Datos funcionales en redes neuronales

Los datos en forma de funciones pueden provenir de diferentes fenómenos, como por ejemplo, la presión arterial medida a tiempo continuo en un intervalo de 24 horas o las curvas generadas por un electrocardiograma. Puede incluso extenderse al caso multivariable, como la escritura de texto a mano sobre una hoja rectangular donde cada dato se puede representar por una función $f : [a, b] \times [c, d] \to [0, 1]$ siendo $(x, y) \in [a, b] \times [c, d]$ un punto de la hoja y $f(x, y)$, la intensidad del trazo en ese punto.

Usualmente no se conoce al dato funcional en todo su dominio, por ejemplo porque sin conocer la fórmula analítica es imposible registrar todos los valores de la función en un dominio continuo, éste es el problema de la representación de los datos. Formas de afrontar eso es interpolando por funciones sobre un espacio adecuado a la estrcutura de los datos trabajados.

Solucionado este problema, se buscará adaptar varios métodos de redes neuronales para regresión del caso de input sobre $\mathbb{R}^q$ al caso de inputs funcionales. En este caso se tratarán las redes de funciones de activación de base radial (Radial-Basis Function Network o RBFN) y el perceptrón multicapa (Multi-Layer Perceptron o MLP).

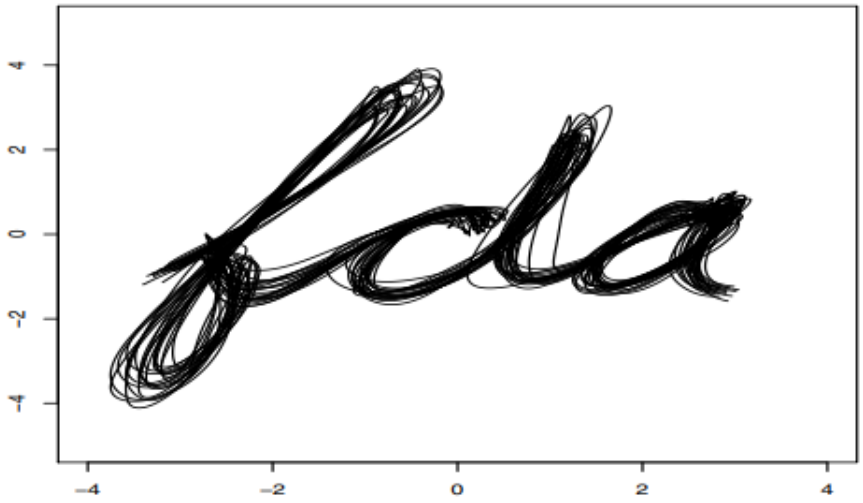

Figura 1: 20 muestras de escritura de la palabra *fda*. La unidad de los ejes está en centímetros.

1. Espacios funcionales

En este caso nos centraremos en el espacio de funciones $f : V \subset \mathbb{R}^p \to \mathbb{R}$ de cuadrado integrable, notado como $L^2(V)$. Lo consideraremos con su producto interno usual $\langle f, g \rangle = \int_V f(x)g(x)dx$, lo que permite tratar un espacio de Hilbert y extender muchas propiedades e ideas de los algoritmos de datos multivariados como el producto escalar, la ortogonalidad y las distancias.

Un ejemplo de esto es el algoritmo de clusterización de k *medias* requiere que el espacio vectorial tenga una métrica. Esto pues partiendo de k puntos *representantes*, se reparten los datos a clusterizar en k grupos donde la característica del grupo $i \in \{1, \dots, k\}$ es que para todos los datos allí el representante más cercano según la métrica usada es el i.

