

# Analysis of Neural Networks. Applications to Interpretability and Uncertainty

Gabriel Marín Sánchez

Directors: Antoni Benseny Ardiaca i Alberto Rubio Muñoz

Barcelona, 8 de Juliol 2020



UNIVERSITAT DE  
BARCELONA

Facultat de Matemàtiques  
i Informàtica



**SCRM**  
INTERNATIONAL HUB

# Continguts

1 Introducció

2 Xarxes Neuronals Directes

3 Aprendentatge Supervisat

4 Aprendentatge No Supervisat

5 Interpretabilitat

6 Incertesa

7 Conclusions

- Inici de les xarxes neuronals:
  - Neurona de McCulloch-Pitts (1943)
  - Perceptró (1958)
- Segona onada: Teoremes d'aproximació universal (Cybenko i Hornik, 1990s)
- Les xarxes neuronals han revolucionat àrees com la visió artificial o el processament de text.
- Tot i això, en general aquesta tecnologia és tractada com una caixa negra.
- Objectius:
  - Donar un tractament matemàtic.
  - Mostrar i desenvolupar mètodes per entendre el perquè del bon funcionament.

# Continguts

1 Introducció

2 Xarxes Neuronals Directes

3 Aprendentatge Supervisat

4 Aprendentatge No Supervisat

5 Interpretabilitat

6 Incertesa

7 Conclusions

# Descripció

- Sigui  $f'$  una funció, l'objectiu de la xarxa neuronal és trobar  $f$  que **aproximi**  $f'$ .
- Funció lineal, on  $w$  es el **pes** i  $b$  el **biaix**:

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

- Composició per components:

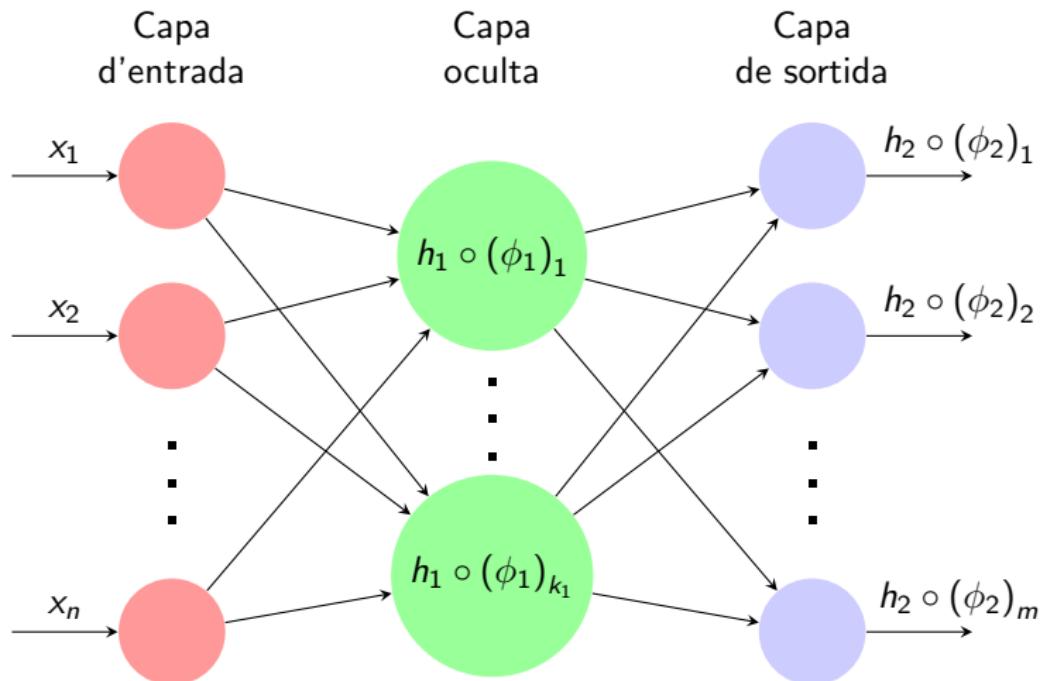
$$(g \bullet f)(x) = ((g \circ f_1)(x), \dots, (g \circ f_m)(x))$$

- Xarxa directa amb  $L + 1$  capes:

$$f_\theta = h_L \bullet \phi_L \circ \cdots \circ h_1 \bullet \phi_1$$

$h_i$  *funcions d'activació*

# Representació



- Funció de pèrdua: Mesura l'error de la prediccó de la xarxa.
- Optimització de la funció de pèrdua: gradient descendent.

# Funció de Pèrdua

Siguin  $x = \{x_i\}_{i=1,\dots,N}$  els valors d'entrada,  $y' = \{y'_i\}_{i=1,\dots,N}$  els objectius i  $f_\theta$  la funció de la xarxa neuronal.

