello, se intenta estimar y modelar la estructura subyacente. Este suavizamiento, se llevará a cabo usando una *expansión en bases funcionales* como lo visto en la ecuación (2.6). Toda la siguiente sección se concentra en darle formas funcionales a las Ψ's. Se usa como referencia el captiulo 5 de (Hastie, Tibshirani y Friedman 2008).

Una expansión en bases de una función $h : \mathbb{R}^d \to \mathbb{R}$ es:

$$h(\mathbf{x}) = \sum_{j=1}^{J} w_j \Psi_j(\mathbf{x}) \tag{2.7}$$

Donde, $\Psi_j(\mathbf{x})$ es la j-ésima transformación no lineal de $\mathbf{x}$ y una vez especificadas y estimadas, el procedimiento (hacia arriba en el modelo) se hace de forma tradicional pues recobra su estructura lineal. Algunos ejemplos son:

- $\Psi_j(\mathbf{x}) = x_j$ donde $j = 1, \ldots, d$ y se tiene el modelo lineal más sencillo.
- $\Psi_j(\mathbf{x}) = \ln x_j$ ó $x_j^{1/2}$ donde se tienen transformaciones no lineales en cada una de las covariables.
- $\Psi_j(\mathbf{x}) = ||\mathbf{x}||$ una transformación lineal de todas las covariables.[7]

7. Como se vio en la ecuación (2.6)

Dependiendo del tipo de datos y de aproximación que se busque, puede ser conveniente usar forma sobre la otra; existen muchas más posibles expansiones de bases. Sin embargo, sobre todo cuando se tiene poca o ninguna experiencia con los datos, se busca una representación más flexible (por no decirla ingenua) de estos. El método más común, es tomar una familia de funciones como los son los polinomios por partes o familias de funciones flexibles que logren representar una gran variedad de patrones. En estos métodos, se cuenta con una gran cantidad de funciones base por lo que se requiere controlar la complejidad; las formas más comunes de lograrlo son:

- **Métodos de Restricción**: como los son los métodos aditivos usados en este trabajo.
- **Métodos de Selección**: como lo son los modelos CART y MARS.
- **Métodos de Regularización**: donde se busca controlar los coeficientes, como los son los modelos *Ridge* y *LASSO*.

Simplificando un poco la exposición, por lo pronto, se puede pensar únicamente en funciones reales, por lo que se deja de usar el subíndice i para indicar el componente del vector $\mathbf{x}$.

Para este trabajo, se aplica el procedimiento de (Denison, Mallick y Smith 1998). Los autores presentan un método revolucionario, que permite estimar con un alto grado de precisión relaciones funcionales entre la variable de respuesta y y el regresor $x \in \mathbb{R}$. Se puede pensar que se busca ajustar una curva tradicional. Esto es, para un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$ se plantea el modelo:

$$y_i = h(x_i) + e_i \quad i = 1, \ldots, n \tag{2.8}$$

con las e_i errores aleatorios de media cero. Este método, combina los procedimientos paramétricos y no paramétricos desarrollados antes para hacer más robusto el algoritmo de (Hastie y Tibshirani 1986). La idea, es ajustar un *polinomio por partes* muy flexible. Estos polinomios, se componen de partes de menor orden entre *nodos* adyacentes. La genialidad del su trabajo es que estos nodos, tradicionalmente fijos, se vuelven parámetros a estimar, usando un paradigma bayesiano. Y no solo eso, sino que permiten *aumentar o disminuir el número de nodos* desarrollando un algoritmo Gibbs sampler trans-dimensional. Esta generalización, logra estimaciones tan robustas, que logran aproximar funciones continuas *casi en todas partes*, como lo son la función Doppler, funciones por bloques y funciones con picos pronunciados.

2.3.1. Polinomios por partes y splines

Antes de llegar a estos polinomios tan flexibles, se busca entender que son los polinomios por partes simplificando la exposición de (Wahba 1990). Sea $x \in [a, b] \subseteq \mathbb{R}$, se busca separar $[a, b]$ en J intervalos. Por lo tanto, se construye una partición correspondiente $\mathcal{P} = \{\tau_1, \tau_2, \ldots, \tau_{J-1}\}$ tal que $a \leq \tau_1 < \ldots < \tau_{J-1} \leq b$. Estas τ's son llamadas *nodos*. Se hace notar, que se puede incluir o no la frontera y que a cada intervalo le corresponde una función Ψ_j. Con estos nodos, se puede representar a la función global h en su expansión de bases como en la ecuación (2.7), donde cada Ψ_j es una función que depende de la partición y de x. Por ejemplo, se puede pensar en un caso sencillo donde se tiene que $J = 3$ y se quiere ajustar funciones constantes en cada intervalo. Entonces, las funciones base correspondientes serían:

$$\begin{aligned}
\Psi_1(x, \mathcal{P}) &= I(x < \tau_1) \\
\Psi_2(x, \mathcal{P}) &= I(\tau_1 \leq x < \tau_2) \\
\Psi_3(x, \mathcal{P}) &= I(\tau_2 \leq x)
\end{aligned}$$

Con $I(\cdot)$ la función indicadora que vale 1 si x se encuentra en la región y 0 en otro caso. Por lo tanto,

$$\begin{aligned}
h(x) &= \sum_{j=1}^{J} w_j \Psi_j(x) \\
&= w_1 I(x < \tau_1) + w_2 I(\tau_1 \leq x < \tau_2) + w_3 I(\tau_2 \leq x)
\end{aligned}$$

Lo cual es una función *escalonada*, en el sentido de que para cada región de x tenemos un nivel w_j.[8] Esta aproximación a mis datos podría servir para datos que estén agrupados por niveles, sin embargo, rara vez será ese el caso.