Otro ejemplo se da en modelos de regresión, donde a una $Y \in \mathbb{R}$ aleatoria se la busca modelar en términos de $X \in \mathbb{R}^p$. Una forma es linealmente como $Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon$, siendo ε una variable aleatoria de esperanza nula e independiente de X. Este modelo lineal se puede reescribir usando el producto interno usual de $\mathbb{R}^p$ como $Y = \langle \beta, X \rangle_{\mathbb{R}^p} + \varepsilon$, siendo $\beta = (\beta_1, \dots, \beta_p)$. Si en cambio se considera un input funcional, por ejemplo $X \in L^2([a,b])$, entonces es razonable proponer un modelo análogo usando ahora el producto interno en $L^2([a,b])$ como $Y = \int_a^b \beta(t)X(t)dt + \varepsilon$, siendo β una función en ese espacio.

2. Representación de funciones y bases usadas

Por lo general se observan n funciones donde a la observación i se la conoce en algunos puntos como $(x_j^{(i)}, y_j^{(i)})_{1 \le j \le m_i}$. El análisis de datos funcionales asume que para cada una de ellas hay una función $g^{(i)} \in L^2(V)$ tal que $y_j^{(i)} = g^{(i)}(x_j^{(i)}) + \varepsilon_j^i \ \forall 1 \le j \le m_i$ siendo $\varepsilon_j^{(i)}$ un ruido aleatorio. Las funciones $g^{(i)}$ son desconocidas y se trabaja con aproximaciones de ellas.

La propuesta para tomar una representación de cada observación es tomarlas sobre un subespacio conocido $\mathcal{A} \subset L^2(V)$ de dimensión finita. Sea $\{\phi_k\}_{1 \le k \le q}$ una base de $\mathcal{A}$, como cada elemento $u \in \mathcal{A}$ queda biunivocamente determinado por el valor de sus coordenadas $(\alpha_k(u))_{1 \le k \le q}$ en esa base, entonces $g^{(i)}$ se la aproxima por $\tilde{g}^{(i)} = \sum_{k=1}^{q} \alpha_k(\tilde{g}^{(i)})\phi_k$ donde sus coordenadas $\alpha_k(\tilde{g}^{(i)})$ se las obtiene minimizando la siguiente suma de cuadrados de los residuos

$$\sum_{j=1}^{m_i} \left(y_j^{(i)} - \tilde{g}^{(i)}(x_j^{(i)}) \right)^2.$$

En general, la minimización de estas funciones cuadráticas para encontrar $\tilde{g}^{(i)}$ tienen un costo $O(m^{(i)}q^2)$.

Trabajar con una base ortonormal de $\mathcal{A}$ conlleva a simplificar operaciones en el algoritmo a realizar pues el producto interno entre dos funcionales u y v coincide con el p.i. de $\mathbb{R}^q$ (el usual) entre sus coordenadas en esa base. En general, el cálculo del producto interno entre u y v está dado por

$$\langle u, v \rangle = \sum_{k=1}^{q} \sum_{l=1}^{q} \alpha_k(u)\alpha_l(v)\langle \phi_k, \phi_l \rangle,$$

que en forma matricial se reescribe como

$$\langle u, v \rangle = \alpha(u)^T \Phi \alpha(v),$$

siendo $\alpha(u)$ y $\alpha(v)$ las coordenadas de u y v respectivamente y Φ la matriz simétrica y definida positiva dada por $\Phi_{kl} = \langle \phi_k, \phi_l \rangle$. Se observa que si la base es ortonormal, Φ

es la matriz identidad y se tiene $\langle u, v\rangle = \langle \alpha(u), \alpha(v)\rangle_{\mathbb{R}^q}$. En otro caso, los productos no coinciden y luego las distancias tampoco. Una propuesta para recuperar esta propiedad es reescalar los coeficientes. Sea la descomposición de Cholesky $\Phi = U^T U$, durante los métodos se trabaja en $\mathbb{R}^q$ con $U\alpha(u)$ pues en este caso se tiene $\langle u, v\rangle = \langle U\alpha(u), U\alpha(v)\rangle_{\mathbb{R}^q}$.

Las bases usadas suelen ser las obtenidas a partir del desarrollo de Fourier o también los B-splines, éste últimos no determinan bases ortonormales, por lo que requiere el cálculo de Φ y su posterior de descomposición. Para computarlo se necesita calcular varias integrales, las cuales se pueden obtener por métodos de cuadraturas o simulaciones de Montecarlo.