- Regressió: **Error quadràtic mig.**

$$J_\theta^{MSE}(x, y') = \frac{1}{N} \sum_{i=1}^N \|y' - f_\theta(x)\|^2$$

- Classificació: **Pèrdua d'entropia creuada.**

- Múltiples classes:

$$J_\theta^{MCE}(x, y') = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in C} y'_c \log(f_{\theta,c}(x))$$

- Cas binari:

$$J_\theta^{MCE}(x, y') = -\frac{1}{N} \sum_{i=1}^N [(y' \log(f_\theta(x)) + (1 - y') \log(1 - f_\theta(x)))]$$

# Optimització

- Mètode del gradient descendent.

$$w_{ij}^l = w_{ij}^l - \eta \frac{\partial J_\theta}{\partial w_{ij}^l}$$

- Altres: Adam, Adagrad, etc.
- Càcul dels gradients: **Retropropagació**.

$$\frac{\partial J_\theta}{\partial w_{ij}^{(l)}} = \beta_j^{(l)} z_i^{(l-1)} \quad ; \quad \frac{\partial J_\theta}{\partial b_j^{(l)}} = \beta_j^{(l)}$$

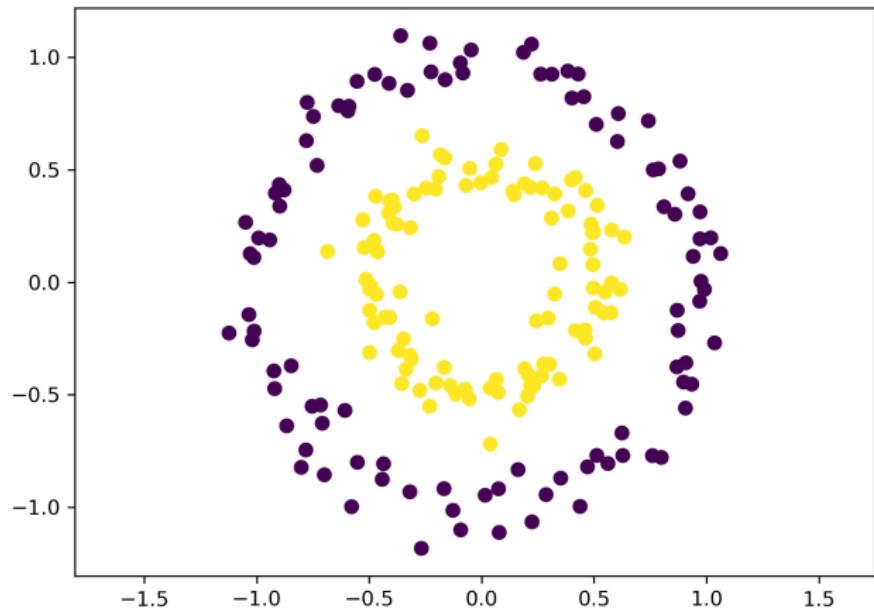
$$\beta_j^{(l)} = \begin{cases} \frac{\partial J_\theta}{\partial z_j^{(l)}} \frac{\partial h_l(u_j^{(l)})}{\partial u_j^{(l)}}, & \text{if } l = L \\ \left( \sum_{a \in A} w_{ja}^{(l+1)} \beta_a^{(l+1)} \right) \frac{\partial h_l(u_j^{(l)})}{\partial u_j^{(l)}}, & \text{otherwise} \end{cases}$$

# Continguts

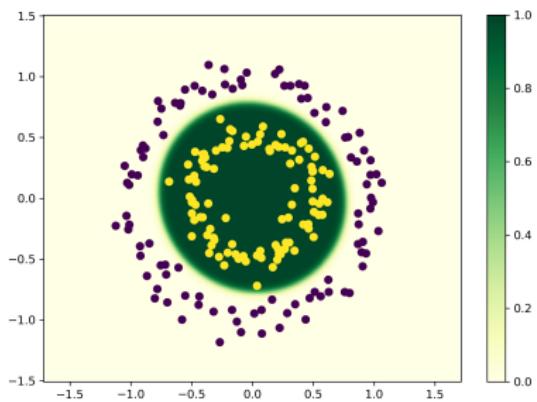
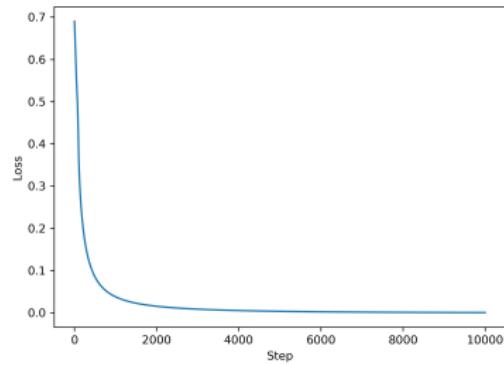
- 1 Introducció
- 2 Xarxes Neuronals Directes
- 3 Aprendentatge Supervisat
- 4 Aprendentatge No Supervisat
- 5 Interpretabilitat
- 6 Incertesa
- 7 Conclusions

- En aquest tipus d'aprenentatge, cada valor d'entrada té associat un objectiu específic.
- La xarxa neuronal prova d'aproximar la funció que mapeja els valors d'entrada amb els objectius.
- Dos problemes de classificació:
  - Classificació binària de dos cercles concèntrics.
  - Classificació binària de punts en un segment de 1D.

