Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2009-18

Predicción de Bancarrota: Una Comparación de Técnicas Estadísticas y de Aprendizaje Supervisado para Computadora

Tonatiuh Peña
Banco de México

Serafín Martínez
Banco de México

Bolanle Abudu
University of Essex

Diciembre 2009

Documento de Investigación
2009-18

Working Paper
2009-18

Predicción de Bancarrota: Una Comparación de Técnicas Estadísticas y de Aprendizaje Supervisado para Computadora*

Tonatiuh Peña†
Banco de México

Serafín Martínez‡
Banco de México

Bolanle Abudu§
University of Essex

Resumen

Estamos interesados en predecir bancarrotas de una forma probabilística. En específico, comparamos el desempeño de varias técnicas estadísticas de clasificación y de aprendizaje supervisado, ellas son: análisis de discriminantes (Z-score de Altman), regresión logística, máquinas de soporte vectorial de mínimos cuadrados y diferentes instancias de Procesos Gaussianos (GP's) – estos son los GP's para clasificación, el discriminante Bayesiano de Fisher y los GP's deformados. Nuestra aportación al campo de las finanzas computacionales consiste en introducir a los GP's como un marco potencialmente competitivo para predecir bancarrota. Datos provenientes del Seguro de Depósito de los Estados Unidos son usados para probar la calidad de las predicciones.

Palabras Clave: Predicción de bancarrota, Inteligencia artificial, Aprendizaje supervisado, Procesos Gaussianos, Z-Score.

Abstract

We are interested in forecasting bankruptcies in a probabilistic way. Specifically, we compare the classification performance of several statistical and machine-learning techniques, namely discriminant analysis (Altman's Z-score), logistic regression, least-squares support vector machines and different instances of Gaussian processes (GP's) -that is GP's classifiers, Bayesian Fisher discriminant and Warped GP's. Our contribution to the field of computational finance is to introduce GP's as a potentially competitive probabilistic framework for bankruptcy prediction. Data from the repository of information of the US Federal Deposit Insurance Corporation is used to test the predictions.

Keywords: Bankruptcy prediction, Artificial intelligence, Supervised learning, Gaussian processes, Z-score.

JEL Classification: C11, C14, C45.

---

*Agradecemos a José A. Murillo, Fabrizio López-Gallo y Pascual O'Dogherty por su apoyo en realizar este trabajo y a Santiago García y Enrique Covarrubias por sus valiosos comentarios. Antonio Noriega y Rocío Elizondo proporcionaron apoyo editorial.

† Dirección General de Investigación Económica. Email: tpena@banxico.org.mx.

‡ Dirección General de Análisis del Sistema Financiero. Email: smartin@banxico.org.mx.

§ University of Essex. Email: bsabud@essex.ac.uk.

1. Introducción

La bancarrota corporativa es un área activa dentro de la investigación financiera, porque un evento de esta naturaleza siempre provocará efectos adversos en la economía y planteará un reto de credibilidad para las autoridades financieras. De hecho, el pronóstico de bancarrota juega un papel de gran importancia para diferentes tipos de organizaciones gubernamentales y comerciales, dado que una compañía quebrada puede causar perturbaciones contagiosas al resto del sistema financiero y así provocar una crisis sistémica. Tal importancia ha aumentado más por regulaciones tales como el Comité Basilea de Supervisión Bancaria (2004) o Basel II, que sugiere que instituciones financieras deberían formar sus portafolios crediticios con base en la evaluación de incumplimiento de sus clientes. Como consecuencia, el desarrollo de herramientas analíticas para determinar cuál información financiera es más relevante para predecir estrés financiero ha ganado popularidad junto con el diseño de los sistemas de alerta temprana que predicen bancarrota.

A lo largo de los años, dos metodologías principales se han desarrollado para asistir en el proceso de estimar estrés financiero (es decir, predecir bancarrotas). La primera metodología usa información contable, mientras que la segunda usa información de mercado. La primer incluye, por ejemplo, el análisis de razones financieras, una técnica que estudia relaciones del tipo X/Y donde $X, Y \in \mathbb{R}$ son variables seleccionadas del estado financiero de una empresa. Aunque no existe consenso acerca de la definición o cálculo de razones financieras, las podemos dividir en cuatro categorías: eficiencia, rentabilidad, y razones de solvencia de corto y largo plazo. Beaver (1966) propuso el trabajo fundamental sobre la predicción de quiebras financieras mediante el análisis de razones financieras, donde éste se puede entender como una técnica de clasificación univariada para estimar la probabilidad de quiebra. Posteriormente, Altman (1968) trabajó en una generalización a través de la estimación de un estadístico multivariado, conocido como *Z-score*.

Mientras que estos dos métodos han comprobado ser útiles por al menos cuarenta años, el advenimiento de nuevas regulaciones tales como Basel II justifica el uso de técnicas más sofisticadas para predecir estrés financiero. Entre estas nuevas metodologías, recientemente se ha desarrollado un grupo con un componente computacional importante. Por ejemplo, los problemas de valoración de bienes, asignación de portafolios y predicción de bancarrota han sido enfocados desde diferentes perspectivas, tales como algoritmos genéticos (GA's, por sus siglas en inglés), redes neurales artificiales (ANN's, por sus siglas en inglés), arboles de decisión, entre otras. Utilizaremos el término *finanzas computacionales* (Tsang y Martinez-Jaramillo, 2004, Chen, 2002) para referir al desarrollo y la aplicación de dichos tipos de técnicas para resolver problemas financieros y algunas referencias sobre estos temas están disponibles en Serrano-Cinca et al.(1993), Back et al. (1996), Joos et al. (1998), Varetto (1998), Atiya (2001), Shin y Lee (2002), Park y Han (2002), Yip (2003) y Quintana et al. (2007).

Hasta donde sabemos, este trabajo es el primero que aplica el formalismo de procesos Gaussianos (GP's, por sus siglas en inglés) desde la perspectiva de

inferencia de datos (O'Hagan, 1978) para estimar las probabilidades de bancarrota. Desde una perspectiva Bayesiana, los GP's constituyen una forma natural para aprender de una función de regresión o clasificación, en términos de distribuciones de probabilidad a priori definidas sobre el espacio funcional y algunas muy buenas monografías sobre este tema se publicaron en años recientes, por ejemplo Rasmussen y Williams (2006). Nuestra aportación al campo consiste en presentar una comparación de técnicas estadísticas clásicas de clasificación con algunos algoritmos recientemente desarrollados de machine learning. Específicamente, introducimos a los GP's como un marco probabilístico competitivo y potencialmente de gran poder para predecir la bancarrota. Como beneficio adicional de trabajar en el ámbito de los GP's, sugerimos una característica que permite determinar la relevancia de diferentes razones financieras de manera automática, algo conocido como determinación automática de relevancia (ARD, por sus siglas en inglés) en la literatura de redes neurales.

Aunque los métodos presentados en este documento se pueden aplicar a cualquier tipo de empresa que maneja razones financieras, debido a la disponibilidad de datos nos enfocamos al sector bancario[1]. Analizar bancarrota en el sector bancario implica considerar que este tipo de instituciones tiene que satisfacer requerimientos legales y de contabilidad muy específicos impuestos por autoridades financieras. Por lo tanto, es adecuado considerarlos como un caso especial en el ámbito de bancarrota corporativa. De hecho, generalizar esta tarea para países diferentes se complica aún más si consideramos que algunas de sus propias regulaciones no contemplan la existencia de bancarrotas.

Lo que resta del documento está organizado de la siguiente manera: la Sección 2 introduce la predicción de bancarrota como un problema estadístico de clasificación. Las Secciones 3 y 4 se dedican a la descripción de algunas técnicas estadísticas bien conocidas utilizadas para la predicción de bancarrota, es decir, análisis de discriminantes y regresión logística. La Sección 5 describe los detalles acerca de cómo una familia de procesos estocásticos, es decir, procesos Gaussianos, se puede utilizar para clasificar datos y por lo tanto, aplicarse a nuestro problema. La Sección 6 describe los experimentos realizados para un conjunto de datos del Seguro de Depósito de los Estados Unidos con el fin de evaluar cómo los procesos Gaussianos se desempeñan con respecto a otros tipos de clasificadores. La Sección 7 discute cómo los GP's podrían ser integrados a modelos de riesgos crediticio comercialmente disponibles. Finalmente, la Sección 8 llega a algunas conclusiones sobre los métodos propuestos y plantea trabajo futuro.

[1] El trabajo de Estrella et al. (2000) tiene un alcance similar al nuestro.

2. Predicción de bancarrota como problema de clasificación

Estamos interesados en predecir la quiebra de bancos y además en asignar un valor de probabilidad para cuantificar nuestro grado de expectativas de que dicho evento ocurrirá. Con el fin de realizar lo anterior, nos acercamos al problema de predicción de bancarrota considerándolo como uno de clasificación binaria, donde cada elemento de un conjunto de datos observados pertenece a un grupo de clases predefinidas (bancarrota o no bancarrota) y el objetivo es intentar separar las dos clases con un error tan pequeño como sea posible. Por lo tanto, nuestro objetivo es tener un sistema que prediga si la institución irá (o no irá) a la quiebra de acuerdo con algún tipo de información financiera, por ejemplo, a través de las razones financieras de las instituciones. Dicho tipo de tarea es conocido como clasificación por la comunidad estadística y como aprendizaje supervisado por la comunidad de machine learning.

En las dos secciones siguientes revisamos algunos de los métodos de clasificación de datos más populares, entre ellos, el análisis de discriminantes de Fisher y la regresión logística. Por lo tanto, para empezar la discusión, asumimos lo siguiente: (i) una tarea de clasificación donde una observación nueva $O^\star$ tiene que asignarse a una de k clases disponibles conocidas *a priori*; (ii) que dichas clases son mutuamente excluyentes; (iii) que por alguna razón el procedimiento de asignación depende de la aplicación de un método indirecto. *Indirecto* significa que se utiliza un vector de características $\mathbf{x}^\star$ en lugar de observaciones $O^\star$. Asumimos la disponibilidad de datos de entrenamiento etiquetados correctamente y en consecuencia, que existe una manera exacta para clasificar las observaciones, pero que por alguna razón no es factible aplicarla. Por ejemplo, en medicina, el diagnóstico (identificar una enfermedad) y la prognosis (estimar la perspectiva de recuperación) son ejemplos típicos donde la clasificación directa no se puede aplicar (MacLachlan, 1991).

Otro caso adecuado para la clasificación indirecta es la determinación del nivel del estrés financiero de una compañía, dado que una evaluación directa no se puede generar; es decir, la solidez financiera de una empresa no se puede determinar de una revisión directa. En lugar de eso, es más apropiado recurrir a métodos indirectos, como las razones financieras de una empresa para determinar si irá (o no) a la quiebra.