3. Redes de funciones de activación de base radial

Las redes involucradas constan de 3 capas: un input $x \in \mathbb{R}^q$, una capa oculta de p neuronas con funciones de activación de la forma $\varphi(d(x, c))$ y un sólo output y que combina linealmente la salida de las neuronas. En síntesis, para un input x, la salida es de la forma

$$y = \sum_{i=1}^{N} a_i \varphi_i(d(x, c_i)),$$

siendo φ_i funciones de una variable que se evalúan respecto de la distancia a un centro $c_i \in \mathbb{R}^q$. Usualmente la función usada es la gaussiana $\varphi_i(r) = e^{-r^2}$ pero en general se consideran funciones tales que no aporten valores cuando se evalúen en distancia grandes, por lo que se pide que $\lim_{r\to\infty} \varphi(r) = 0$.

La ventaja de este tipo de redes es la poca cantidad de capas que usa, disminuyendo la costo computacional de su implementación como la propiedad de que son aproximadores universales bajo ciertas condiciones. Para mostrar esto último, reescribimos la relación como

$$y = \sum_{i=1}^{N} a_i K_i \left(\frac{x - c_i}{\sigma_i} \right),$$

donde $K_i : \mathbb{R}^q \to \mathbb{R}$ son tales que $\lim_{||x||\to\infty} K(x) = 0$ y $K_i(x) = K_i(y)$ para todo x, y tales que $||x|| = ||y||$. Previo a demostrar las propiedades de aproximadores universales, precisamos el siguiente lema de Análisis Real.

Lema 1. *Sean $f \in L^p(\mathbb{R}^q)$, $p \in [1, \infty)$ y $K \in L^1(\mathbb{R}^q)$ tal que $\int_{\mathbb{R}^q} K(x)dx = 1$. Definiendo para cada $\varepsilon > 0$, $K_\varepsilon(x) := \frac{1}{\varepsilon^q}\varphi(\frac{x}{\varepsilon})$ entonces $||K_\varepsilon * f - f||_{L^p} \to 0$ cuando $\varepsilon \to 0$.*

Demostración. Ver *Measure and Integral, An Introduction to Real Analysis* de R. Wheeden y A. Zygmund (1977), pág 148.

Sea K una función de activación con la propiedad antes mencionada y f una función cualquiera de $L^p(\mathbb{R}^q)$ a aproximar, se quiere ver si existe una función $\rho(x) = \sum_{i=1}^{N} a_i K\left(\frac{x-c_i}{\sigma_i}\right)$ tal que ρ esté tan cerca de f como se desee. Formalmente, sea una métrica d fija, se quiere

ver que para cualquier $\varepsilon > 0$ (parámetro de cercanía) existen $a_i, c_i, \sigma \in \mathbb{R}$, con $\sigma > 0$, $N \in \mathbb{N}$ tales que $d(\rho, f) < \varepsilon$. Definiendo el conjunto

$$S_K = \left\{ \sum_{i=1}^{N} a_i K\left(\frac{x - c_i}{\sigma}\right) : a_i, c_i, \sigma \in \mathbb{R}, \sigma > 0, N \in \mathbb{N} \right\},$$

hay que probar que S_K es denso en $L^p(\mathbb{R}^q)$ con la métrica d.

Teorema 1. *Sea $K \in L^1(\mathbb{R}^q)$ tal que $\int_{\mathbb{R}^q} K(x)dx \neq 0$, acotada y continua c.t.p, entonces S_K es denso en $L^p(\mathbb{R}^q)$ con la norma usual de ese espacio.*

Demostración. Ver *Universal Approximation Using Radial-Basis-Function Networks* de J. Park e I. Sandberg (1991), pág 250.

Este resultado se puede extender para aproximar funciones continuas, en este caso usando otra métrica basada en la del supremo.