Entre cada pareja de nodos, se puede buscar ajustar un polinomio de grado arbitrario. Adicionalmente, se pueden construir polinomios con restricciones como continuidad en las derivadas, lo cual logra una estimación más robusta. Esta es la magia de los polinomios por partes, que se les puede pedir cuanta *suavidad* queramos, entendido como la continuidad de la K-ésima derivada. Tradicionalmente, se construyen polinomios cúbicos con segunda derivada continua en los nodos. Esto, pues resulta en funciones suaves al ojo humano que logran aproximar una gran cantidad de funciones.

8. Sin entrar en el detalle, usando una función de perdida cuadrática, es fácil demostrar que cada $\hat{w}_j = \bar{x}_j$ es decir, para cada región, el mejor estimador constatne, es el promedio de los puntos de esa región.

Orígenes y justificación de su uso

La palabra *spline*, [9] se usa para designar a este grupo de polinomios por parte. Sin embargom dependiendo de como se definan pueden denotar funciones muy diversas; hasta ahora no hay consenso en la literatura. Para este trabajo se denota a un *spline de grado M* como un polinomio por partes de grado $M-1$ con continuidad hasta la $M-2$ derivada. (Wasserman 2007). Se hace notar, que se tienen definiciónes de splines muy diferentes a las presentadas aquí, todo está en la definición de la partición que además, puede ser tan flexible como se requiera, por ejemplo los B-Splines.

Los splines, surgen en (Schoenberg 1964) donde se plantea el problema: encontrar h en el espacio de Sovolev W_{M-1} de funciones con $M-2$ derivadas continuas y M-ésima derivada integrable al cuadrado que minimice:

$$\int_a^b (h^{(m)}(x))^2\,dx$$

sujeta a que interpole los puntos, ie: $h(x_i) = h_i \quad i = 1, 2, \ldots, n$. Sin embargo, se hace notar que sea como sea la especificación, se tiene un problema con la la naturaleza global de los polinomios, es decir, se necesita controlar lo que pasa más allá de los nodos de la frontera. Por lo que usualmente se escogen condiciones adicionales o linealidad pasando los nodos. En un contexto estadístico, el problema (2.8) se puede plantear como encontrar la función h que minimice:

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda \int_a^b (h^{(m)}(x))^2\,dx \tag{2.9}$$

para alguna $\lambda > 0$, donde la solución se demuestra que son *splines naturales* que se estudian más adelante, específicamente *splines cúbicos naturales* si $m = 2 \quad (M = 4)$. Cabe mencionar, que esta formulación del problema engloba muchas de técnicas estadísticas interesantes, además de conceptos de optimización. El lector reconocerá que el primer término claramente es la *suma de residuales cuadrados (RSS)* y el segundo término del sumando es un procedimiento conocido como *regularización*. No es el enfoque entrar a detalle en cada uno de estos pues merecen una tesis por si mismas, sin embargo, se definen en el Apéndice D. Por lo pronto, lo esencial, es que al tratar de minimizar el RSS se puede caer en problemas de sobreajuste en donde los parámetros no estén capturando efectos y patrones subyacentes, sino solo estén tratando de seguir los datos. Para compensar, se penaliza el modelo con segundo termino que controla la complejidad del modelo y la suavidad deseada mediante la λ. Esto se logra, incorporando un termino de penalización

9. A diferencia de el texto tradicional de (Boor 1978)

el cual crece a medida que h se vuelve más complicada.[10]

Una vez más, Hastie y Tibshirani, con su modelo aditivo y función de perdida cuadrática con penalización en la segunda derivada:

$$y = \sum_{j=0}^{d} h_j(x_j) + \epsilon$$

$$\text{RSS}(h_0, h_1, \ldots, h_d) = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{d} h_j(x_{ij})]^2 \; + \; \sum_{j=1}^{d} \lambda_j \int h_j^{''}(t_j)\, dt_j$$

más la convención de que h_0 es una constant y las λ_j los parámetros de suavizamiento, muestran que $h_j \quad j = 1, \ldots, d$ son splines cúbicos. Sin embargo, el modelo no es identificable pues h_0 puede ser arbitraria. Por lo que se necesita una restricción adicional para que el mínimo sea único, esta es:

$$\sum_{i=1}^{n} h_j(x_{ij}) = 0 \quad \forall j \tag{2.10}$$

es decir, las funciones promedian en cero sobre los datos. Lo que nos lleva a que: $h_0 = \bar{y}$. Por lo que si viéramos cada dimensión j, tendríamos que las h_j estarían centradas alrededor de la media $\bar{y}$. Además, la implementación de esta formulación, es conocida como *Algoritmo Backfitting* que será revisado en el Capítulo **??**.

Formalización matemática de splines

Retomando la discusión de la página 18, se busca definir un polinomio de grado $M-1$ por partes en J intervalos. Tomando una expansión de bases para cada intervalo, como en el ejemplo anterior, el número de funciones base aumenta en J por cada grado que se agregue, dando un total de $J*M$ bases funcionales, y en consecuencia, el mismo número de parámetros por estimar. Esto ocurre porque necesitamos definir $\mathcal{B}_j = \left\{1, x, x^2, \ldots, x^{M-1}\right\}$ para cada j. Sin embargo, esto llevaría a polinomios que se comportan de forma independiente en cada intervalo y no se conectan. La primera condición que se le impone es continuidad en los nodos, lo cual devuelve $J-1$ parámetros que corresponden a los $J-1$ nodos donde se da la continuidad. Cada grado de continuidad en las derivadas que se le pida al polinomio, se restringe el modelo y por ende, el número de funciones bases necesarias a un total de:

10. Si el lector tiene una intuición de análisis, notará que integrar la función al cuadrado, corresponde con el producto interno de las funciones pertenecientes al espacio de Hilbert $\mathcal{L}_2([a, b])$. Más detalles de esto en la Sección **??**

$$N^*(M, J, K) = M * J - K * (J - 1) \tag{2.11}$$

donde K es el número de restricciones para cada nodo ($K \leq M - 1$), definiendo hasta que número de derivada es continua. Independientemente de la base que escojamos, $N*$ será la *dimensión mínima de la base* es decir, el número de funciones necesarias para representar un polinomio en función de M que define su grado, el número de intervalos J, por ende el número de nodos y K. Por lo pronto, se centra la discusión cuando $K = M - 1$ regresando a la definición de spline: polinomios de grado $M - 1$ con continuidad hasta la $M - 2$ derivada. Por ende, la dimensión queda: $N^* = M + J - 1$

Ahora, para definir la expansión de bases, se define la función auxiliar *parte positiva*:

$$x_+ = \text{máx}\,\{0, x\}$$

quedando una expansión en bases truncada:

$$\begin{aligned} h(x) &= \sum_{i=1}^{M+J-1} w_i \Psi_i(x, \mathcal{P}) \\ &= \sum_{i=1}^{M} w_i \; x^{i-1} + \sum_{i=1}^{J-1} w_{M+i} \; (x - \tau_i)_+^{M-1} \end{aligned} \tag{2.12}$$

El primer sumando de (2.12), representa el *polinomio base* de grado $M - 1$ que afecta a todo el rango. El segundo sumando, está compuesto únicamente de funciones parte positivas que se van *activando* a medida que x se mueve a la derecha y va pasando por los nodos. Estas funciones parte positiva, capturan el efecto de todos los intervalos anteriores que, al combinarlos con el primer sumando definen un polinomio cúbico en todo $[a, b]$. Se hace notar, que esta derivación de las bases, surge cuando se integra un polinomio por partes constante $M - 1$ veces. En cada iteración, se juntan las constantes y se integran por si solas, independientemente de los intervalos, lo cual lleva a este *polinomio base*. De forma más explicita, tenemos las bases:

$$
\begin{aligned}
&\Psi_1(x, \mathcal{P}) = 1 \\
&\Psi_2(x, \mathcal{P}) = x \\
&\vdots \\
&\Psi_M(x, \mathcal{P}) = x^{M-1} \\
&\qquad \text{el } \textit{polinomio base} \\
&\Psi_{M+1}(x, \mathcal{P}) = (x - \tau_1)_+^{M-1} \\
&\vdots \\
&\Psi_{M+J-1}(x, \mathcal{P}) = (x - \tau_{J-1})_+^{M-1} \\
&\qquad \text{la base truncada}
\end{aligned}
$$

las cuales forman un espacio lineal de funciones $(M + J - 1)$-dimensional.

A estos splines, se les conoce como splines cúbicos y son los más usados cuando se buscan funciones suaves.

A pesar de la utilidad de los splines por su suavidad, todos sufren de problemas más allá del rango de entrenamiento. Su naturaleza global hace que, fuera de la región con nodos, los polinomios crecen o decrecen rápidamente. Por lo tanto, extrapolar con polinomios o splines es peligroso y podría llevar a estimaciones erróneas. Para corregir esto, en ocasiones, se puede imponer la restricción de que el polinomio deba ser lineal más allá de los nodos frontera. Para designarlos, se les agrega el adjetivo de *natural*. Esta modificación, libera $2 * (M - 2)$ funciones bases, pues quita todas las bases de orden mayor a 1 en los dos nodos frontera. Adicionalmente, es razonable que se mejore la fuerza predictiva fuera de el dominio de entrenamiento. Todo depende de los datos y el tipo de funciones que se esté tratando de estimar. Su expansión en bases, también se deriva de la ecuación (2.12).

Hasta ahora, se han usando los parámetros M, J y K para definir el número de funciones base N^*, ecuación (2.11), pero también, sirven para definir los *grados de libertad* que se tienen. Esto se debe a que no solo nos dicen el número de funciones bases y la dimensión del espacio lineal, sino que nos indican el número de parámetros w_j's a estimar.

Otra consideración, es que, al estar trabajando en espacios funcionales, la elección de base es relativamente arbitraria y podemos intercambiarla como lo haríamos con una transformación de coordenadas en un

espacio euclidiano. Cada base tiene sus beneficios y simplicidad. Aquí se escoge una expansión de bases truncadas pues es explicada fácilmente y tiene una forma funcional relativamente sencilla. Sin embargo, no es óptima computacionalmente cuando J es grande. En la practica, usualmente se implementan B-Splines[11] que se derivan de los vistos anteriormente.

2.3.2. Polinomios por parte flexibles

Independientemente de la selección de parametros en la construcción del polinomio, se tiene el problema de la selección de los nodos. Existen procedimientos adaptativos, como los propuestos en (Friedman 1991). Sin embargo, y como ya se mencionó más atrás, en 1998, Denison, Mallik y Smith proponen un método bayesiano más atractivo.