Un enfoque común para aplicar esta estrategia se basa en la recuperación de las razones financieras observadas de un número N de empresas a lo largo de un intervalo de tiempo T; donde la razón financiera de cada banco se denota por un vector $\mathbf{x}_{n,t}$; con subíndice $n \in \{1, N\}$ denotando al banco y $t \in \{1, T\}$, al tiempo cuando los datos se observaron. Dado que las bancarrotas se repiten raramente, los datos están considerados usualmente invariantes en el tiempo y tal dependencia se omitió en este trabajo, es decir, se asume el siguiente supuesto $\mathbf{x}_{n,t} = \mathbf{x}_n$, que básicamente consiste en hacer los datos independientes e idénticamente distribuidos. La investigación sobre bancarrota ha asumido ampliamente este supuesto, por ejemplo, Altman (1968) y Atiya (2001). Cabe

señalar que eliminando el subíndice t, los datos se consideran efectivamente como una "fotografía fija" y de hecho, en el presente trabajo se utilizó dicho enfoque debido a las limitaciones de los datos. Enfoques alternativos para estimar el estrés financiero de una corporación pero que incorporan la dimensión temporal de datos son, por ejemplo, McDonald y van de Gucht (1999) o Duffie et al. (2007).

3. Análisis de Discriminantes de Fisher

El análisis de discriminantes es una técnica popular de clasificación desarrollada por Fisher (1936) con el fin de resolver un problema de clasificación de restos humanos que se le pidió resolver[2]. Esta técnica encuentra la relación que existe entre el conjunto de datos y sus correspondientes etiquetas (MacLachlan, 1991) y su objetivo es especificar la relación en términos de una función que idealmente separa cada elemento dentro del conjunto de datos de entrenamiento de acuerdo al valor de su etiqueta. En el resto del documento, denominamos al análisis de discriminantes como FDA (por sus siglas en inglés). En la presente sección revisamos brevemente FDA para el caso especial de una clasificación binaria y de tal forma que siente la base para la introducción de la regresión logística y de los procesos Gaussianos para clasificación. Nos enfocamos en el análisis de discriminantes porque constituye la base del *Z-score* de Altman que es una de las mejores y más conocidas técnicas para evaluar estrés financiero.

3.1. Definición del problema

Se considera un conjunto de datos de entrenamiento $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{\mathbf{x}^n, y^n\}_{n=1}^N$ donde una observación única en un espacio de d-dimensiones se denota por $\mathbf{x}^{(n)}$ y la variable categórica o la etiqueta asignada a dicha observación se denota por $y^{(n)} \in \{1, 0\}$. Un dato $\mathbf{x}^{(n)}$ consiste en un conjunto de razones financieras registradas en un punto fijo en el tiempo para un banco dado n, que estuvo (o no) en bancarrota en este momento, es decir, $y^{(n)}$. En términos matemáticos, el objetivo del análisis de discriminantes es asignar una nueva observación $O^\star$ en una de las $k = 2$ clases disponibles y el discriminante hará eso a través de encontrar un vector de parámetros $\mathbf{w}$ que será óptimo de alguna manera. De hecho, el espacio $\mathbb{R}^d$ se dividirá en k regiones por hiperplanos en $\mathbb{R}^{d-1}$ para realizar la separación.

El proceso se explica mejor de manera pictórica. La Gráfica 1 muestra un conjunto de datos compuesto de dos clases separado por una función de discriminantes $D(\mathbf{w})$ perpendicular a $\mathbf{w}$. Cada punto de datos $\mathbf{x}^{(n)}$ es proyectado sobre $\mathbf{w}$, tal que la distancia entre la media proyectada $d = (\mu_0 - \mu_1)$ es lo más amplia posible, mientras que la dispersión alrededor de las proyecciones $\left(\sigma_0^2 + \sigma_1^2\right)$ es lo más pequeña posible. La proyección se realiza al tomar el producto escalar $f^{(n)} = \mathbf{w}^T \mathbf{x}^{(n)}$ $(\forall n)$, por lo tanto, la calidad de la solución depende

[2]Se requirió determinar el sexo de algunos restos humanos descubiertos en un sitio funerario en Egipto, es decir, si pertenecieron al espécimen masculino o femenino (Fisher, 1936).

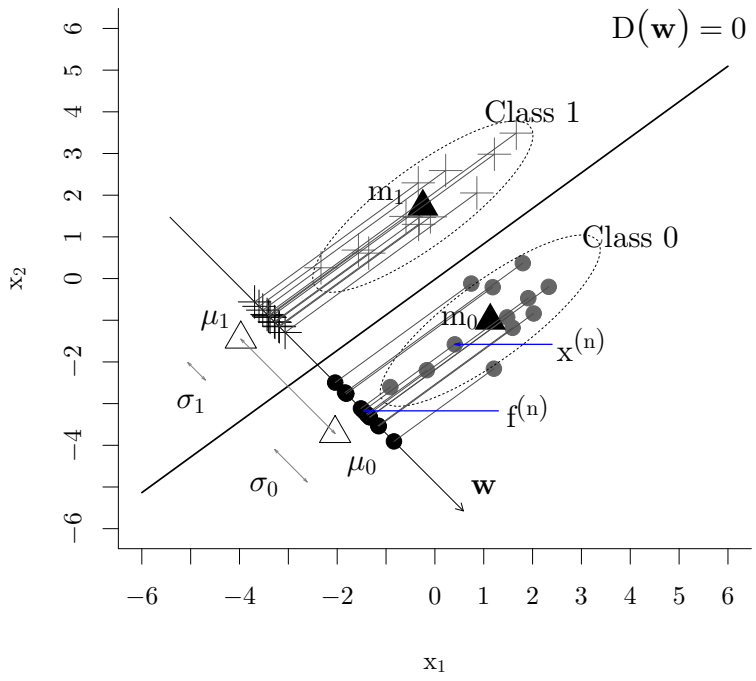


Figura 1: Ejemplo de análisis de discriminantes de Fisher. Dos agrupamientos de datos son proyectados en dirección de discriminación $\mathbf{w}$. Los miembros de cada clase se representan como signo 'más' o 'puntos'. La calidad del discriminante depende de la separación entre las medias proyectadas μ_0 y μ_1y la dispersión de las clases proyectadas σ_0^2 y σ_1^2. En la gráfica, $f^{(n)}$se refiere a la proyección de $\mathbf{x}^{(n)}$ sobre $\mathbf{w}$.

de la inclinación del vector $\mathbf{w}$. Obsérvese que se puede obtener un clasificador mediante la verificación de los signos de los puntos proyectados con respecto a $D(\mathbf{w})$, es decir, asignar cada elemento a la clase 1 si $D(\mathbf{w}) \geq 0$ y a la clase 0 en el caso complementario. Probabilidades a posteriori para cada clase, i.e. $p(\mathcal{C}_1|\mathbf{x})$ y $p(\mathcal{C}_0|\mathbf{x}) = 1 - p(\mathcal{C}_1|\mathbf{x})$, también se pueden derivar asumiendo que las proyecciones resulten de densidades Gaussianas.

Bajo este escenario, Fisher (1936) fue el primero que concluyó que el vector $\mathbf{w}$ resulta de la maximización de la razón de la varianza entre clases y la varianza dentro de clases,

$$J = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2}. \tag{1}$$

Donde utilizamos el subíndice $q = \{0, 1\}$ para denotar elementos que pertenecen a una de las clases. Dado que

$$\mu_q = \sum_{n \in q} \frac{1}{N_q} \mathbf{w}^T \mathbf{x}_q^{(n)}$$

y

$$\sigma_q^2 = \sum_{n \in q} \frac{1}{N_q} \left(\mathbf{w}^T \mathbf{x}_q^{(n)} - \mu_q \right)^2,$$

el coeficiente J se puede expresar en términos de $\mathbf{w}$ y con una manipulación directa llegamos a

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{\Sigma}_B \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma}_w \mathbf{w}}, \tag{2}$$

donde las matrices

$$\mathbf{\Sigma}_B = (\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T$$

y

$$\mathbf{\Sigma}_w = \sum_{q \in \{0,1\}} \sum_{n=1}^{N_q} \left(\mathbf{x}_q^{(n)} - \mathbf{m}_q \right) \left(\mathbf{x}_q^{(n)} - \mathbf{m}_q \right)^T$$

son conocidas como matrices de covarianza *entre* y *dentro* de clases, respectivamente. Una solución para el problema de discriminantes consiste en tomar la derivada de la Ecuación 2 con respecto a $\mathbf{w}$ y resolverla. Igualando a cero el gradiente y a través de algunas reordenaciones obtenemos

$$\hat{\mathbf{w}} \quad \propto \quad \mathbf{\Sigma}_w^{-1} (\mathbf{m}_0 - \mathbf{m}_1), \tag{3}$$

que es la expresión que buscamos.

Por lo tanto, las predicciones de clases para nuevas observaciones $\mathbf{x}^\star$ están fácilmente disponibles a través de la proyección de los datos sobre la dirección estimada del discriminante $\hat{\mathbf{w}}$ y la verificación del signo de la proyección, es decir,

$$f^\star = \hat{\mathbf{w}}^T \mathbf{x}^\star + b \geq D(\hat{\mathbf{w}}); \tag{4}$$

donde b es la ordenada al origen. Nótese que FDA no produce una estimación directa de probabilidades de las clases y en este sentido es un método no probabilístico.

4. Modelos discriminativos para la clasificación

En la presente sección nos enfocamos en métodos probabilísticos para clasificación, o sea, queremos que las predicciones de datos directamente tomen la forma de probabilidades de clases y no de valores que requieran una etapa de post-procesamiento para interpretarlos, así como en el caso de FDA. Primero, observamos que los problemas de clasificación pueden ser atendidos en términos similares que los de regresiones estándar, es decir, mediante la especificación explícita de una función de verosimilitud (o función de costos) que modela el proceso de generación de datos de las observaciones procediendo con la estimación de parámetros a través de la aplicación de técnicas tales como máxima verosimilitud. En esta sección introducimos la regresión logística, la cual probablemente es uno de los métodos más populares para clasificación.