Teorema 2. *Sea $K \in L^1(\mathbb{R}^q)$ tal que $\int_{\mathbb{R}^q} K(x)dx \neq 0$ continua, entonces S_K es denso en $C(\mathbb{R}^q)$ con la métrica d dada por*

$$d(f, g) = \sum_{n=1}^{\infty} 2^{-n} \frac{||f - g||_{L^\infty([-n,n]^q)}}{1 + ||f - g||_{L^\infty([-n,n]^q)}}$$

Demostración. Ver *Universal Approximation Using Radial-Basis-Function Networks* de J. Park e I. Sandberg (1991), pág 253.

Estos resultados son más potentes en sentido de que no requieren hipótesis como la simetría radial en la función K, aunque precisa que no haya más de un K_i ni σ_i.

Volviendo al caso de inputs funcionales, en este caso, la distancia usada sería correspondiente a L^2, es decir

$$d(x, c_i) = \left(\int_V (x(t) - c_i(t))^2 dt \right)^{1/2}.$$

De hecho, se puede dar una forma más general. Así como la distancia euclídea en $\mathbb{R}^q$ dada por $||v-w|| = [(v-w)^T(v-w)]^{1/2}$ se puede generalizar con $||v-w||_A = [(v-w)^T A(v-w)]^{1/2}$ con A simétrica y definida positiva, la distancia anteriorL^2 se puede generalizar como

$$d(x, c_i) = \left(\int_V \int_V (x(t) - c_i(t)) w_i(t, t') (x(t') - c_i(t')) dt dt' \right)^{1/2},$$

siendo w_i una función en $L^2(V \times V)$ definida positiva.

4. Perceptrón multicapa

Este tipo de redes neuronales, en el caso de datos multivariados consisten en capas ocultas cuyas neuronas reciben las salidas de todas las neuronas de la capa anterior, escribamos $x \in \mathbb{R}^q$ (la capa anterior posee q neuronas) y devuelve el valor

$$\phi(w^T x + b),$$

siendo $\phi : \mathbb{R} \to \mathbb{R}$ una función (de activación) no lineal, $w \in \mathbb{R}^q$ y $b \in \mathbb{R}$. Esto se puede extender al caso funcional de $L^2(V)$, reemplazando el producto $w^T x$ por el producto interno usual de ese espacio, es decir, se consideran neuronas cuyo input es $x \in L^2(V)$ de la forma

$$\phi\left(\int_V w(t)x(t)dt + b\right), \tag{1}$$

siendo ϕ y b como antes y $w \in L^2(V)$. Esto también se puede generalizar al caso de q inputs funcionales: dados $x_1, \ldots, x_q \in L^2(V)$, la neurona considerada es

$$\phi\left(\sum_{i=1}^{q} \int_V w_i(t)x_i(t)dt + b\right),$$

con $w_1, \ldots, w_q \in L^2(V)$.

Como el output de estas neuronas generalizadas son valores numéricos, este tipo de neuronas sólo se usa en la primer capa oculta que es la que recibe los inputs funcionales. El resto de capas posteriores usan las neuronas usuales.

Estas redes neuronales también traen la propiedad de ser aproximadores universales, esto quiere decir que para cualquier función $G : L^2(V) \to \mathbb{R}$ ($(L^2(V))^q$) si se consideran más inputs funcionales) y una métrica d fija sobre ese espacio, se quiere ver que para cualquier $\varepsilon > 0$ (parámetro de cercanía) existe otra función allí $\rho(f) = \sum_{i=1}^{N} a_i \phi\left(\int_V w(t)_i f(t)dt + b_i\right)$ (aproximación por una sola capa oculta) tales que $d(\rho, G) < \varepsilon$. Aquí se presentará una versión más general sobre espacios $L^p(V)$. En particular se trabajará aproximando G con funciones en el siguiente conjunto:

$$S_\phi^W = \left\{\rho(f) = \sum_{i=1}^{N} a_i \phi\left(\int_V w_i(t)f(t)dt + b_i\right) : w_i \in W, a_i, b_i \in \mathbb{R}, N \in \mathbb{N}\right\},$$

siendo W subconjunto de las funciones de $\mathbb{R}^q$ en $\mathbb{R}$ y $\phi : \mathbb{R} \to \mathbb{R}$.