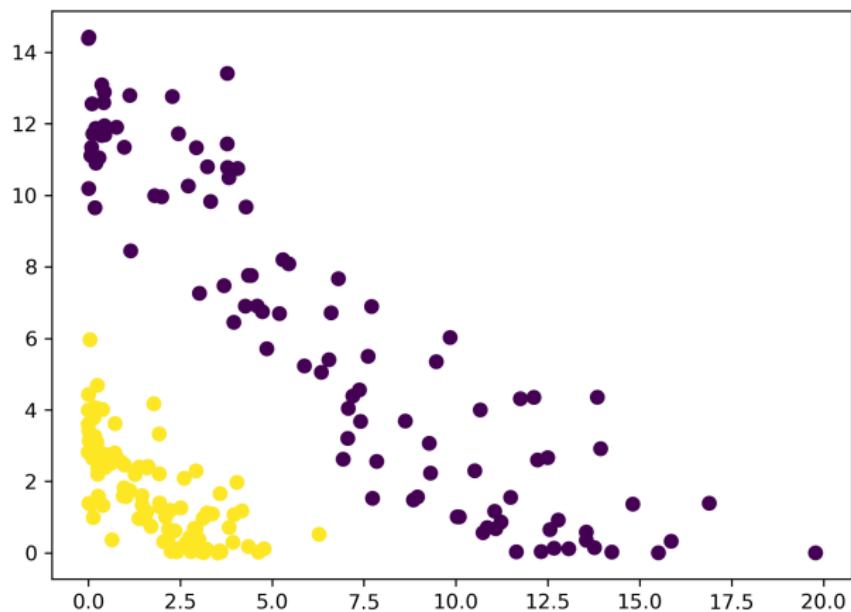
# Dos Cercles Concèntrics



# Funció d'activació: $x^2$



# Funció d'activació: $x^2$



# Funció d'activació: tanh

# Funció d'activació: tanh

# Segment 1D

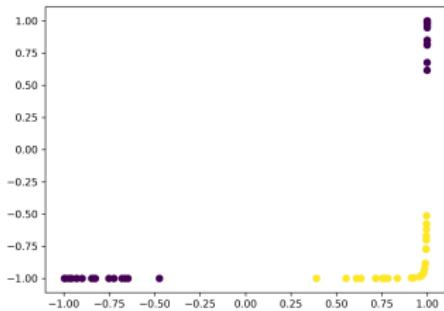
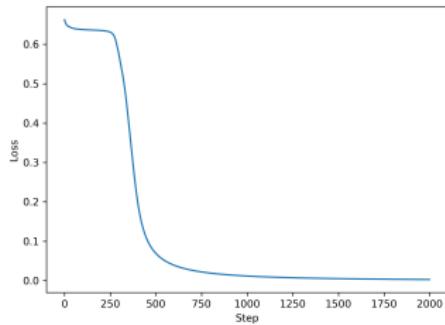
- Dividim un segment de 1D en  $s$  subgrups de 30 punts cada un.
- Assignem la categoria 0 als subgrups d'índex imparell i 1 a la resta.



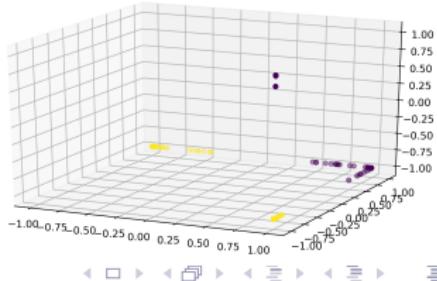
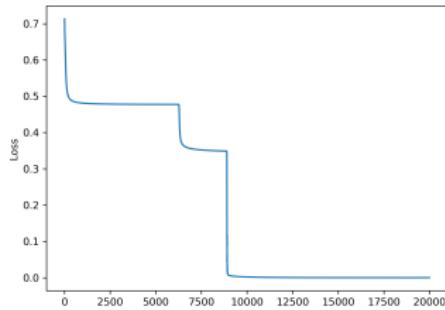
- Considerem dos escenaris:
  - Funció d'activació injectiva.
  - Funció d'activació no injectiva.

# Funció d'activació injectiva: tanh

$s = 3$



$s = 4$

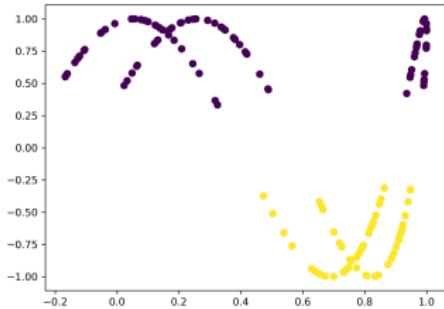
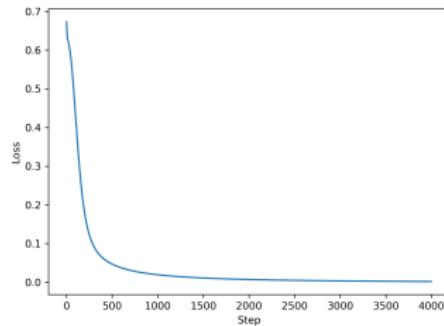


## Funció d'activació no injectiva: sin

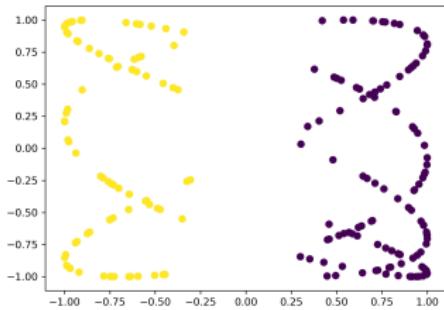
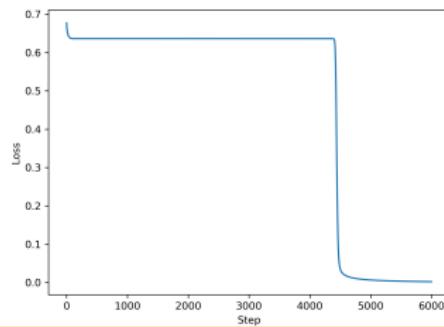
$$s = 4$$