4.1. Regresión logística

Volviendo al problema de asignación de la Sección 2, todavía queremos realizar una asignación de clases para la observación O y el enfoque más evidente es considerar $\mathbf{x}$ y y como variables aleatorias y trabajar con la densidad conjunta $p(\mathbf{x}, y)$ que de ellas se produce.[3] Aplicando las reglas de probabilidad, la probabilidad conjunta se puede denotar como $p(\mathbf{x}|\, y)\, p(y)$ y como $p(y|\, \mathbf{x})\, p(\mathbf{x})$ y de estas representaciones resultan los dos enfoques diferentes para la clasificación probabilística de datos. El primer enfoque usualmente se conoce como *generativo* porque modela el proceso de generación de datos en términos de la densidad condicional de clase $p(\mathbf{x}|\, y)$ que combinada con la distribución de probabilidad a priori (en lo subsecuente también denominada *previo*) de la clase $p(y)$ permite obtener la distribución de probabilidad a posteriori (en lo subsecuente denominada *posterior*):

$$p(y|\, \mathbf{x}) \quad = \quad \frac{p(\mathbf{x}|\, y)\, p(y)}{p(\mathbf{x}|\, y=1)\, p(y=1) + p(\mathbf{x}|\, y=0)\, p(y=0)} \ .$$

El segundo enfoque se llama *discriminativo* porque se enfoca en modelar $p(y|\, \mathbf{x})$ directamente y es en el que nos concentraremos en el presente documento. En ambos enfoques, el generativo y el discriminativo, es necesario asumir supuestos para la modelación, por ejemplo, decidir qué tipo de densidad se utiliza para representar $p(\mathbf{x}|\, y)$ o $p(y|\, \mathbf{x})$.

Una manera directa de obtener un clasificador discriminativo es la transformación de una función de regresión en la probabilidad de la cual se espera la clase, por ejemplo, aplicando una función de respuesta.[4] Es decir, considerando una función de regresión $f(\cdot)$ cuyo domino es $(-\infty, \infty)$, adaptándola al rango $[0, 1]$, se habrá obtenido el clasificador deseado. Un ejemplo es el modelo de regresión logística

$$p(y = 1|\, \mathbf{x}) \quad = \quad g\left(\mathbf{w}^T \phi(\mathbf{x})\right), \tag{5}$$

[3] Recordamos que $\mathbf{x}$ es un vector de características observadas obtenido de manera indirecta, mientras que y es una variable canonica que representa la clase.

[4] La función de respuesta es la inversa de la función de *enlace* utilizada en estadística.

cuya función de respuesta es

$$g(z) = \frac{1}{1+\exp(-z)}. \tag{6}$$

Nótese que (5) es una combinación de un modelo lineal, parametrizado por $\mathbf{w}$, una función base $\phi(\cdot)$ y la función de respuesta logística g. Una función alternativa es la función acumulativa Gaussiana $g(z) = \int_{-\infty}^{\infty} \mathcal{N}(x|\,0,1)\,dx$ que produce el modelo probit.

Dado el conjunto de entrenamiento $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, con $y^{(n)} \in \{1, 0\}$, podemos utilizar la definición del problema de la Sección 3 para interpretar cómo funciona la regresión logística. Podemos volver a pensar que el objetivo es encontrar un vector de pesos, tal que las proyecciones de datos sobre el vector estarán separadas al máximo de acuerdo con un criterio especificado. Sin embargo, el criterio ya no será el cociente de (1) sino la función de verosimilitud (5) y por lo tanto, surgirá un nuevo problema de optimización,

$$\begin{aligned} \hat{\mathbf{w}} &= -\arg\min_{\mathbf{w}} \ln p(\mathbf{y}|\,\mathbf{X}, \mathbf{w}) \\ &= -\arg\min_{\mathbf{w}} \sum_{n=1}^{N} \left\{ y^n \ln \sigma(a_n) + (1-y^n) \ln(1-\sigma(a_n)) \right\}, \end{aligned} \tag{7}$$

donde $a_n = \mathbf{w}^T \mathbf{x}^{(n)}$.

El estimado $\hat{\mathbf{w}}$ se puede obtener a través de métodos numéricos (Bishop, 2006). En contraste con FDA, las predicciones están disponibles cuando se incorpora la estimación de $\hat{\mathbf{w}}$ y el punto de prueba $\mathbf{x}^\star$ en la función logística (6) y esta vez se producirá automáticamente una probabilidad de pertenencia a una clase. Suponiendo que la base $\phi(\cdot)$ es la identidad, la probabilidad se vuelve $p(y^\star = 1|\,\mathbf{x}^\star) = g\left(\hat{\mathbf{w}}^T \mathbf{x}^\star\right)$.

5. Procesos Gaussianos para regresión y clasificación

Los procesos Gaussianos son una generalización de densidades multivariadas Gaussianas a conjuntos de funciones infinitas continuas (Rasmussen, 2004) y se han utilizado para tareas de inferencia de datos por al menos durante los últimos cien años; por ejemplo, Thiele (1903) fue uno de sus primeros proponentes. Sin embargo, las aplicaciones modernas de GP's iniciaron con el trabajo del ingeniero de minas, Krige (1966), y después con el de Kimeldorf y Wahba (1970), O'Hagan (1978) y Wahba (1990). Se utiliza el término *proceso* para referirse a una colección de variables aleatorias indexadas $\left[f^{(1)}, \ldots, f^{(N)}\right]$ que (i) se pueden definir a través de una densidad de probabilidad común, en este caso una Gaussiana, y (ii) que satisface las propiedades de consistencia y permutación; véase Grimmett y Stirzaker (2004) para una definición de un proceso consistente.

Los procesos Gaussianos mantienen estrechas conexiones con ANN's siempre y cuando ambos se estudien de una perspectiva Bayesiana (Neal, 1996). Sin embargo, en contraste con las ANN's, los procesos Gaussianos conllevan la ventaja de modelar flexiblemente y sin la limitación de tener que adaptar un gran número de parámetros, algo que comúnmente restringió la aplicación de ANN's en muchos problemas. Un trabajo de finanzas computacionales que específicamente enfoca la predicción de bancarrota en términos de ANN's es el de Atiya (2001).

En la presente sección, discutimos la regresión lineal y su enfoque complementario, regresión funcional o con GP's, ambas desde la perspectiva Bayesiana. De hecho, se puede mostrar que ambos enfoques son equivalentes, pero bajo ciertas circunstancias es más conveniente aplicar uno en lugar de otro. La siguiente discusión permitirá la introducción de algunas formas diferentes de GP's para la clasificación de datos: clasificadores con procesos Gaussianos, máquinas de soporte vectorial de mínimos cuadrados, entre otras. En documentos como por ejemplo Box y Tiao (1973) se discute el enfoque Bayesiano para la regresión lineal, mientras que la regresión con GP's en los más modernos, por ejemplo, Mackay (2003) y Rasmussen y Williams (2006).

5.1. Regresión lineal Bayesiana: el enfoque en el espacio de parámetros

La siguiente discusión se basa principalmente en Williams (1999). Consideramos lo que se puede llamar regresión lineal generalizada porque utilizaremos un conjunto fijo de funciones base $\{\phi_i(\mathbf{x})\}_{i=1}^{m}$. Suponiendo, además, un conjunto de datos de entrenamiento $\mathcal{D} = \left\{\left(\mathbf{x}^{(n)}, t^{(n)}\right)\right\}_{n=1}^{N}$ y una función latente f que nos interesa inferir y que las entradas y objetivos están relacionados de manera lineal a través de $t^{(n)} = f^{(n)} + \epsilon$; con $f^{(n)} = \mathbf{w}^T\phi\left(\mathbf{x}^{(n)}\right)$ y $\epsilon \sim \mathcal{N}\left(0, \sigma_v^2\right)$. Entonces, una representación de la información extraída de los datos está dada por la distribución a posteriori sobre los parámetros $\mathbf{w}$, que se expresa en términos de regla de Bayes de la siguiente manera:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\,p(\mathbf{w})}{p(\mathcal{D})}; \tag{8}$$

donde $p(\mathcal{D}|\mathbf{w})$ se conoce como una función de verosimilitud y $p(\mathbf{w})$ como la distribución a priori. Si las observaciones son independientes e idénticamente distribuidas[5], la verosimilitud se puede representar muy bien por $t^{(n)} \sim \mathcal{N}\left(\mathbf{w}^T\phi\left(\mathbf{x}^{(n)}\right), \sigma_v^2\right)$, mientras que al previo como $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{wt})$. Bajo dichos supuestos es muy fácil mostrar que el posterior tomará la siguiente forma:

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}_{MAP}, \mathbf{A}_r),$$

donde el vector media

$$\mathbf{w}_{MAP} = \beta\mathbf{A}_r^{-1}\mathbf{\Phi}^T\mathbf{t}, \tag{9}$$

[5]Que es uno de los principios en este trabajo.

y la matriz de covarianza $\mathbf{A}_r = \mathbf{\Sigma}_{wt}^{-1} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$, con $\mathbf{\Phi}$ siendo la matriz de diseño y $\beta = 1/\sigma_v^2$, la precisión. Notamos que la covarianza posterior $\mathbf{\Sigma}_{wt}$ es una matriz de dimensiones $m \times m$.

Desde una perspectiva de modelación de datos, el objetivo principal no es derivar la distribución a posteriori, más bien predecir $f^\star$ para datos no observados $\mathbf{x}^\star$, que en el presente caso se realiza mediante la evaluación de

$$\begin{aligned} p\left(f^\star | \mathcal{D}\right) &= \int p\left(f^\star | \mathcal{D}, \mathbf{w}\right) p\left(\mathbf{w} | \mathcal{D}\right) \partial \mathbf{w} \\ &= \mathcal{N}\left(\bar{f}^\star, \left(\sigma^\star\right)^2\right). \end{aligned} \tag{10}$$

Nótese que la integral anterior es el promedio ponderado de expectativas condicionales sobre la distribución posterior[6]. La media y covarianza se denotan por las siguientes expresiones:

$$\bar{f}^\star = \mathbf{w}_{MAP}^T \phi\left(\mathbf{x}^\star\right), \tag{11}$$

y

$$\sigma_f^2\left(\mathbf{x}^\star\right) = \phi\left(\mathbf{x}^\star\right)^T \mathbf{A}_r^{-1} \ \phi\left(\mathbf{x}^\star\right), \tag{12}$$

respectivamente. Con respecto al resultado de la media, si consideramos la definición de la clasificación, es fácil mostrar que $\mathbf{w}_{MAP}$ (9) es equivalente a $\hat{\mathbf{w}}_{FDA}$ (3) si simplemente se fijan los objetivos a los valores de las etiquetas (Bishop, 1995). Cabe señalar que es necesario agregar σ_v^2 a $\sigma_f^2\left(\mathbf{x}^\star\right)$ para obtener la varianza predictiva $\operatorname{var} t\left(\mathbf{x}^\star\right)$ y para considerar la varianza adicional debido al ruido, ya que las dos fuentes de variación no están correlacionadas (Williams, 1999).