Para poder explicar su método, se tiene que hacerle una modificación a la ecuación (2.12) para convertirla, de un spline, a un polinomio por partes más general con grado arbitrario de continuidad en las derivadas. En su expansión de bases se tiene:

$$h(x) = \sum_{i=1}^{N^*} w_i^* \ \Psi_i(x, \mathcal{P}) \qquad \text{con } N^* = J * M - K * (J-1) \tag{2.13}$$

$$= \sum_{i=1}^{M} w_{i,0} \ x^{i-1} + \sum_{i=K}^{M-1} \sum_{j=1}^{J-1} w_{i,j} \ (x - \tau_j)_+^i \tag{2.14}$$

Dado que se tiene una doble suma, es necesario incluir un segundo índice, al menos temporalmente, a los pesos. De, modo que el primer índice, denotado por i refleja el grado asociado a su término[12]. Por lo tanto, si $i = 2$ entonces, $w_{2,j}$ está asociado a una término de grado 1. El segundo índice j denota al nodo al que está asociado el peso. Como convención, si $j = 0$, se hace referencia al *polinomio base* que siempre tendrá efecto. En el segundo sumando de (2.14) la primera suma comienza en K. Recordando, K es el número de restricciones de continuidad que se imponen al polinomio en los nodos. Por ejemplo, $K = 0$ implicaría que cada polinomio es independiente; $K = 2$, se tiene continuidad en la función y en la primera derivada, etc. En el caso de que $K = M - 1$ regresamos a la ecuación (2.12) y tenemos una vez más splines que, por construcción, son suaves. La suavidad, aunque importante, no siempre es requerida. Existen muchas funciones con primera y segunda derivada que varían rápidamente e incluso funciones discontinuas que no se podrían estimar usando splines; todo depende de los datos. Esta constucción, con su doble suma,

11. Vease el Capítulo 5.5 de (Wasserman 2007) o el Apéndice del Capítulo 5 en (Hastie, Tibshirani y Friedman 2008)

12. Desgraciadamente y para ser consistentes con la notación anterior, no se puede indexar directamente, es decir, se le tiene que restar 1 para obtener el grado en los primeros terminos, pero en los posteriores si es directo.

w_j^*	$w_{n,m}$	$\Psi_j(x,\mathcal{P})$		
Subíndice j	Subíndices n,m	Función Base		
1	$1,0$	1	M elementos	
2	$2,0$	x		
$\vdots$	$\vdots$	$\vdots$		
M	$M,0$	x^{M-1}		
$M+1$	$K,1$	$(x-\tau_1)_+^K$	$M-K$	$J-1$ veces
$M+2$	$K+1,1$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(M-K)$	$M-1,1$	$(x-\tau_1)_+^{M-1}$		
$M+(M-K)+1$	$K,2$	$(x-\tau_1)_+^K$	$M-K$	
$M+(M-K)+2$	$K+1,2$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+2(M-K)$	$M-1,2$	$(x-\tau_1)_+^{M-1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-2)(M-K)+1$	$K,J-1$	$(x-\tau_{J-1})_+^K$	$M-K$	
$M+(J-2)(M-K)+2$	$K+1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-1)(M-K)$	$M-1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		

Figura 2.4: Biyección entre w_j^*, $w_{n,m}$ y sus correspondientes funciones base Ψ_j

permite tener $M-K$ términos por nodo, codificando así las continuidades en las derivadas[13]. En la ecuación (2.13) se usa w_i^* solamente para denotar que se puede seguir expresando como una combinación lineal. Finalmente, hágase que $h(x)$ sea igual a $f_i(x_i)$. Con este cambio de notración,(2.13) es equivalente a (2.4). Este era el último componente fundamental por definir del modelo, completando así su exposición.

En la Figura 2.4 de la página 24, se hace un compendio de los polinomios por partes. Esto ayuda no solo a esclarecer las cosas, sino a formar una biyección entre w_i^*, $w_{n,m}$ y Ψ_i que posteriormente ayudará a expresar todo de forma matricial en su implementación en código.

Por lo tanto, se termina teniendo $N^* = M+(J-1)(M-K) = JM-K(J-1)$ términos una vez más. Y por construcción, como se vio anteriormente, la biyección, es consistente con la definición en (2.12) para el caso especifico que $K = M-1$.

13. Esta codificación es sutil pues, al hacer los cálculos de continuidad, tenemos que considerar los límites izquierdos y derechos, los cuales existen siempre. Sin embargo, los términos $(x-\tau)_+^K$ se desvanecen únicamente hasta la K-esima derivada. Para la $(K+1)$-derivada, el coeficiente correspondiente se suma a la función y rompe la continuidad pues no corresponde con el límite izquierdo

Antes de cerrar el capítulo, se centra la atención en los nodos τ. A estos, se les ha dado poca importancia hasta el momento, pues se han considerado como fijos. Como ya se mencionó antes, en (Denison, Mallick y Smith 1998) se desarrolla, además de la ecuación (2.14) un paradigma bayesiano para que los nodos, sean tratados como parámetros y por ende sus posiciones son variables. La ventaja de que estos estén indeterminados, es que se pueden concentrar en los lugares donde la función varia más. Y al contrario, si la función es relativamente plana para alguna sección, se necesitan pocos nodos. En el Capitulo **??**, se continua con esta exposición y se detalla el proceso para la selección de la posición de los nodos. Sin embargo, cabe recalcar que a diferencia del trabajo original donde el número de nodos era variable, en este trabajo se usa J fija. Esto corresponde a que se busca simplificar el algoritmo sin tener que usar métodos que aumenten o disminuyan el número de dimensiones. En la practica, la J se tiene que calibrar, sin embargo, no ha resultado ser un problema adicional pues normalmente, se busca suavizar más que estimar estimar funciones especificas complejas como era el objetivo del trabajo original.

Consideraciones finales

Al tener en mente que se tienen d covariables, y por ende d polinomios por partes, además de la estructura lineal de (2.13) podemos sustituir (2.4) dentro de (2.3) y se tiene la siguiente estructura con doble suma:

$$\begin{aligned} f(\mathbf{x}) &\approx \sum_{i=0}^{d} \beta_i f_i(x_i) \\ &\approx \beta_0 + \sum_{i=1}^{d} \beta_i \left[\sum_{j=1}^{N^*} w_{i,j} \Psi_{i,j}(x_i, \mathcal{P}_i) \right] \end{aligned}$$

Lo cual, es perfectamente lineal. Se tienen $1 + d * N^*$ términos y se pueden acomodar en un solo vector. Sin embargo, se tiene un cruce de parámetros interesante, la multiplicación de la $\beta_i \quad \forall i$ contra $w_{i,j} \quad \forall j$. Tradicionalmente, no se usan β's y se deja que se capture ese efecto dentro de las f_i como en los modelos aditivos normales. Sin embargo, dado que el objetivo de este trabajo es la predicción, más que la estimación de funciones, se opta por dar una nueva capa de suavizamiento con las β's. No existe forma de garantizar ortogonalidad de las β's contra las w's, por lo tanto, se le da prioridad a la correcta estimación de w pues captura un mayor efecto además de que, por construcción de los polinomios por partes, si está garantizada la ortogonalidad contra las funciones bases Ψ'.