# Funció d'activació no injectiva: sin

$s = 5$



$s = 7$



# Continguts

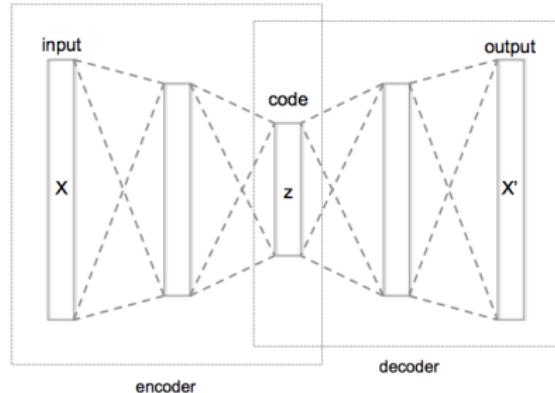
- 1 Introducció
- 2 Xarxes Neuronals Directes
- 3 Aprendentatge Supervisat
- 4 Aprendentatge No Supervisat
- 5 Interpretabilitat
- 6 Incertesa
- 7 Conclusions

# Aprendentatge No Supervisat

- En aquest tipus d'aprendentatge, els valors d'entrada **no tenen associats objectius específics.**
- La xarxa neuronal prova d'aprendre **propietats de les dades.**
- Problema de representació utilitzant sèries temporals i l'arquitectura d'Autoencoder.

# Autoencoder

- L'objectiu és obtenir una **aproximació** de les dades inicials.
- Aquest model **codifica** les dades a un espai de dimensió inferior i el **descodifica** a l'original.
- Actua com a **coll d'ampolla d'informació**.



# Sèries Temporals

- Dades borsàries diàries de 500 companyies aleatòries, obtingudes a través de *Yahoo! finance*.
- Un total de 3017 sèries temporals de longitud 365, corresponent al nombre de dies en un any.

# Sèries Temporals: Transformació

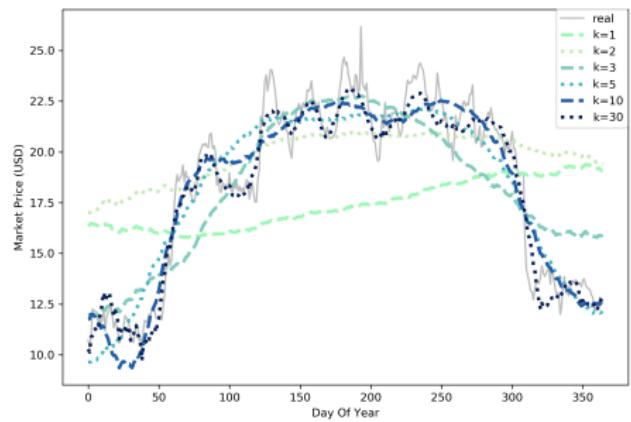
- Les sèries temporals poden tenir valors molt diferents i aleshores el **rang** de valors pot ser molt gran.
- Dues transformacions:
  - **Normalització:** Valors entre 0 i 1.

$$x' = \frac{x - \min}{\max - \min}$$

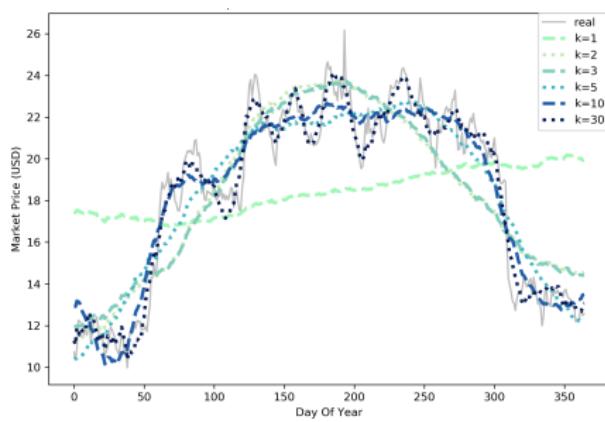
- **Estandardització:** Mitjana 0 i variància 1.

$$x' = \frac{x - \mu}{\sigma}$$

# Sèries Temporals: Resultats

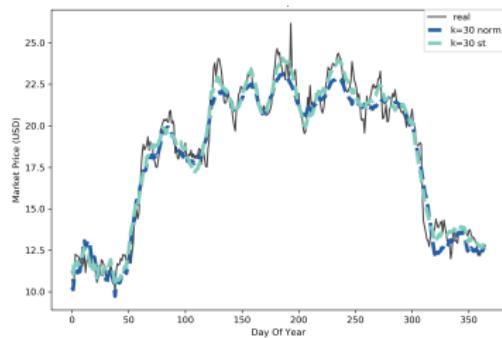
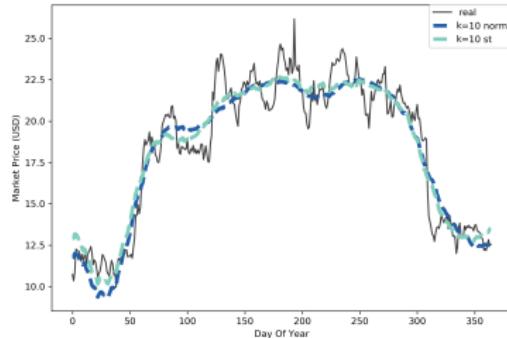
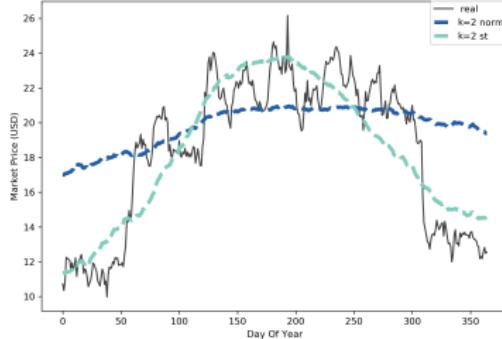
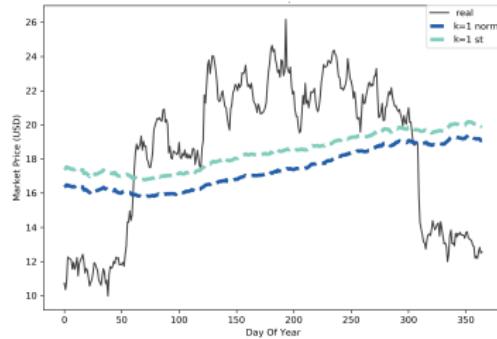


(a) Normalització.