5.2. Procesos Gaussianos para regresión: el enfoque en el espacio funcional

En la sección previa observamos cómo la incertidumbre en un problema típico de regresión se describió en términos de una distribución de probabilidad sobre los parámetros $\mathbf{w}$. También es posible manejar directamente la incertidumbre con respecto a los valores de la función en los puntos que nos interesan, lo que representa la perspectiva del espacio funcional (o GP) del problema, como expresa Williams (1999). El punto clave de por qué se dejó de usar el enfoque basado en parámetros para modelar datos es que las proyecciones también se pueden manejar como variables aleatorias. Específicamente, asumiendo una generación finita de instancias $\mathbf{f} = \left[f^{(1)}, \ldots, f^{(N)}\right]^T$ definidas de manera consistente obtendremos un proceso aleatorio, que será un GP, si $\mathbf{f}$ se describe por una densidad (Mackay, 1998).

En particular, asumimos que cada $f^{(n)}$ depende de una entrada $\mathbf{x}^{(n)}$ con índice n, así que $f^{(n)} = f\left(\mathbf{x}^{(n)}\right)$. Nótese que dicha definición implica que la parametrización de f con $\mathbf{w}$ es irrelevante para el proceso de modelado. Sin embargo, la justificación del supuesto GP es corroborada por el hecho de que

[6] Omitamos dependencias de $\mathbf{x}^\star$ para mantener la notación simplificada.

poniendo un previo Gaussiano sobre los párametros $\mathbf{w}$ induce un distribución Gaussiana previa sobre el conjunto de instancias $\mathbf{f}$, dado que $\mathbf{f}$ es una función lineal de $\mathbf{w}$.

Por lo tanto, asumiendo datos de entrenamiento $\mathcal{D}$ observados, se requerirá inferir una distribución a posteriori en términos similares a aquellos presentados en la Sección 5.1. Con respecto a la especificación de una distribución a priori del tipo GP, se definirá por una función de la media $m(\mathbf{x})$ y de la covarianza $k(\mathbf{x}, \mathbf{x}')$, es decir, $p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ con matriz $\mathbf{K} \in \mathbb{R}^{N \times N}$ poblada con entradas de la forma $k\left(\mathbf{x}^i, \mathbf{x}^j\right) \forall i, j$. Si la verosimilitud $p(\mathcal{D}|\mathbf{f})$ es Gaussiana, o sea, si $\mathcal{D}$ se compone por un conjunto de observaciones con ruido $t^{(n)} = f^{(n)} + \epsilon$, con $\epsilon \sim \mathcal{N}\left(0, \sigma_v^2\right)$, se puede mostrar que la aplicación de la regla de Bayes resultará en:

$$\begin{aligned} p(\mathbf{f}|\mathcal{D}) &\propto p(\mathcal{D}|\mathbf{f})\, p(\mathbf{f}) \\ &= \mathcal{N}\left(\mathbf{K}\left(\sigma_v^2\mathbf{I} + \mathbf{K}\right)^{-1}\mathbf{t}, \sigma_v^2\left(\sigma_v^2\mathbf{I} + \mathbf{K}\right)^{-1}\mathbf{K}\right), \end{aligned} \tag{13}$$

donde el vector $\mathbf{f} = \left[f\left(\mathbf{x}^{(1)}\right), \ldots, f\left(\mathbf{x}^{(N)}\right)\right]^T$ y $\mathbf{t} = \left[t^{(1)}, \ldots, t^{(N)}\right]^T$, (Seeger, 2004). Por lo tanto, la distribución posterior está influida por la distribución previa y eso se detecta en (13) a través de la observación que la media y covarianza posterior dependen de la matriz $\mathbf{K}$, que es la covarianza previa.

Anteriormente, se ha inferido la distribución posterior sobre los datos de entrenamiento, i.e. $p(\mathbf{f}|\mathcal{D})$ pero la tarea más importante es predecir puntos de prueba, lo que requiere determinar la distribución predictiva posterior para un punto $f^\star = f(\mathbf{x}^\star)$, fuera del conjunto de entrenamiento, una vez que se ha observado D.

Lo anterior se realiza fácilmente aplicando

$$\begin{aligned} p(f^\star|\mathcal{D}) &= \int p(f^\star|\mathbf{f})\, p(\mathbf{f}|\mathcal{D})\, \partial\mathbf{f}, \\ &= \mathcal{N}\Big(\mathbf{k}(\mathbf{x}^\star)^T\left(\mathbf{K} + \sigma_v^2\mathbf{I}\right)^{-1}\mathbf{t}, k(\mathbf{x}^\star, \mathbf{x}^\star) + \mathbf{k}(\mathbf{x}^\star)^T \\ &\qquad \left(\mathbf{K} + \sigma_v^2\mathbf{I}\right)^{-1}\mathbf{k}(\mathbf{x}^\star)\Big) \end{aligned} \tag{14}$$

donde el vector $\mathbf{k}(\mathbf{x}^\star) \in \mathbb{R}^{N \times 1}$ tiene como elementos escalares de la forma $k\left(\mathbf{x}^{(n)}, \mathbf{x}^\star\right)$, para $n = 1, \ldots, N$. Remitimos al lector interesado a Williams (1999) para la demostración de la equivalencia de los resultados (10) y (14).

Dado que las perspectivas del espacio de pesos y del espacio funcional para hacer una regresión son equivalentes, vale preguntar cuál es más conveniente por aplicar. Desde una perspectiva computacional, ambos enfoques se basan en la inversión de una matriz, la cual en el enfoque en el espacio de pesos es la de $\mathbf{\Sigma}_{wt}$, una matriz de $m \times m$ (Sección 5.1); mientras que en el enfoque del espacio funcional es $\mathbf{K}$, una matriz $N \times N$. En general, para muchos tipos de regresión, $m \ll N$ y el enfoque en el espacio de pesos sería preferible. Sin embargo, para ciertos tipos de predicciones lineales, m es infinito y el único enfoque posible se vuelve el de perspectiva del espacio funcional.

Considerando la complejidad computacional de invertir una matriz $\mathbf{M} \in \mathbb{R}^{\ell \times \ell}$ es $\mathcal{O}\left(\ell^3\right)$, donde ℓ es el número de filas (Golub y Van Loan, 1996). Williams

(1999) y Schölkopf y Smola (2002) presentan aspectos adicionales acerca de conveniencia del enfoque en el espacio funcional para la regresión.

5.2.1. La función de covarianza

La mayoría de las aplicaciones de GP's asumen que la función de la media $m(\mathbf{x})$ se centra alrededor de $\mathbf{0}$, así que el punto clave del formalismo se encuentra en el tipo de función de covarianza utilizada. Por lo tanto, cabe analizar algunas características de dicha función, por ejemplo, en este trabajo solamente utilizamos funciones *isotrópicas* de la forma $k(\mathbf{x}, \mathbf{x}') = k(r)$, con $r = \|\mathbf{x} - \mathbf{x}'\|$. En covarianzas isotrópicas, la correlación entre las observaciones es independiente de su posición absoluta; solamente cuenta su diferencia en términos de una norma. Por ejemplo, utilizando una norma Euclidiana aseguramos que los puntos $\mathbf{x}$ y $\mathbf{x}'$, que se encuentran cerca el uno de otro, se reflejarán en una correlación alta. Por lo tanto, $f(\mathbf{x})$ y $f(\mathbf{x}')$ también estarán cerca de uno a otro. Un ejemplo de una covarianza isotrópica que utilizamos es

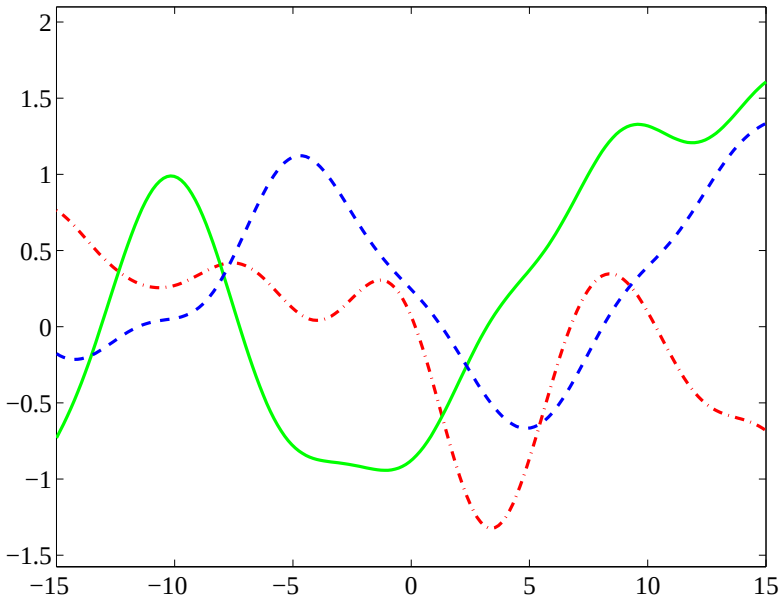
$$k\left(f\left(\mathbf{x}^i\right), f\left(\mathbf{x}^j\right)\right) \quad = \quad k\left(\mathbf{x}^i, \mathbf{x}^j\right) \quad = \quad \theta_1 \exp\left(-\frac{\theta_2}{2}\left|\left|\mathbf{x}^i - \mathbf{x}^j\right|\right|^2\right) \tag{15}$$

también conocida como covarianza de función de base radial o RBF, por sus siglas en inglés. Los parámetros $\mathbf{\Theta}_k = \{\theta_1, \theta_2\}$ ajustan la longitud escalar de la función radial, que en este caso es Gausssiana. La inversa de θ_2 también es conocida como el parámetro de ancho de banda σ.