Capítulo 3

Paradigma bayesiano e implementación

El objetivo de este capitulo es ilustrar como se pasa de un modelo completamente teórico, a una implementación con código. Específicamente se desarrollará el código en el software estadístico `R`. Además, se hace una exposición de algunos de los métodos más recientes y sofisticados para la estimación bayesiana y frecuentista de los parámetros del modelo.

De forma global, el algoritmo trata de estimar de forma conjunta los cuatro parámetros: $\boldsymbol{\tau}, \mathbf{w}, \boldsymbol{\beta}$ y σ^2. A diferencia de la exposición del modelo y dada su estructura, el algoritmo debe de construir de *abajo hacia arriba*. La idea, es que el algoritmo pueda entrenarse de tal forma, que los parámetros reflejen, hacia abajo y hacia lo largo, las estructuras en las covariables, que lleven a predecir si y es exito o fracaso. Se considera, una buena forma de entender el algoritmo es *visualizando* tanto los datos como los objetos que componen el modelo, por lo tanto se hace un paréntesis notacional.
Tenemos datos $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ que podemos representar en una tabla (o matriz):

$$\left[\begin{array}{c|ccc} y_1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,d} & \cdots & x_{n,d} \end{array}\right]$$

Donde el vector de observaciones binarias es $\mathbf{y}$, y la matriz de covariables $\mathbf{X}$. Se habla de que la estimación debe reflejar los patrones *hacia abajo* y *hacia lo largo*. Hacia abajo, pues se ve que cada f_j con $j = 1, \ldots, d$ mediante $\mathbf{w}$ y $\boldsymbol{\tau}$ codifica toda la columna de datos j. Hacia lo largo, pues la función de proyección f suma (de forma ponderada y a través de las *betasn*), estos efectos individuales hacia lo largo de la tabla de datos. Este balance es fundamental para la correcta predicción.
En forma de pseudocódigo el algoritmo tiene la siguiente forma:

```
Parametros inciales:
```

```
WHILE (...)
    Transformación de X -> Phi -> F (Función: estimate_PWP)
    Simulación de betas (Función simulate_beta)
```

- Usamos los nodos iniciales en cuantiles determinados.

- taus: HMC
- β Estimar por máxima verosimilitud pero dentro del Gibbs con el método ABC
- $w's$ BAyesianas + importantes que las betas.

3.1. Fundamentos de la estadística bayesiana

- Hacer preambulo bayesiano y justificación de la fislosofía bayesiana - Sacar posterior-ish - Hacer derivación de la condicional para β paper de Albert + Chibb - Argumentar por qué es igual para w - Distros a Priori

3.2. Especificación para el modelo

Para los parámetros, se usan las siguientes distribuciones *apriori*:

$$\beta = (\beta_0, \beta_1, \ldots, \beta_d)^t \sim N_d(\mu_0, \Sigma_0) \tag{3.1}$$

$$w^{(i)} = (w_1^{(i)}, \ldots, w_J^{(i)})^t \sim N_J(\mu_0^{(i)}, \Sigma_0^{(i)}) \quad i = 1, \ldots, d \tag{3.2}$$

3.2.1. Gibbs Sampler para datos binarios

3.3. Funciones de probabilidad condicional completas

3.4. Algoritmo

- Explicar Alortimo y hacer pseudocódigo de cada sección - Explicar lógica del algoritmo - Explicar desarrolo de paquetes en R

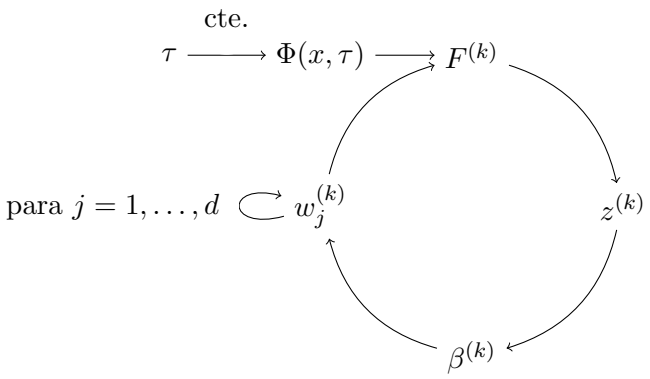

Figura 3.1: Esquema del algoritmo

Capítulo 4

Ejemplos y resultados

4.1. Análisis a fondo de un ejemplo sencillo

- Hacer todo el análisis de los grupos simulados con las normales bivariadas.

- Hacer varios tipos de polinomios con varias fronteras - Comparar contra un GLM probit normal

4.2. Otros resultados interesantes

- Dar gráficas y resultados de: la normal modificada, datos parabolicos, circulares y hopefully Ying-Yang.

4.3. Prueba con datos reales

Hacer prueba con datos SVSS y con datos de Hastie and Tibsh

Capítulo 5

Conclusiones

5.1. Consideraciones adicionales y posibles mejoras

- Y después? Mejoras al modelo - Seleccion de varaibles con SVSS - Automatización de selección de parámetros , $Jy\tau$ usando Mallik, M usando exploración previa de datos por dimensión.