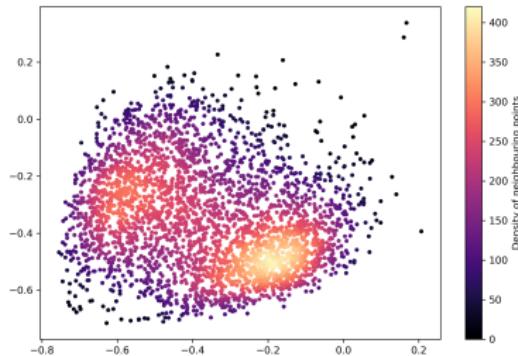


(b) Estandardització.

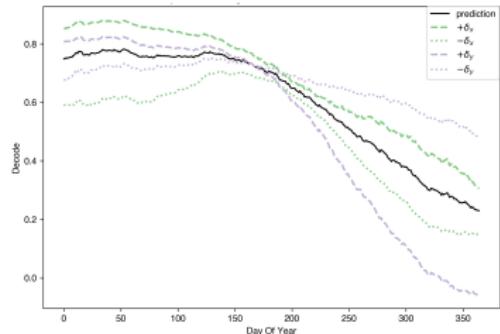
# Sèries Temporals: Comparació



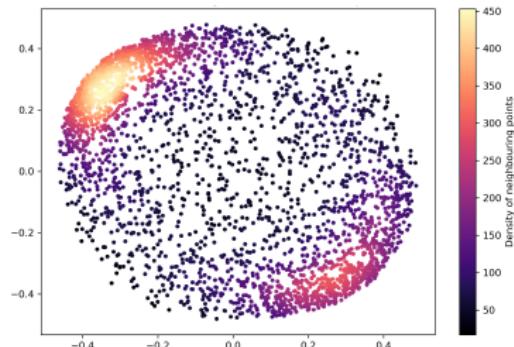
# Sèries Temporals: Espai Latent — $k = 2$



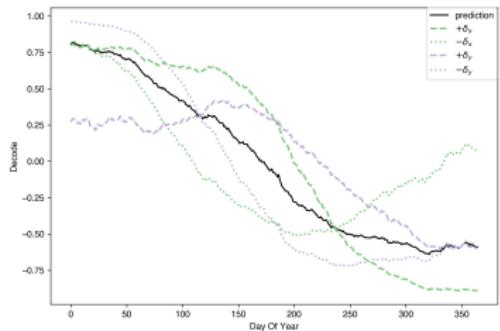
(a) Espai Latent Normalització.



(b) Decodificació Normalització.

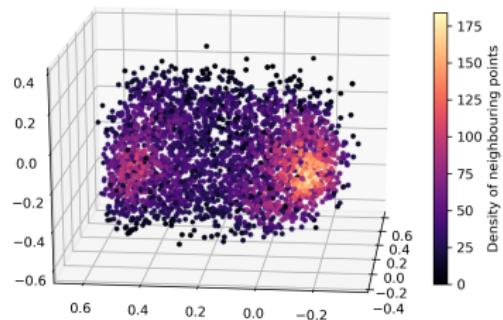


(c) Espai Latent Estandardització.

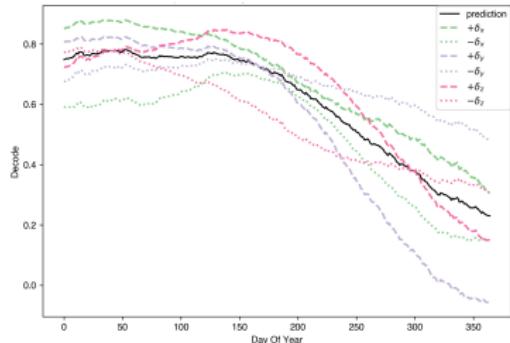


(d) Decodificació Estandardització.

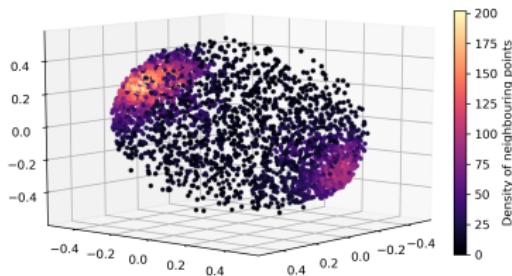
# Sèries Temporals: Espai Latent — $k = 3$



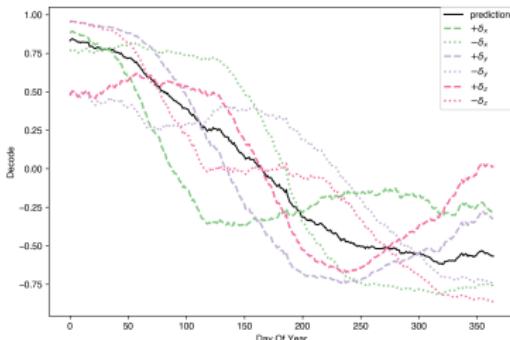
(a) Espai Latent Normalització.