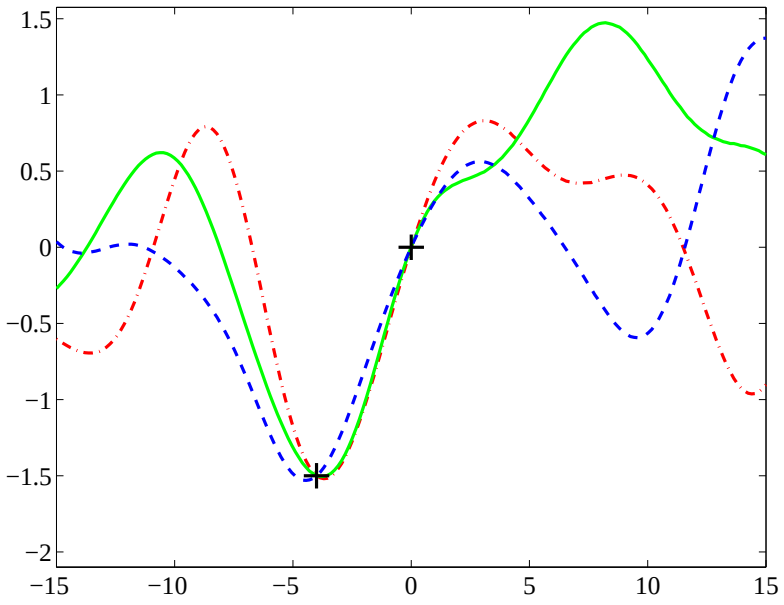
Para comparar cómo influye la elección de la covarianza a los GP's previo y posterior, la Gráfica 2 presenta muestras de ambas, donde la primera se define como $p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ y la última como $p(f^\star|\mathcal{D})$, especificada en la Ecuación 14. Se utilizó una covarianza RBF (15) para tomar las muestras. En la Gráfica (a) las funciones pueden tener cualquier forma, dado que son suaves, mientras que en la Gráfica (b), las funciones también tienen que ser suaves pero son atraídas por los puntos observados. En ambos casos, el ancho de banda de la RBF fue ajustado a $\log\theta_2 = \log\sigma^{-1} = -2{,}3026$.

5.3. Procesos Gaussianos para clasificación

Podemos pensar en la regresión GP como la generalización de la más conocida regresión lineal Bayesiana y en términos similares se puede entender la clasificación GP como la generalización de la regresión logística. Nótese que en la Sección 4 la función de activación logística se denotó por $a = \mathbf{w}^T\phi(\mathbf{x})$, y así, siguiendo una lógica similar a la de la sección previa, un proceso Gaussiano permite no linealizar la función a para trabajar directamente sobre el espacio funcional. Por lo tanto, considerando una colección de variables latentes a_n para $n \in \{1, N\}$, podemos reemplazar los modelos lineales $\mathbf{w}^T\phi\left(\mathbf{x}^{(n)}\right)$ por un proceso Gaussiano $\mathbf{f}$. Además, dada la observación $\mathbf{x}^\star$ nos interesa determinar la probabilidad de pertenencia a una clase $\pi(\mathbf{x}^\star) = p(y = 1|\mathbf{x}^\star) = \sigma(f(\mathbf{x}^\star))$. El



(a) Previo



(b) Posterior

Figura 2: La gráfica ilustra 3 muestras de GP's a priori y a posteriori. (a) Muestras del previo $p(\mathbf{f}) = N(\mathbf{0}, \mathbf{K})$. (b) Dados unos datos de entrenamiento D, la gráfica contiene muestras obtenidas del posterior predictivo $p(f^{\star} | \mathcal{D})$ de la Ecuación 14. En ambas gráficas se utilizó una covarianza RBF (15) para computar la matriz K. Nótese que en (b) las funciones continúan siendo suaves, pero en este caso están condicionadas por los puntos observados.

proceso de inferencia se realiza análogamente a lo descrito previamente y por lo tanto, la distribución sobre $f^\star$ se calcula como

$$p\left(f^\star|\,\mathcal{D}\right) \;=\; \int p\left(f^\star|\,\mathcal{D},\mathbf{f}\right)p\left(\mathbf{f}|\,\mathcal{D}\right)\partial\mathbf{f}\;, \tag{16}$$

donde $p\left(\mathbf{f}|\,\mathcal{D}\right) \propto p\left(\mathcal{D}|\,\mathbf{f}\right)p(\mathbf{f})$ es el posterior obtenido a través de la aplicación de la regla de Bayes. Sin embargo, en contraste con el caso de la regresión de la Sección 5.2, el modelo de ruido que requiere ser especificado es el de clasificación, es decir, una distribución Bernoulli de la forma

$$p\left(\mathcal{D}|\,\mathbf{f}\right) \;=\; \prod_{n=1}^{N}\sigma\left(f^n\right)^{y^n}\left[1-\sigma\left(f^n\right)^{(1-y^n)}\right]\;. \tag{17}$$

Esta densidad es equivalente a aquella presentada como argumento en el problema de optimización de la Ecuación 7, pero con la parametrización de la forma $\mathbf{w}^T\phi\left(\mathbf{x}\right)$ reemplazada por f.

El posterior (16) se utiliza subsecuentemente para estimar una predicción probabilística de la etiqueta de la clase, que es

$$\pi^\star \;=\; p\left(y^\star=1|\,\mathcal{D},\mathbf{x}^\star\right) \;=\; \int p\left(y^\star|\,f^\star\right)p\left(f^\star|\,\mathcal{D}\right)\partial f^\star. \tag{18}$$

Ambas integrales (16) y (18) no son analíticamente computables y así se tienen que calcular por aproximación. Sin embargo, mientras que la Ecuación 16 usualmente se computa a través de métodos estocásticos, tales como Markov Chain Monte Carlo o enfoques determinísticos como la aproximación de Laplace o inferencia variacional; la Ecuación 18 que es de una dimensión se puede evaluar mediante técnicas numéricas estándar como cuadratura. Más referencias de procesos Gaussianos para clasificación, o GPC's, se encuentran en Williams y Barber (1998).

5.4. Algunos otros tipos de GP's

El método de aproximación más directo es tal vez la realización de una expansión cuadrática alrededor del modo de la distribución posterior $p\left(\mathbf{f}|\,\mathcal{D}\right)$ lo que comúnmente se denota como la aproximación de Laplace. Sin embargo, varios autores (por ejemplo, Minka, 2001) han comprobado que dichos tipos de aproximación muchas veces no logran capturar la verdadera naturaleza de la distribución y producen malos resultados de predicción. En la literatura existen varios métodos alternativos, entre ellos, la aproximación de los modos de la verosimilitud de Bernoulli $p\left(\mathcal{D}|\,\mathbf{f}\right)$ con densidades Gaussianas. Este método genera un clasificador con características similares a aquellas de FDA y, como mostraron Peña Centeno and Lawrence (2006), que puede producir resultados competitivos en algunos problemas. En lo siguiente referiremos a dicho método como Discriminante Bayesiano de Fisher (BFD, por sus siglas en inglés).

Otro tipo de técnica de GP es la máquina de soporte vectorial de mínimos cuadrados de Suykens y Vandewalle (1999), la cual se formula como problema

de optimización con restricción de igualdad. La motivación de este método, llamado LS-SVM, por sus siglas en inglés, es encontrar una manera más rápida y más sencilla de resolver el problema de Programación Cuadrática (QP, por sus siglas en inglés) que incluye resolver máquinas de soporte vectorial estándar (Cortes y Vapnik, 1995). La simplificación consiste en reemplazar la restricción de la desigualdad de las máquinas de soporte vectorial estándar con la de una igualdad. De esta manera el LS-SVM se puede resolver de una forma menos intensiva en términos de poder de cómpúto, a costo de perder esparcidad en la solución.

Finalmente, una de las principales desventajas de aplicar la regresión GP resulta del hecho de que asume ruido Gaussiano e infortunadamente la mayoría de los problemas no muestran dicha característica. Snelson et al. (2003) generalizan el marco de GP para la regresión mediante el aprendizaje de una transformación no lineal de los resultados, de tal forma que ruido no Gaussiano todavía se puede modelar con un GP. Snelson et al. mencionaron que la generalización consiste en el aprendizaje de *un regresor de GP en espacio latente y simultáneamente una transformación o deformación del espacio para las salidas*; de esta manera se toman en cuenta otros tipos de ruido. Dicha estrategia se denominará Proceso Gaussiano Deformado o WGP, por sus siglas en inglés.

Estos tres métodos (BFD, LS-SVM y WGP) solamente son un conjunto de herramientas algorítmicas que ha desarrollado la comunidad de machine learning para resolver problemas de regresión y clasificación. En la Sección 6 recurriremos a estas herramientas y probaremos su efectividad con respecto al problema de clasificar un conjunto de datos reales.

5.5. Adaptación de hiperparámetros

En todos los métodos presentados basados en GP, solamente después de obtener una solución para la distribución predictiva posterior $p\left(f^{\star}|\,\mathcal{D}\right)$ se puede abordar el asunto de fijar los hiperparámetros $\boldsymbol{\Theta}_k$ de la función de covarianza. La metodología Bayesiana prescribe que dichos parámetros deberían inferirse de manera jerárquica. Sin embargo, las distribuciones de parámetros condicionales resultantes de una covarianza del tipo especificado en (15) no se prestan para el muestreo de Gibbs. Por lo tanto, expertos han buscado métodos más directos para la estimación de parámetros. Williams (1999), por ejemplo, recomienda el uso de máxima verosimilitud o validación cruzada generalizada. En esta sección se proporcionan más detalles sobre la estimación de máxima verosimilitud, mientras que la aplicación de la validación cruzada generalizada se encuentra en Rasmussen y Williams (2006). En el presente trabajo, seleccionamos hiperparámetros para todos los algoritmos basados en GP's a través de la máxima verosimilitud.

En el ejemplo más sencillo, el caso de regresión, dados algunos datos de entrenamiento $\mathcal{D} = (\mathbf{X}, \mathbf{t})$, un modelo con ruido de forma $p\left(\mathcal{D}|\,\mathbf{f}\right) = \mathcal{N}\left(\mathbf{f}, \sigma_v^2\mathbf{I}\right)$ y un previo tipo GP de la forma $p\left(\mathbf{f}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$, se puede comprobar que la verosimilitud marginal es

$$
\begin{aligned}
p\left(\mathcal{D}|\,\boldsymbol{\Theta}_k\right) &= \int p\left(\mathcal{D}|\,\mathbf{f}\right) p\left(\mathbf{f}|\,\boldsymbol{\Theta}_k\right) \partial \mathbf{f} \\
&= \frac{1}{(2\pi)^{N/2}\left|\mathbf{K}+\sigma_v^2\mathbf{I}\right|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{t}^T\left(\mathbf{K}+\sigma_v^2\mathbf{I}\right)^{-1}\mathbf{t}\right\}.
\end{aligned}
$$

Por lo tanto, el log de $p\left(\mathcal{D}|\,\boldsymbol{\Theta}_k\right)$ se puede computar analíticamente como

$$
\begin{aligned}
l &= \log p\left(\mathcal{D}|\,\boldsymbol{\Theta}_k\right) \\
&= -\frac{1}{2}\log\left|\mathbf{K}+\sigma_v^2\mathbf{I}\right| - \frac{1}{2}\mathbf{t}^T\left(\mathbf{K}+\sigma_v^2\mathbf{I}\right)^{-1}\mathbf{t} - \frac{N}{2}\log 2\pi.
\end{aligned} \tag{19}
$$

Dado que no existe una solución de forma cerrada para la maximización de l con respecto a $\boldsymbol{\Theta}_k$, se tiene que recurrir a métodos numéricos, tales como gradientes conjugados para encontrar el máximo local. En efecto, el gradiente de (19) se utilizará y se expresa como

$$
\frac{\partial l}{\partial \theta_i} = -\frac{1}{2}\mathbf{t}^T\left(\mathbf{K}+\sigma_v^2\mathbf{I}\right)^{-1}\mathbf{t} + \mathbf{t}^T\left(\mathbf{K}+\sigma_v^2\mathbf{I}\right)^{-1}\frac{\partial \mathbf{K}}{\partial \theta_i}\left(\mathbf{K}+\sigma_v^2\mathbf{I}\right)^{-1}\mathbf{t}.
$$

La estrategia para la especificación de parámetros en el caso de clasificadores y otras variantes de procesos Gaussianos (es decir, GPC, BFD y WGP) sigue la misma estructura a la de regresión, o sea, la idea consiste en maximizar la verosimilitud de los parámetros, pero con el modelo específico de ruido definido para cada método. Por ejemplo, en el caso de GPC's será la Ecuación 17.