- Problemas que tuve - Derivación bayesiana de las w0s

Listado de Assumptions - Assumptions: - 0. Y se distribuye bernoulli - 1. Existe y funciona f (aproximaxición) - 2. Las dimensioens son independientes entre si

5.2. Extensiones y alternativas al modelo

- Después: - Las redes neuronales se basan en coass parecidas. Ish hacer análogía - Cita de Articulo de ML para Cosas financieras

Apéndice A

Análisis Funcional

A.1. Convergencia del modelo

Habiendo entendido la estructura de las funciones que componen este proyecto, se busca demostrar (o dar una idea) por que se tiene una aproximación a f y a f_i's y no una igualdad escrita. Esto se logra, usando principios de álgebra lineal y análisis funcional. Esta discusión sigue los principios planteados por (Bergstrom 1985) en donde se demuestra el caso univariado y una pequeña discusión sobre los *Espacios de Hilbert de Kernel Reproductivo* (RKHS)[1]

Para entender el teorema de convergencia, necesitamos considerar los Espacios de Hilbert. Como introducción a estos, se usa el ejemplo clásico de plantearlo como una generalización del caso euclidiano. En este espacio euclidiano normal $\mathbb{R}^n$ podemos representar cualquier vector como una combinación lineal de un conjunto de bases ortogonales. En espacios abstractos de dimensión infinita, en particular en espacios de funciones, se busca representar *cualquier función*, en este caso f_i's, como una combinación lineal de bases. Los espacios de Hilbert dan idea de que los espacios vectoriales pueden ser suficientemente abstractos para que los vectores no sean simplemente listas ordenadas de números como lo son en $\mathbb{R}^n$. Los vectores pueden ser cualquier objeto en este caso, funciones. Una vez definido el espacio vectorial y sus objetos (que cumplan los correspondientes 8 axiomas presentados en el Apéndice D) se les puede denota de *producto interno* y por consecuente una *métrica* la cual induce una topología.

1. Reproducing Kernel Hilbert Space a falta de una mejor traducción.

Formalización matemática y teorema de convergencia

Se dice que $\mathcal{H}$ es un Espacio de Hilbert si $\mathcal{H}$ es un espacio vectorial con producto interno que también es un espacio metrico completo. (Rudin 1987).

Para hacer la prueba de convergencia, se considera únicamente a las funciones f_i y no a la f general. Se estudia en particular el espacio de Hilbert $\mathcal{H} = \mathbf{L}_2(\mathbf{R}, \mu)$ el *espacio de funciones integrables al cuadrado en* $\mathbf{R} = [a, b]$ con medida de Lesbegue ordinaria μ. Es decir:

$$f \in \mathcal{H} \iff \int_a^b f(x)^2 \, dx < \infty$$

Donde el producto punto es:

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx$$

y su norma inducida:

$$||f||_{\mathcal{H}} = \langle f, f \rangle = \int_a^b f(x)^2 \, dx$$

Antes de que se presente el teorema de Bergstrom, se tienen que presentar los tres supuestos fuertes que hace:

1. Las variables aleatorias $y_1, \ldots, y_n$ son generadas por la ecuación:

$$y_i = h(x_i) + e_i \quad \forall i = 1, \ldots, n$$

y los valores de las covariables están dados por las ecuaciones:

$$\begin{aligned} x_1 &= a + \frac{b-a}{2n} \\ x_{i+1} &= x_i + \frac{b-a}{n} \end{aligned} \tag{A.1}$$

Con a, b los extremos del intervalo $b > a$ y e_i son ruido aleatorio cumpliendo:

$$\begin{aligned} \mathbb{E}[e_i] &= 0, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i^2] &= \sigma^2, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i e_j] &= 0, \quad \forall i, j = 1, \ldots, n \; i \neq j \end{aligned} \tag{A.2}$$

2. La función $h(x)$, está definida en el intervalo cerrado $[a, b]$ es acotada y continua en casi todas partes.

3. El conjunto contables de funciones base $\Psi_1(x), \Psi_2(x), \ldots$ es un conjunto *ortonormal* en $\mathcal{H}$. Es decir, estas funciones cumplen:

$$\begin{aligned} \int_a^b \Psi_j^2(x)\, dx &= 1 \quad \forall j \\ \int_a^b \Psi_j(x)\Psi_k(x)\, dx &= 0 \quad \forall k, j \; k \neq j \end{aligned} \tag{A.3}$$

Estos supuestos son bastante fuertes y hay ciertas ecuaciones que no se cumplen *per se* en el modelo propuesto, sin embargo, vale la pena analizar el resultado pues lleva a cosas aún más interesantes. El primer supuesto es el más problemático. Aunque el modelo generador es idéntico a (2.8) y el ruido aleatorio es un supuesto aceptable (y común) el problema está en (A.2) pues para este trabajo no se asume que el estadista fija las x's sino que se asume una muestra aleatoria de datos. Sin embargo, en la prueba, este supuesto se usa para argumentar que, si $n \to \infty$, los datos cubren de manera homogénea todo el intervalo aproximando una integral. Aunque el propósito es completamente diferente que el de este trabajo en el que se busca suavizar sobre datos dispersos, se decide obviar por ahora el supuesto, en interés de presentar el teorema en su forma más rigurosa.