(b) Decodificació Normalització.



(c) Espai Latent Estandardització.



(d) Decodificació Estandardització.

# Continguts

1 Introducció

2 Xarxes Neuronals Directes

3 Aprendentatge Supervisat

4 Aprendentatge No Supervisat

5 Interpretabilitat

6 Incertesa

7 Conclusions

- Interpretabilitat és el grau en què una persona pot comprendre la causa d'una decisió.
- Es desconeix el motiu concret pel qual una xarxa neuronal pren una decisió.
- Aquesta falta de coneixement converteix aquesta tecnologia en una caixa negra.
- Els algorismes d'interpretabilitat intenten resoldre aquest problema.

# Algorismes d'Interpretabilitat

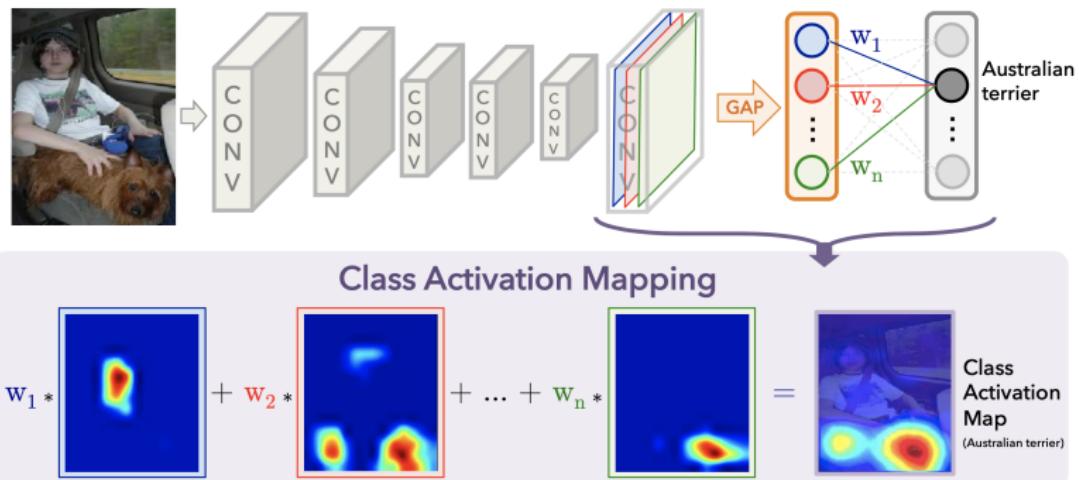
- Interpretabilitat a nivell de **valors d'entrada**, de neurona i de capa.
- *Baseline*: **Punt de referència**, representa informació nul·la.
- Dos axiomes:
  - Invariància respecte de la implementació.
  - Sensibilitat.
- Retropropagació. **Integrated Gradients**:

$$(x_i - x'_i) \times \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

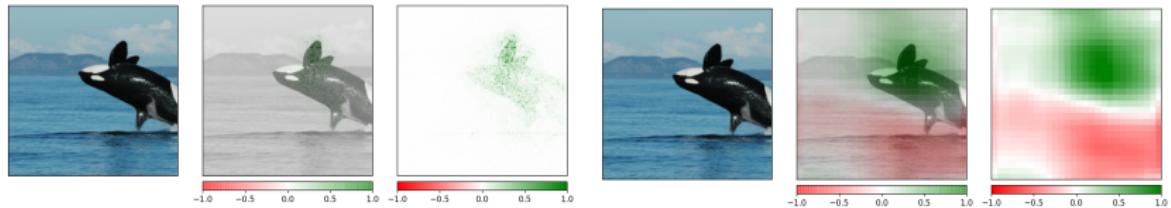
- Modificació de l'input: **Occlusion**.

# Algorismes d'Interpretabilitat

- Atenció: Class Activation Mapping.



# Algorismes d'Interpretabilitat



(a) Integrated Gradients.

(b) Occlusion.



(c) Class Activation Mapping.

# Continguts

1 Introducció

2 Xarxes Neuronals Directes

3 Aprendentatge Supervisat

4 Aprendentatge No Supervisat

5 Interpretabilitat

6 Incertesa

7 Conclusions

- Una predicció **sense un error d'estimació** es pot convertir en **poc útil**.
- Exemple: estimació del número de vendes d'un producte.
- Presentem dos mètodes per afegir error a les prediccions de regressions:
  - Xarxa neuronal que aprèn a **determinar tant la regressió com l'error**.
  - Xarxa neuronal que aprèn a **afegir incertesa a un regressor qualsevol**.