5.6. Determinación automática de relevancia

La adaptación de valores de hiperparámetros es importante si se quiere tener buenos resultados de generalización y un mejor entendimiento de los datos. En efecto, para algunas familias de funciones de covarianza existe un hiperparámetro asociado con cada dimensión de entrada, tal que cada uno representa la longitud escalar característica de los datos [7]. Por lo tanto, aplicando un método de adaptación de parámetros como el de máxima verosimilitud, se inferirá la importancia relativa de las entradas.

Por ejemplo

$$
k\left(\mathbf{x}^i,\mathbf{x}^j\right) = \theta_1 \exp\left(-\frac{\theta_2}{2}\left(\mathbf{x}^i-\mathbf{x}^j\right)^T\boldsymbol{\Theta}_{ard}\left(\mathbf{x}^i-\mathbf{x}^j\right)\right) + \theta_3\delta_{ij}, \tag{20}
$$

es una función que pondera cada componente de $\boldsymbol{\Theta}_{ard} = \mathrm{diag}\left(\theta_4,\ldots,\theta_{4+d-1}\right)$ -con d como la dimensión de los datos cuando se realiza el entrenamiento. El

[7] Como expresaron Rasmussen y Williams (2006), la longitud escalar característica se puede interpretar como la distancia requerida para moverse a lo largo de cada eje para obtener entradas no correlacionadas.

parámetro δ_{ij} es la delta de Kronecker que, para un valor θ_3 suficientemente grande, asegura que $\mathbf{K}$ sea positiva definida y por lo tanto, invertible en todo momento.

Este tipo de característica fue propuesta inicialmente por Mackay (1995) y Neal (1996) en el contexto de redes neuronales y usualmente se denomina como *determinación automática de relevancia* o ARD, por sus siglas en inglés. Si la selección de covarianza de la distribución a priori es adecuada, entonces ARD puede ser un método muy útil para clasificar y seleccionar las características dado que ordena efectivamente las entradas de acuerdo a su importancia y elimina aquellas que no se consideran importantes. Esta característica puede ser muy útil para el problema de predicción de bancarrota porque se puede utilizar para clasificar las razones financieras según su importancia, como se realizará más adelante.

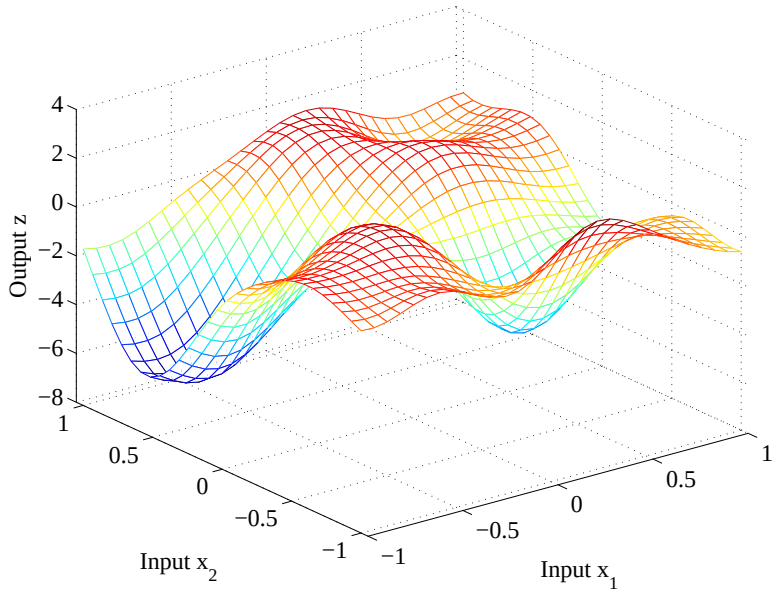
Para entender mejor a ARD, la Gráfica 3 ilustra muestras de la covarianza de la forma (20) con entradas bidimensionales. El panel (a) presenta una muestra donde ambas entradas x_1 y x_2 tienen los mismos pesos asociados θ_4 y θ_5. Por lo tanto, en promedio el conjunto de muestras tendrá aproximadamente el mismo grado de variación a lo largo de los ejes x_1 y x_2. Por el contrario, el panel (b) presenta una muestra donde el valor $\theta_4 > \theta_5$, produciendo un resultado que varía más en dirección x_1 que x_2. Por lo tanto, en ambos casos, observando ciertos datos $\mathcal{D}$, la distribución a posteriori ajustada tendrá pesos θ_4 y θ_5 que reflejan su importancia "real" para la regresión.

6. Datos y Experimentos

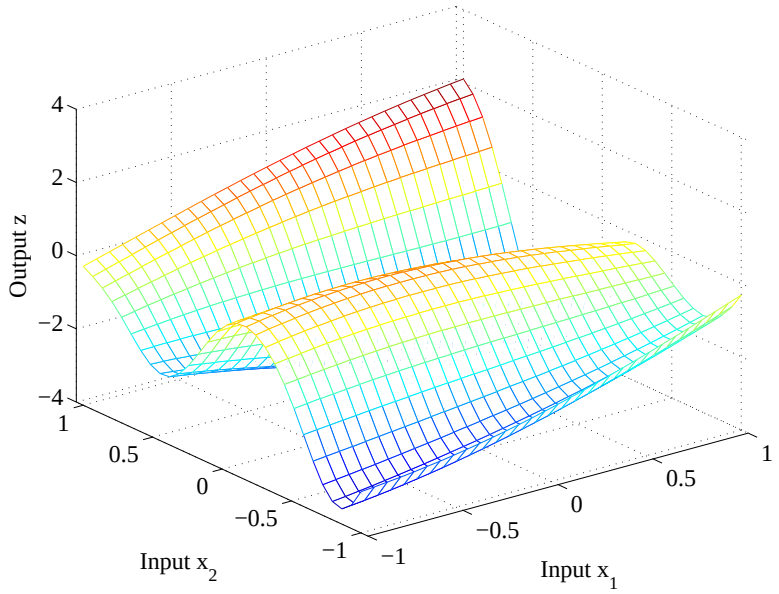
La presente sección describe los experimentos que se realizaron para comparar el desempeño predictivo de los algoritmos propuestos con respecto al análisis de discriminantes y a la regresión logística. Como se mencionó previamente, utilizamos datos del Seguro de Depósito de los Estados Unidos (FDIC, por sus siglas en inglés) y un análisis breve de los resultados sigue. Cabe destacar que se realizaron experimentos bajo un esquema limitado y consecuentemente los resultados no son conclusivos estadísticamente, pero sin embargo proporcionan cierta evidencia acerca del potencial poder de los GP's.

6.1. Datos de FDIC

La Universidad de Essex (Reino Unido) amablemente proporcionó un conjunto de datos que comprende 280 observaciones multivariadas y cuyas características principales son las siguientes. La muestra contiene un igual número de bancos estadounidenses que fueron y no fueron a la quiebra y que reportaron sus estados financieros al FDIC del 31 de marzo de 1989 al 30 de septiembre de 2002. Se formaron pares de cada uno de los bancos que fueron a la quiebra con un banco que no fue a la quiebra, de acuerdo a criterios de tamaño de activos, tipo de institución y ubicación; un procedimiento estándar para este tipo de trabajo (Tam y Kiang, 1992); y además, se excluyeron observaciones



(a) $\theta_4 = 1{,}0$ y $\theta_5 = 1{,}0$



(b) $\theta_4 = 1{,}0$ y $\theta_5 = 0{,}01$

Figura 3: Las funciones fueron obtenidas de un previo tipo GP de dos dimensiones con una función de covarianza ARD de la forma (20). El Panel (a) muestra una función con dos entradas x_1 y x_2 de igual importancia, mientras que en el panel (b), la entrada x_1 varia más rápido que x_2. ARD puede ayudar a determinar la relevancia de una característica (por ejemplo, razón financiera) en una tarea de clasificación.

Financial ratios	
1. Margen de interés neto	7. Razón de eficiencia
2. Ingreso -sin los incurridos por intereses- a activos productivos	8. Activos fijos/no corrientes más otros inmuebles en posesión de activos
3. Gastos sin interés a activos productivos	9. Razón de efectivo más obligaciones de gobierno y del tesoro de E.E.U.U.
4. Ingreso operativo neto a activos	10. Capital propio/de patrimonio neto a activos
5. Rentabilidad económica	11. Razón de capital primario
6. Rentabilidad financiera	

Cuadro 1: Las razones financieras utilizadas en el experimento de clasificación. Los datos se obtuvieron del Seguro de Depósito de los Estados Unidos (FDIC, por sus siglas en inglés) y fueron proporcionados por el Centro de Finanzas Computacionales y Agentes Económicos (*Centre for Computational Finance and Economic Agents, CCFEA*), Universidad de Essex. El Apéndice A describe cada razón.

cuando faltaron atributos. El tamaño de activos de la muestra fue entre $\$2m$ y $\$700m$ y las razones financieras utilizadas se obtuvieron en un periodo de cuatro trimestres antes de la quiebra de un banco dado. Después de excluir los datos incompletos, el conjunto final de datos comprendió 234 observaciones que se dividieron de manera aleatoria en un conjunto de entrenamiento que consistió en 145 observaciones y un conjunto de prueba con 89 observaciones. El número de características en cada observación es 11, el número de razones financieras a analizar. Cabe recordar que el enfoque utilizado en este trabajo considera cada observación como independiente e idénticamente distribuida, independientemente de que el periodo de observación cubre alrededor de trece años. Debido a la cantidad limitada de datos, tuvimos que confiar en particiones aleatorias de los datos de entrenamiento y de prueba con el fin de promediar nuestros resultados y reducir en todo lo posible cualquier efecto de varianza. Formamos 100 pares diferentes de conjuntos de entrenamiento y en los datos de prueba de las 234 observaciones disponibles, manteniendo la misma proporción entre los datos de entrenamiento y los datos de prueba como en el conjunto de datos original. Utilizando la partición aleatoria de los datos para reducir la varianza de estimaciones no es inusual y es justificado por el trabajo de Efron (1979) y Stone (1974) sobre *boot-strapping* y validación cruzada, respectivamente. Rätsch et al. (1998) utilizaron un enfoque similar al nuestro. Esta definición también ayuda a reducir los efectos de cualquier sobreajuste posible, aunque eso no está completamente garantizado.