El segundo supuesto no es nada descabellado y se ha usado con anterioridad. Además, aún permite aproximar un número grande de funciones y es igual de flexible que el modelo anterior. Sin embargo, este supuesto si implica que $h \in \mathcal{H}$. Por lo que esta puede ser representada en su combinación lineal de bases funcionales, es decir:

$$h(x) = \sum_{i=1}^{\infty} w_i \Psi_i(x)$$

diferente a la expansión de bases en de la ecuación (2.7). Esto se deriva, de que ahora se busca encontrar

una representación *exacta* de h.[2] El último supuesto, implica la construcción de una base *ortonormal*. El trabajo original, sugiere que se puede lograr una base completa, aplicando un proceso de ortonormalización a las bases canónica polinomial $\{1, x, x^2, \ldots\}$. Por lo tanto, al menos de forma teórica, la base escogida para este trabajo definida en (2.14) también es ortonormalizable independientemente de la elección de J, N y K. Por lo que se cumple el supuesto. Finalmente:

Teorema 1 *Sea $\hat{h}_n^{N^*}(x)$ el estimador de h definido por:*

$$\hat{h}_n^{N^*}(x) = \hat{w}_1\Psi_1(x) + \ldots + \hat{w}_{N^*}\Psi_{N^*}(x) \tag{A.4}$$

que depende del número de datos n y el número de bases funcionales N^*. Y, con $\hat{w}_i \quad i = 1, \ldots, N^*$ los estimadores de mínimos cuadrados, es decir, los valores de $w_i \quad i = 1, \ldots, N^*$ que minimizan la expresión:

$$\sum_{i=1}^{n} [y_i - w_1\Psi_1(x_i) - \ldots - w_{N^*}\Psi_{N^*}(x_i)]^2 \tag{A.5}$$

Para todo $\epsilon > 0$, bajo los supuestos 1 a 3, existe un entero N^* y una función $n_\epsilon(N^*)$ tal que:

$$\mathbb{E}\left[\int_a^b \left(\hat{h}_n^{N^*}(x) - h(x)\right)^2 dx\right] < \epsilon \qquad \forall N \geq N^* \ y \ \forall n \geq n_\epsilon(N^*) \tag{A.6}$$

Detrás de toda esta verborrea y notación aparatosa, el corazón del teorema está en que, bajo ciertos supuestos, rigurosos más no descabellados, y en caso de existir una función h que genere los datos, está se puede aproximar a un grado de precisión arbitraria.

Aunque no es el objetivo del trabajo, vale la pena hacer una mención a lo sublime que es la demostración, pues utiliza conceptos de análisis, álgebra lineal, optimización y estadística. Los detalles y la utilización de los supuestos, son sutiles, sin embargo es una prueba rigurosa en todo el sentido de la palabra. Además, (Bergstrom 1985) va mucho más allá de únicamente demostrar la existencia. Se demuestran tres teoremas más, para dar estimaciones (bajo un supuesto adicional) de el tamaño de muestra necesario n y el número de bases N^* necesarias para la aproximación arbitraria de h. Sin embargo, este procedimiento, aunque elegante, no es nada practico pues depende de poder generar a merced las x's (con su correspondiente nivel y) aumentando y disminuyendo el tamaño de muestra. Además, se requiere ir generando todas las bases Ψ's de forma que sean ortonormales y sus correspondientes coeficientes de Fourier $w_j = \int_a^b h(x)\Psi_j(x)\,dx \quad \forall j$.

2. Antes se buscaba, más que aproximarla, suavizar los datos. Además, se usa el signo de igualdad para no introducir confusión en la exposición.

El resultado, es más bien teórico en el sentido de que, justifica que estos modelos tienen sentido. Para este trabajo, en especifico, da la *intuición* de que funcionará (obviando un poco el primer supuesto) pues, con una muestra suficientemente grande, las f_i's serán identificables y aunque sean aproximaciónes, estaremos captando los patrones subyacentes y con suerte, podremos hacer predicciones.

Se hace notar que el primer supuesto, da la intuición de *granularidad* de el intervalo $[a, b]$. Bajo la construcción de Bergstrom, tenemos las condiciones exactas para estimar $h(x)$. Sin embargo, en la practica es raro que el estadista tenga el control sobre las covariables y siempre tendremos datos aleatorios. A pesar de esto, al estar trabajando sobre intervalos cerrados, se puede suponer que $X \sim \mathrm{U}(a, b)$ de donde tenemos que si $n \to \infty$ cubrimos todo el intervalo y tenemos algo análogo al supuesto 1 y por ende, el teorema es válido.

Otro teorema de convergencia y RHKS

Este resultado de Bergstrom, es uno de los muchos teoremas que se probaron en la época para justificar la existencia de estos modelos. Otro resultado interesante del mismo año, viene dado por (Stone 1985). El, discute varios modelos posibles dada una estructura de datos practica. Plantea un GAM[3] tradicional $h^*(\mathbf{x}) = \sum h_i^*(x_i)$ con la restricción (2.10), y prueba que

$$\mathbb{E}[(h(x) - h^*(x))^2]$$

es mínimo, con la igualdad si h es en verdad aditiva. Además, los estimadores de h_j^* que da, son splines (los mismos que usan Hastie y Tibshirani).

Finalmente, se hace notar, que toda esta teoría y resultados, son casos específicos de una teoría más general y mucho más compleja, llamada *Espacios de Hilbert de Kernel Reproductivo* (RKHS) desarrollada en los años 90. Esta va mucho más allá del enfoque de este trabajo, sin embargo, vale la pena mencionarla pues engloba muchos de los modelos usados hoy en día en un marco matemático riguroso y basado en el análisis funcional. Muchas de las ideas presentadas en este trabajo, como lo son la regularización, las expansiones de bases y los espacios de Hilbert, se elevan varios niveles y llevan a resultados todavía más generales. Se recomienda el Capítulo 5.8 de (Hastie, Tibshirani y Friedman 2008) y el libro de (Wahba 1990) donde se discuten a detalle todas las consideraciones de los RKHS. Además, todos estos resultados

3. T. Hastie y R. Tibshirani publicaron preámbulos a los GAM antes del trabajo citado aquí de 1986. Además, de que la cercanía, Stone en Berkley y ellos en Stanford, ayudó a su colaboración.

son *deterministas* en sus parámetros, en contrapuesta de la filosofía bayesiana. Sin embargo, esto no le quita validez al modelo ni mucho menos a los resultados.