# NormalLoss

- Sigui la distribució normal  $\mathcal{N}(\mu, \sigma^2)$ , la funció de densitat de probabilitat és:

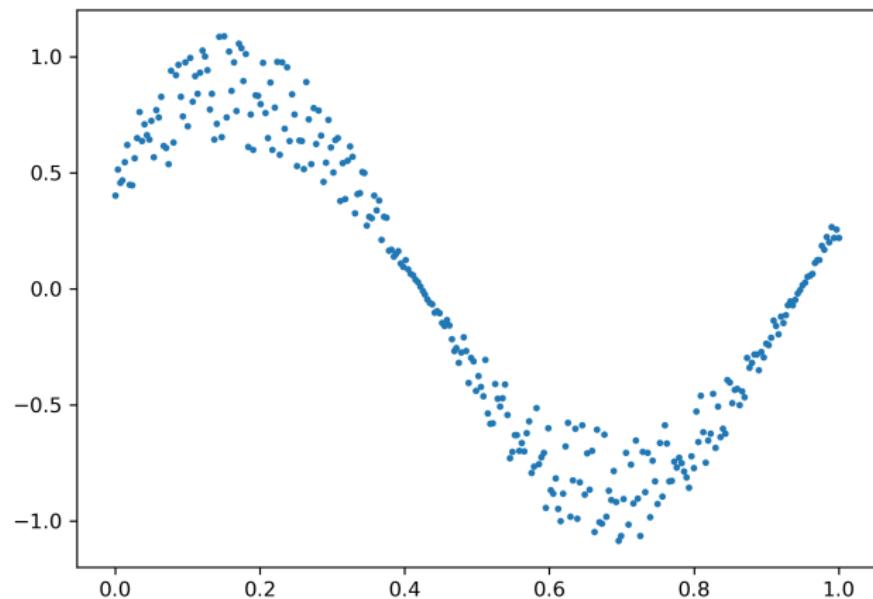
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Indica quant de probable és que un punt es trobi en aquesta distribució.
- Es pot transformar en una funció de pèrdua:

$$J(x, \mu, \sigma) = \frac{1}{N} \sum_{i=1}^N \left( \log(\sigma_i) + \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right)$$

on  $x = \{x_i\}_{i=1,\dots,N}$ ,  $\mu = \{\mu_i\}_{i=1,\dots,N}$  i  $\sigma = \{\sigma_i\}_{i=1,\dots,N}$ .

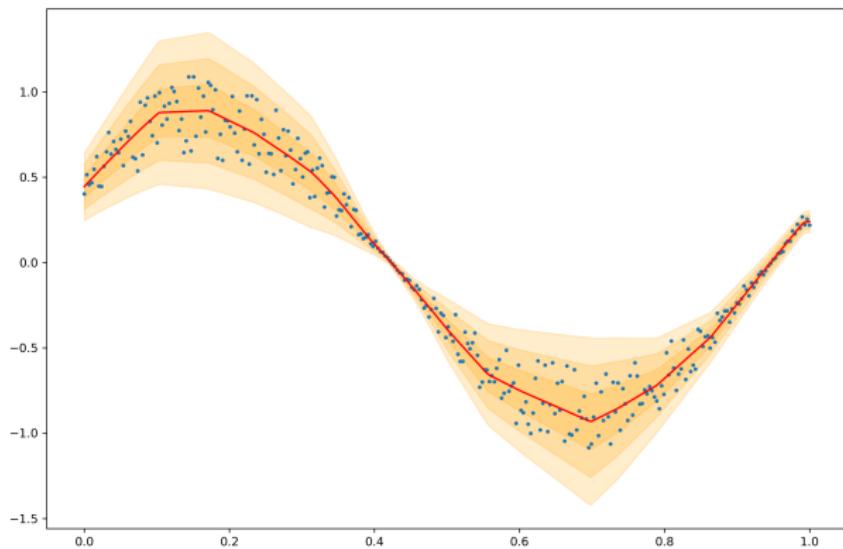
# Problema de Regressió



# Cas 1: Regressió i Error

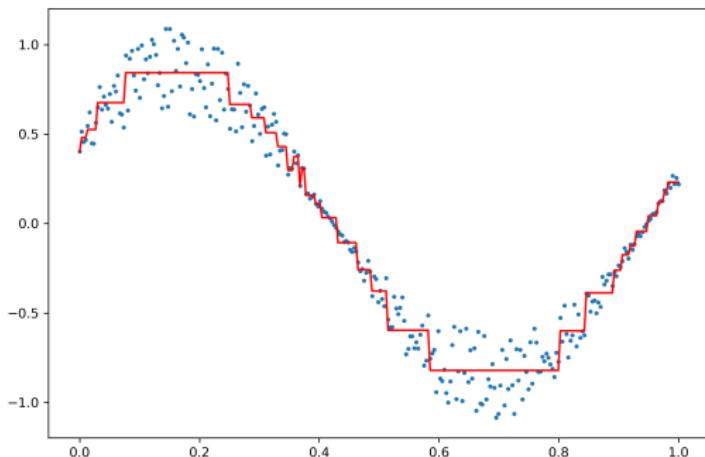
- Xarxa neuronal amb dues capes ocultes de 128 neurones cada una.
- **NormalLoss** com a funció de pèrdua.
- Donat  $x$ , volem determinar la millor distribució  $\mathcal{N}(\mu, \sigma^2)$ .