6.2. Definición experimental

Probamos cinco algoritmos diferentes de los datos referidos: análisis de discriminantes de Fisher (FDA), máquinas de soporte vectorial de mínimos cuadrados

(LS-SVM, por sus siglas en inglés), GP's para clasificación (GPC), GP's deformados (WGP) y discriminante Bayesiano de Fisher (BFD). Cada conjunto fue normalizado con media cero y desviación estándar de uno. Dada la partición de los datos, los algoritmos se repitieron 100 veces y se probaron 100 veces más y por lo tanto, consideramos más conveniente reportar el desempeño de clasificación promedio sobre las 100 particiones, en términos de las áreas bajo las curvas ROC (receiver operating characteristic, por sus siglas en inglés), las cuales serán subsecuentemente denominadas AUC.

De hecho, AUC's presentan una manera muy conveniente para medir el desempeño de un clasificador siempre y cuando el umbral de decisión no esté definido, dado que una curva ROC traza sobre un espacio de dos dimensiones el número de falsos positivos (FP's) y negativos (FN's) obtenidos bajo umbrales de clasificación diferentes. Recordando la función $D(w)$ de la Gráfica 1 y de la Ecuación 4, el número de FP's y FN's variará de acuerdo con el valor de ordenada al origen b. La aplicación de curvas ROC y técnicas relacionadas se remontan a varias décadas, con unos ejemplos tempranos como Egan (1975) y Bamber (1975). Más recientemente, ROC's se aplicaron para estudiar problemas económicos y financieros, por ejemplo, Rodriguez y Rodriguez (2006) las aplican para predecir la reprogramación de deuda soberana. Dado que no conocemos los costos de clasificar incorrectamente un banco al asignarlo a la clase de bancarrota, las ROC's presentan una buena manera de medir el desempeño del conjunto de algoritmos propuestos. Algunas buenas monografías sobre curvas ROC y AUC's se encuentran en Engelmann et al. (2003) y Fawcett (2003) (2006).

6.3. Implementación y resultados

Los clasificadores de FDA, logit y probit se implementaron con la función de Matlab `classify` (*Statistics toolbox*, versión 5,0,1). Mientras que para LS-SVM utilizamos el toolbox LSSVMlab de Suykens et al. (2002). Para el entrenamiento se utilizaron los parámetros de validación cruzada por defecto de 10 veces. BFD se implementó mediante el toolbox de Peña Centeno y Lawrence (2006). Mientras que la implementación de WGP utilizó la de Snelson et al. (2003), con el parámetro I fijado en 5 funciones. Dado que WGP's están diseñados para la regresión pero no lo están para la clasificación, fijamos los objetivos a los valores de las etiquetas. Finalmente, para los GPC's utilizamos el código de Rasmussen y Williams (2006). Para todos estos métodos generamos curvas ROC con los valores de salidas dados por las probabilidades a posteriori de clase, excepto para FDA y WGP's.

El Cuadro 2 reporta los promedios de AUC's sobre todas las 100 instancias de prueba de los datos de FDIC. En dicha comparación, LS-SVM, GPC y BFD fueron entrenados con una función de covarianza de la forma (15). Nótese que FDA supera a todos los demás métodos, en términos de los parámetros de la media y de la mediana. Debido a estos resultados pensamos que el conjunto de datos de FDIC se podría separar por una tendencia lineal en lugar de una función no lineal. Por lo tanto, utilizamos una covarianza lineal de la forma

$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{\Theta}_{ard} \mathbf{x}'$, con $\mathbf{\Theta}_{ard} = \{\theta_1, \ldots, \theta_d\}$ para los experimentos reportados en el Cuadro 3. Este tipo de covarianza se conoce como ARD (Sección 5.6) porque asigna un hiperparámetro θ_i a cada dimensión i de los datos.

	FDA	Logistica	Probit	LS-SVM (rbf)	BFD (rbf)	GPC (rbf)
Media	0.866	0.839	0.825	0.823	0.817	0.815
Mediana	0.877	0.841	0.838	0.818	0.816	0.815
Max	0.962	0.949	0.940	0.956	0.950	0.949
Min	0.672	0.679	0.678	0.687	0.681	0.676
STD	0.051	0.056	0.055	0.055	0.051	0.050

Cuadro 2: Resultados promedio de clasificación de los datos del Seguro de Depósito de los Estados Unidos. Reportamos la media, la mediana, el máximo, el mínimo y la desviación estándar del área porcentual bajo la curva ROC (AUC) sobre todas las instancias de prueba de los datos. Los algoritmos comparados son: Análisis de discriminantes de Fisher (FDA), regresiones logísticas y de probit, máquinas de soporte vectorial de mínimos cuadrados (LS-SVM) y dos instancias de procesos Gaussianos (GP's): el discriminante Bayesiano de Fisher (BFD) y de los clasificadores GP (GPC's). Se puede observar que FDA supera al resto de los algoritmos.

	BayLS-SVM	BFD (linard)	GPC (linard)
Media	0.839	0.832	0.869
Mediana	0.853	0.831	0.873
Max	0.952	0.964	0.982
Min	0.627	0.720	0.578
STD	0.061	0.048	0.051

Cuadro 3: Resultados promedio de clasificación de los datos del Seguro de Depósito de los Estados Unidos con algoritmos que tienen un previo ARD (Sección 5.6). Reportamos la media, la mediana, el máximo, el mínimo y la desviación estándar del AUC porcentual sobre todas las instancias de prueba. Los métodos comparados son: máquinas bayesianas de soporte vectorial de mínimos cuadrados (BayLS-SVM), discriminante bayesiano de Fisher (BFD) y Procesos Gaussianos deformados (WGP's). Compárense estos resultados con los de Cuadro 2.

En el segundo experimento (Cuadro 3) observamos resultados mucho mejores para GPC y un mejoramiento moderado para la versión bayesiana de LS-SVM's (Van Gestel et al., 2002) y para BFD *(linard)*, en comparación con las cifras del Cuadro 2. En este caso, la media de GPC es ligeramente mayor que la de FDA, aunque la mediana es todavía menor.

Como experimento final decidimos probar el algoritmo de WGP de Snelson et al. (2003), porque en algunos dominios puede tener más poder expresivo que otros métodos. Los resultados de AUC se muestran en el Cuadro 4. Se puede observar que WGP tiene un mejor desempeño predictivo que el resto de los

modelos comparados, incluyendo FDA y GPC (véase Cuadros 2 y 3); aunque con una varianza mayor.

	WGP
Mean	0.914
Median	0.978
Max	1.000
Min	0.541
STD	0.114

Cuadro 4: Resultados promedio de clasificación de los datos del Seguro de Depósito de los Estados Unidos con el algoritmo de procesos Gaussianos deformados (WGP) de Snelson et al. (2003). Las cifras reportadas son la media, la mediana, el máximo, el mínimo y la desviación estándar del área porcentual bajo la curva ROC (AUC) sobre todas las instancias de prueba de los datos.

Los resultados reportados proporcionan una guía de cómo los GP's pueden ser una herramienta útil para hacer predicciones en la vida real, sin embargo, existen algunas cuestiones que todavía se tienen que considerar. Por ejemplo, la partición aleatoria permitió reducir los efectos de varianza de un conjunto de datos pequeño como el nuestro. Sin embargo, lo anterior conllevó el costo de superponer las muestras, lo que en sentido estricto causa que el cómputo de la desviación estándar en los Cuadros 2-4 sea invalida. En adición, la aplicación de pruebas estadísticas estándar de significancia es difícil debido a esta superposición. Como se discute en la Sección 8, esperamos tener un conjunto de datos más completo en el futuro para superar dichas limitaciones. También notamos que la desviación estándar de los resultados reportados implica que ni GPC's ni WGP's son sustancialmente mejores que FDA. Finalmente, la desviación estándar de WGP's es la mayor de todas (Cuadro 4) mostrando que el algoritmo se desempeña sustancialmente mejor en algunos casos y sustancialmente peor en otros.

6.4. Análisis de características

La presente sección describe brevemente los resultados de aplicar un previo tipo ARD al conjunto de datos FDIC. El estudio se realizó para GPC's, LS-SVM's, BFD y WGP's debido a los resultados de los experimentos reportados en los Cuadros 3 y 4. Sin embargo, solamente se reportan los WGP's porque es el método que presentó los mejores resultados. Debido a la partición aleatoria de los datos, se obtuvieron cien clasificaciones diferentes, cada miembro del conjunto siendo el ordenamiento de 11 razones financieras. Por lo tanto, se consideró más apropiado resumir el número de veces en las que una característica fue asignada a un rango en particular a través de histogramas, como lo presentan las Gráficas 4 y 5. Es importante considerar que para hacer el ranking, ARD mide el grado de variación de una característica y aunque también asume independencia entre las características, esto no necesariamente significa que un rango bajo

sea necesariamente irrelevante para la clasificación. El Apéndice A incluye una descripción de cada razón financiera.

Las Gráficas 4 y 5 muestran cierta regularidad en las primeras cuatro, así como las últimas tres posiciones ocupadas por las características. En el primer grupo, las características seis, cinco, siete y cuatro, correspondientes a Rentabilidad Financiera (ROE) , Rentabilidad Económica (ROA) , Razón de Eficiencia (ER) e Ingreso Operativo Neto (NOI), son las más frecuentemente clasificadas. ROE es un razón financiera relevante para determinar la solidez financiera porque mide la eficiencia de una empresa en generar ganancias por cada dólar del patrimonio de los accionistas. ROA también es una característica plausible porque se usa frecuentemente para comparar el desempeño de las instituciones financieras, por ejemplo los bancos, aunque posiblemente no es una característica tan útil para comparar otros tipos de instituciones, por ejemplo, empresas aseguradoras, que en particular tienen requerimientos de reserva específicos. Como se menciona en el Apéndice A, no existe consenso acerca de la manera de computar ER, sin embargo, un mayor valor de este parámetro usualmente se interpreta como signo de estrés corporativo y por eso, es una buena candidata para predecir bancarrota. Por último, en general se percibe NOI como medida confiable del desempeño de una empresa. Por lo tanto, constituye otra selección razonable.