Apéndice B

Paquete en R. Desarrollo y Lista de Funciones

- Hacer un resumen de como se desarrolló todo el paquete. - Citar al buen H Wickham y su libro - Hacer un listado y un diagrama de las funciones que existen en el y como descargarlo - Mencionar los métodos S3 de `plot` y `summary`

Apéndice C

Notación

- **Datos:** $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$
- $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ variables de respuesta binarias.
- $\mathbf{x}_i \in X^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ covariables o regresores.
- $d \in \mathbb{N}$ dimensionalidad de mis regresores.
- $z_i \in \mathbb{R}$ variables latentes.
- $\Phi(\cdot) : \mathbb{R} \to (0, 1)$ la función de distrubución acomulada de una normal estandar.
- **Parámetros**: $\theta = (\sigma^2, \{\phi_i\}_{i=1}^d\}, \{\tau_j\}_{j=1}^J)$ donde,
- σ^2 es la varianza global de mis datos.
- $\{\phi_i\}_{i=0}^J$ los pesos de mi combinación lineal.
- $\{\tau_i\}$ los parámetros de mi función Ψ
- $f(\mathbf{x}_i)$ función de media para Z
- J orden de la expansión en bases.

Apéndice D

Definiciones

Espacio con producto interno Un espacio vectorial dotado de una estructura adicional llamada *producto interno*: $\langle \cdot, \cdot \rangle$, que asocia cada par de vectores con una cantidad escalar sobre F. Es decir, $\langle \cdot, \cdot \rangle : V \times V \to F$. Que cumple, para x, y, z vectores en V y a en F:

- $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- $\langle ax, y \rangle = a\langle x, y \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, x \rangle \geq 0$
- $\langle x, x \rangle = 0 \Leftrightarrow x = \mathbf{0}$

Espacio Funcional Un espacio funcional es un espacio vectorial cuyos elementos son funciones.

Espacio Metrico Un espacio métrico es un espacio donde la distancia (norma) inducida por el producto punto está definido sobre todos sus elementos. Norma: $||x|| = \sqrt{\langle x, x \rangle}$ la raiz no negativa del producto interno.

Espacio Métrico Completo Un espacio métrico es completo si todas las secuencias de Cauchy, convergen a puntos dentro del espacio.

Espacio Vectorial Un espacio vectorial sobre un campo F es un conjunto V, dotado de dos operaciones, *suma* $+$ y *multiplicación escalar* $\cdot$ que cumple los siguientes axiomas. Sean x, y, z vectores en V, y a, b escalares en F

1. $x + (y + z) = (x + y) + z$
2. $x + y = y + x$
3. $\exists 0 \in V$ tal que, $x + 0 = x$

4. $\forall x \in V \quad \exists -x \in V$ tal que, $x + (-x) = 0$
5. $a(bx) = (ab)x$
6. $\exists 1 \in F$ tal que, $1x = x$
7. $a(x + y) = ax + ay$
8. $(a + b)x = ax + bx$

Ortogonalidad Dos elementos son ortogonales (en cierto espacio) si $\langle x, y \rangle = 0$. Denotado $x \perp y$

Bibliografía

Albert, J.H., y S. Chib. 1993. "Bayesian analysis of binary and polychotomous response data". *Journal of the American Statistical Association:* 669-679.

Barber, D. 2012. *Bayesian reasoning and machine learning.* Cambridge University Press.

Bergstrom, A. R. 1985. "The Estimation of Nonparametric Functions in a Hilbert Space". *Econometric Theory* 1 (01): 7-26.

Bishop, C M. 2006. *Pattern Recognition and Machine Learning.* Springer.

Boor, C De. 1978. *A Practical Guide to Splines.* 346. New York, Springer-Verlag.

Box, George E. P. 1979. *Robustness in the Strategy of Scientific Model Building.* p. 74. May. RL Launer / GN Wilkinson.

Denison, DGT, BK Mallick y AFM Smith. 1998. "Automatic Bayesian curve fitting". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (2): 333-350.

Friedman, Jerome H. 1991. "Multivariate adaptive regression splines". *The Annals of Statistics:* 1-67.

Härdle, Wolfgang, Marlene Müller, Stefan Sperlich y Axel Werwatz. 2004. *Nonparametric and semiparametric models.* Springer Verlag.

Hastie, T., R. Tibshirani y J. Friedman. 2008. *The elements of statistical learning,* volumen 1. Springer Series in Statistics.

Hastie, Trevor, y Robert Tibshirani. 1986. "Generalized additive models". *Statistical science:* 297-310.

James, Gareth, Daniela Witten, Trevor Hastie y Robert Tibshirani. 2013. *An introduction to statistical learning.* Springer.

MacCullagh, P., y J. A. Nelder. 1989. *Generalized Linear Models.* Chapman & Hall, London.

Rudin, Walter. 1987. *Real and complex analysis.* Tata McGraw-Hill Education.

Schoenberg, I J. 1964. "Spline Interpolation And The Higher Derivatives." *Proceedings of the National Academy of Sciences of the United States of America* 51, número 1 (): 24-8.

Stone, Charles J. 1985. "Additive regression and other nonparametric models". *The annals of Statistics:* 689-705.

Sundberg, Rolf. 2016. *Statistical Modelling by Exponential Families - Lecture Notes.* Stockholm University.

Wahba, G. 1990. *Spline Models for Observational Data.* Society for Industrial / Applied Mathematics.

Wasserman, Larry. 2007. *All of nonparametric statistics.* Springer.