# Cas 1: Regressió i Error



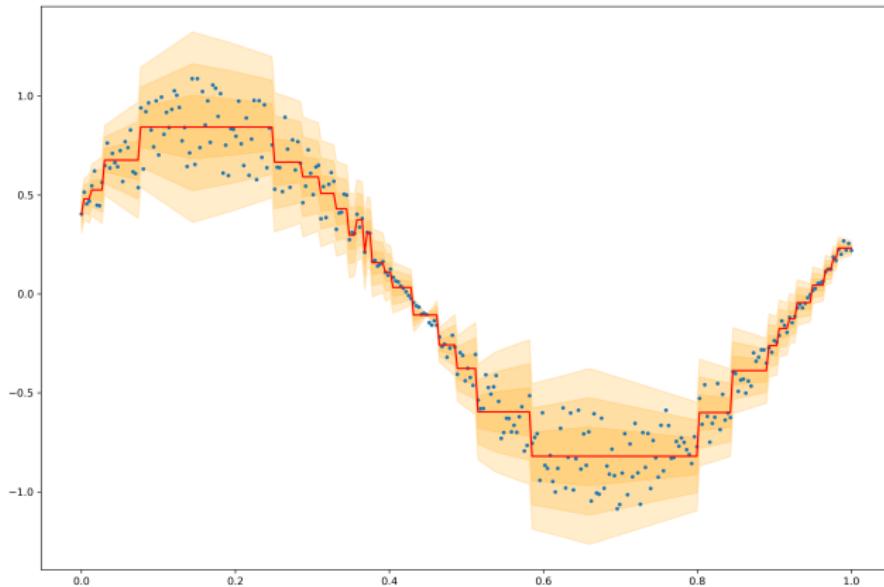
## Cas 2: Regressor

- Considerem un arbre de decisió amb màxima profunditat de 5.
- Tindrem, com a molt,  $2^5 = 32$  valors diferents.

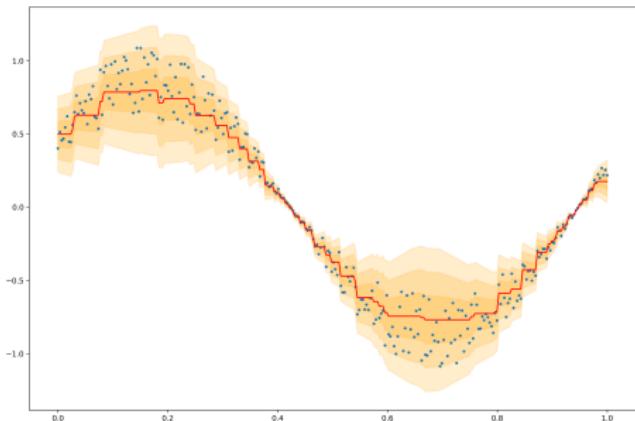


## Cas 2: Xarxa Neuronal

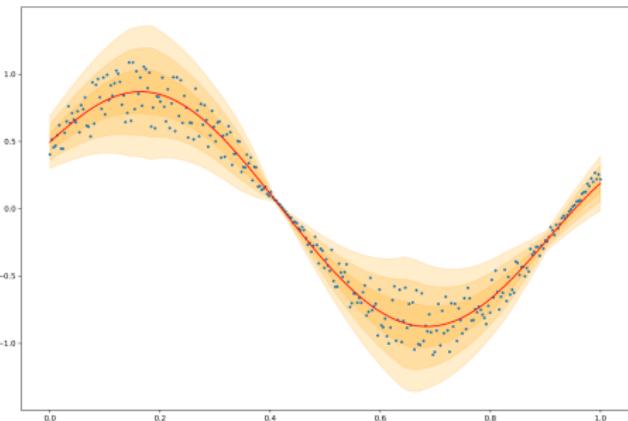
- Xarxa neuronal amb dues capes ocultes de **64 neurones** cada una.
- **NormalLoss** com a funció de pèrdua, però ara els valors de  $\mu$  vindran determinats pel regressor.



## Cas 2: Diferents Regressors



(a) Gradient Boosting amb 25 estimadors.



(b) SVR amb RBF kernel.

# Incertesa: Conclusions

## Uncertainty Estimation for Black-Box Classification Models: A Use Case for Sentiment Analysis

José Mena<sup>1,3</sup>, Axel Brando<sup>2,3</sup>, Oriol Pujo<sup>3</sup>, and Jordi Vitrià<sup>3</sup>

<sup>1</sup> Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain  
jose.mena@eurecat.org

<sup>2</sup> BBVA Analytics Data & Analytics, Madrid, Spain  
axel.brando@bbvadata.com

<sup>3</sup> Universitat de Barcelona, Barcelona, Spain  
{axelbrando,oriol.pujol,jordi.vitria}@ub.edu

(a) Setembre 2019.

## Building uncertainty models on top of black-box predictive APIs

AXEL BRANDO<sup>1,2</sup>, DAMIÀ TORRES<sup>2</sup>, JOSE A. RODRÍGUEZ-SERRANO<sup>2</sup> AND JORDI VITRIÀ.<sup>1</sup>

<sup>1</sup>Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB) (e-mails: axelbrando@ub.edu, jordi.vitria@ub.edu)

<sup>2</sup>BBVA Data and Analytics, Spain, (e-mails: axel.brando@bbvadata.com, joseantonio.rodriguez.serrano@bbvadata.com)

Corresponding author: Axel Brando (e-mail: axelbrando@ub.edu).

(b) Juliol 2020.

# Continguts

- 1 Introducció
- 2 Xarxes Neuronals Directes
- 3 Aprendentatge Supervisat
- 4 Aprendentatge No Supervisat
- 5 Interpretabilitat
- 6 Incertesa
- 7 Conclusions

# Conclusions

- Desenvolupat una descripció matemàtica de les xarxes neuronals directes.
- Analitzats els tipus d'aprenentatge supervisat i no supervisat.
- Aplicacions en el camp de la interpretabilitat i la incertesa.
- Presentats mètodes per a entendre el funcionament i comportament de les xarxes neuronals.
- Futura recerca:
  - Nous mètodes d'interpretabilitat.
  - Desenvolupament del camp de la incertesa.
  - Analitzar diferents arquitectures (GANs, aprenentatge per reforç, etc.)
  - **Aplicacions en camps de recerca concrets.**