En el extremo opuesto, el grupo de características menos "relevantes" incluye Margen de interés neto (NIM), Ingreso sin ingresos por interés (NII) y la razón de capital (CR), correspondientes a las características con los números uno, dos y once. El bajo rango de NIM parece contra-intuitivo porque de alguna manera mide la solidez financiera de una institución. Sin embargo, en general se opina que bancos modernos deberían confiar menos en este parámetro debido a las ganancias competitivas que el sector financiero obtuvo recientemente. Con respecto a NII, esta razón no parece tener una relación directa con los síntomas típicos de la tensión financiera que un banco puede tener. Por lo tanto, se requiere un análisis adicional. Finalmente, aunque CR es probablemente una de las razones más importantes para evaluar solidez financiera, fue aquella que ocupa los rangos más bajos. Sin embargo, esta observación puede llevar a conclusiones erróneas si se considera que CR también es una candidata para ocupar el rango seis (de once) y lo que es más importante, el hecho de que el supuesto de independencia entre las características mencionado antes de ARD no necesariamente se cumple para este conjunto de datos.

Sería útil probar si las características *relevantes*: Rentabilidad Financiera, Rentabilidad Económica, Razón de Eficiencia e Ingreso Operativo Neto continúan siendo tan relevantes a través de diferentes periodos de tiempo. Lo anterior en relación con lo dicho por Goodhard (1975) de que el contenido informativo de indicadores económicos y financieros se pierde una vez que se utilizan para hacer predicciones. Eso es una debilidad común de todos los métodos presentados en este documento: FDA, regresión logística y los diferentes tipos de GP's; aunque opinamos que dicho asunto puede ser atendido utilizando métodos que rompen con el supuesto de que los datos son independientes e idénticamente distribuidos. No obstante, en el caso especifico de FDA, es notable que los pesos

estimados de *Z-score* se han mantenido virtualmente sin cambios por aproximadamente cuatro décadas.

7. Riesgo crediticio en los portafolios

Presentamos una nueva familia de técnicas algorítmicas no conocidas por la comunidad de economía computacional, la de procesos Gaussianos interpretados como una distribución a priori en espacio de funciones y cómo se pueden aplicar para la predicción de bancarrota en términos de una tarea de clasificación. Se utilizan algunos productos comerciales, tales como CreditMetrics^{TM}, para cuantificar el riesgo crediticio completo, es decir, estimar las pérdidas de un portafolio a través de la aplicación de un conjunto de técnicas diferentes; incluyendo FDA. En efecto, mediante el enfoque de CreditMetrics (JP, 1997) reconocemos que los GP's son perfectamente adecuados para la integración en tal producto en la forma de un módulo de clasificación binaria. Algo similar pasaría con otros tipos de productos, tales como Moody's KMV^{TM} por ejemplo.

8. Conclusiones y trabajo futuro

Este trabajo ha presentado una revisión comprensiva de algunos métodos estadísticos para la clasificación y su aplicación al problema de predicción de bancarrota. También se incluyó una comparación con herramientas desarrolladas recientemente, tales como varios tipos de procesos Gaussianos para clasificación. Probar nuevas técnicas se justifica debido a que los enfoques estándar para la estimación de un clasificador se basan en métodos parámetricos. Sin embargo, se demostró que utilizando un enfoque parámetrico se omitió una clase de modelos más completos y flexibles, entre ellos, lo de modelos no parámetricos con los procesos Gaussianos.

Los GP's son una generalización de la densidad Gaussiana a espacios funcionales de dimensión infinita y se aplican naturalmente para tareas de inferencia Bayesiana, dadas sus propiedades analíticas simples y facilidad de uso. Sin embargo, estas características no les impiden ser aplicadas en problemas complejos, como por ejemplo la separación de datos en distintas clases. En este trabajo, utilizamos datos del Seguro de Depósito de los Estados Unidos para mostrar cómo instancias diferentes de GP's generaron resultados de clasificación potencialmente competitivos con respecto a técnicas bien establecidas como el *Z-score* de Altman (es decir, análisis de discriminantes) y regresión logística; aunque admitimos que la definición experimental estuvo lejos de ser óptima.

Un corolario interesante del formalismo Bayesiano es que ciertos previos llevan al *ranking* y a la eliminación automática de características si se realiza la inferencia y los GP's no son la excepción. Dicho corolario se conoce como determinación automática de relevancia y se realiza siempre y cuando un parámetro de la distribución a priori se asigne a cada dimensión de los datos (en nuestro caso las dimensiones dadas por cada una de las razones financieras de los

datos del FDIC). Con el objetivo de entender mejor cuáles razones financieras han sido más importantes en la tarea de clasificación, se probaron algunas funciones de covarianza tipo ARD y los resultados mostraron que para los procesos Gaussianos deformados, la Rentabilidad Financiera (ROE), la Rentabilidad Económica (ROA), Razón de Eficiencia (ER) y el Ingreso Operativo Neto (NOI) tuvieron los rangos más altos. La Razón de Capital (CR), una razón generalmente considerada relevante para la evaluación de solidez financiera, se clasificó en bajas posiciones.

Planeamos expandir el presente trabajo en varias direcciones. Primero, la evaluación de la solidez financiera de instituciones bancarias mexicanas con algunas de estas herramientas basadas en GP's sería útil para realizar predicción automatizada de bancarrota, la cual todavía está en sus primeras etapas en este país. Segundo, expandir nuestro conjunto de datos para incluir más razones financieras y otros tipos de variables ayudaría a aumentar nuestro entendimiento de la tarea de predicción de bancarrota. De hecho, quisiéramos probar cuál es el efecto de revisar el estado financiero en el tiempo $t + 1$, si originalmente se publicó en t. Tercero, introduciendo un componente dependiente del tiempo, dichos tipos de modelos podrían volverse útiles para la alerta temprana y tal vez ayudar a superar las limitaciones impuestas por la ley de Goodhard. Además sería útil analizar episodios particulares de estrés financiero, como la crisis económica global (la cual aún permanece) y ver cuál algoritmo se desempeña mejor. Finalmente, quisiéramos aplicar un mejor diseño experimental para comparar los algoritmos sobre una mejor base, dado que personas como Verikas et al. (2009) han observado que cada nuevo método propuesto es casualmente mejor que todos los anteriores.

A. Apéndice

Una descripción breve de las razones financieras que componen los datos de FDIC.

Razón 1. Margen de interés neto (NIM) es la diferencia entre el interés pagado por prestatarios y el interés pagado a sus prestadores.

Razón 2. Ingreso -sin ingresos por intereses- a activos productivos (NII) es la suma de los siguientes tipos de ingresos: basados en honorarios, comercio, resultando de actividades fiduciarios y otros no relacionados con el interés.

Razón 3. Gastos sin interés a activos productivos (NIX) comprende básicamente tres tipos de gastos: gastos personales, tenencia y otros gastos operativos.

Razón 4. Ingreso operativo neto a activos es relacionado al ingreso bruto de una empresa asociado con sus propiedades menos los gastos operativos.

Razón 5. Rentabilidad Económica (ROA es un indicador de qué tan rentable es una empresa en relación con sus activos totales. ROA se calcula como la razón entre las ganancias totales de una empresa durante el año y sus activos totales.

Razón 6. Rentabilidad Financiera (ROE) es una medida de la tasa de rendimiento del patrimonio de los propietarios de acciones comunes. ROE se estima como las ganancias netas del año (después/excluyendo dividendos de acciones preferentes pero antes/incluyendo dividendos de acciones comunes) divididas por patrimonio total (excluyendo acciones preferentes).

Razón 7. Razón de Eficiencia (ER) es una razón utilizada para medir la eficiencia de una empresa, aunque no siempre se calcula de la misma manera.

Razón 8. Activos fijos/no corrientes más otros inmuebles en posesión de activos (NCA) son aquellos que no se pueden convertir fácilmente en efectivo, por ejemplo, inmuebles, maquinaria, inversión de largo plazo o patentes.

Razón 9. Es la razón de efectivos más obligaciones de gobierno y del tesoro de EE.UU. a activos totales.

Razón 10. Capital Propio/ de Patrimonio neto a activos (EC) es el capital levantado por los propietarios.

Razón 11. Razón de capital (primario) (CR) también conocida como Razón de Apalancamiento se calcula como el capital de Tier 1 dividido por el promedio de activos totales consolidados.

Referencias

A. F. Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions On Neural Networks*, 12:929–935, July 2001.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, USA, 2006.

B. Efron. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7:1–26, 1979.

A. Estrella, S. Park, and S. Peristiani. Capital ratios as predictors of bank failure. *Federal Reserve Bank of New York Economic Policy Review*, pages 33–52, July 2000.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179, 1936.

C. Goodhard. Monetary relationships: A view from Threadneedle Street. *Papers in Monetary Economics*, 1975. Reserve Bank of Australia.

D. G. Krige. Two-dimensional weighting moving average trend surfaces for ore evaluation. *Journal of the South African Institute of Mining and Metallurgy*, 1966.

D. J. C. Mackay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

D. J. C. Mackay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series*, pages 133–165. Springer-Verlag, Berlin, Germany, 1998.

G. J. MacLachlan. *Discriminant Analysis and Pattern Recognition*. John Wiley & Sons Ltd., New York, USA, 1991.

T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, USA, 1996.

T. Peña Centeno and N. D. Lawrence. Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. *Journal of Machine Learning Research*, 7:455–491, 2006.

C. E. Rasmussen. Gaussian processes in machine learning. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science/Artifical Intelligence*. Springer-Verlag, 2004.

M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):69–106, 2004.

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped gaussian processes. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, USA, 2003. MIT Press.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147, 1974.

J. A. Suykens, T. Van Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

T.Ñ. Thiele. *Theory of Observations*. Layton, London, UK, 1903. Reprinted in *Annals of Mathematical Statistics* 2:165-308, 1931.

T. Van Gestel, J. A. Suykens, G. Lanckriet, A. Lambrechts, B. de Moor, and J. Vandewalle. Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel discriminant analysis. *Neural Computation*, 14(5):1115–1147, 2002.

A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey. *Soft Computing - A Fusion of Foundations, Methodologies and Applications (Online)*, September 2009.

C. K. Williams. Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning in Graphical Models*, D, Behavioural and social sciences 11. Kluwer, Dordrecht, The Netherlands, 1999.