La primera propiedad de aproximación universal es la siguiente, que es más general pues se considera aproximaciones con inputs en L^p

Teorema 3. *Sean $p, p' \in (1, \infty)$ tales que $\frac{1}{p} + \frac{1}{p'} = 1$, M un subconjunto denso de $L^{p'}(\mathbb{R}^q)$ y $\phi : \mathbb{R} \to \mathbb{R}$ Riemann integrable sobre compactos, entonces para todo $K \subset L^p(\mathbb{R}^q)$ compacto, S_ϕ^M contiene un subconjunto denso en $C(K, \mathbb{R})$ con la métrica del supremo.*

Demostración. Ver *Theoretical Properties of Functional Multi Layer Perceptrons* de F. Rossi, B. Conan-Guez, F. Fleuret (2002), pág 3.

Se puede extender este resultado también sobre datos funcionales de L^1. Este resultado nos muestra que, en nuestro caso, si consideramos inputs sobre un compacto de $L^2(V)$, entonces G es aproximable por funciones de $S_\phi^{L^2(V)}$.

Recordemos que nuestra idea era trabajar no con datos de $L^2(V)$ sino con representantes de estos en dimensión finita. Es decir, Sea $\{\varphi_i\}_{i\in\mathbb{N}}$ una base ortonormal en $L^2(V)$, en lugar de trabajar con $f \in L^2(V)$, fijando $n \in \mathbb{N}$ usaremos una proyección sobre el subespacio $\langle\varphi_1, \ldots, \varphi_n\rangle$:

$$\Pi_n(f) = \sum_{i=1}^{n} \beta_i \varphi_i,$$

que en realidad $\beta_i = \int_V f(t)\varphi_i(t)dt$. En la práctica, por desconocimiento del input completo f y por sólo conocerlo en algunos puntos, se pueden estimar los coeficientes β_i por mínimos cuadrados. Esto da lugar a un estimador $\widehat{\Pi}_n(f) = \sum_{i=1}^{n} \widehat{\beta}_i \varphi_i$ que es fuertemente consistente para $\Pi_n(f)$ (con la convergencia asociada a L^2). También en la red usada, el peso $w(t)$ se "trunca", es decir, sea $\{\psi_j\}_{j\in\mathbb{N}}$ una base de $L^2(V)$ y $p \in \mathbb{N}$, se considera $w \in \langle\psi_1, \ldots, \psi_p\rangle$ por lo que queda unívocamente determinado los p valores de sus coeficientes en esa base. A continuación, se da un resultado sobre la aproximación de redes usando proyecciones.

Teorema 4. *Sean ϕ una función de activación continua no polinómica, $G \in C(K, \mathbb{R})$ con K un compacto de $L^2(V)$, y $\varepsilon > 0$, existe $n, p \in \mathbb{N}$ y $\rho \in S_\phi^{L^2(V)}$ tales que*

$$\sup_{f\in K} |\rho(\Pi_n(f)) - G(f))| < \varepsilon$$

Demostración. Ver *Multi-Layer Perceptrons for Functional Data Analysis: a Projection Based Approach* de F. Rossi, B. Conan-Guez (2002), pág 4.

5. Bibliografía

- Wheeden R., Zygmund. A (1977) *Measure and Integral, An Introduction to Real Analysis*
- Ramsay, J., Silverman B. (1997) *Functional Data Analysis*
- Park, J., Sandberg I. (1991) *Universal Approximation Using Radial-Basis-Function Networks.*
- Rossi, F., Delannay, N., Conan-Guez, B., Verleysen M. (2004) *Representation of Functional Data in Neural Networks.*
- Rossi, F., Conan-Guez, B., Fleuret, F. (2002) *Theoretical Properties of Functional Multi Layer Perceptrons.*
- Rossi, F., Conan-Guez, B. (2002) *Multi-Layer Perceptrons for Functional Data Analysis: a Projection Based Approach.*