

Aprendizaje Profundo

Maestría en Modelación y Optimización de Procesos
CIMAT-Aguascalientes

Dra. Lilí Guadarrama Bustos¹
Dr. Isidro Gómez-Vargas²

¹Cátedra CONACyT CIMAT-Aguascalientes

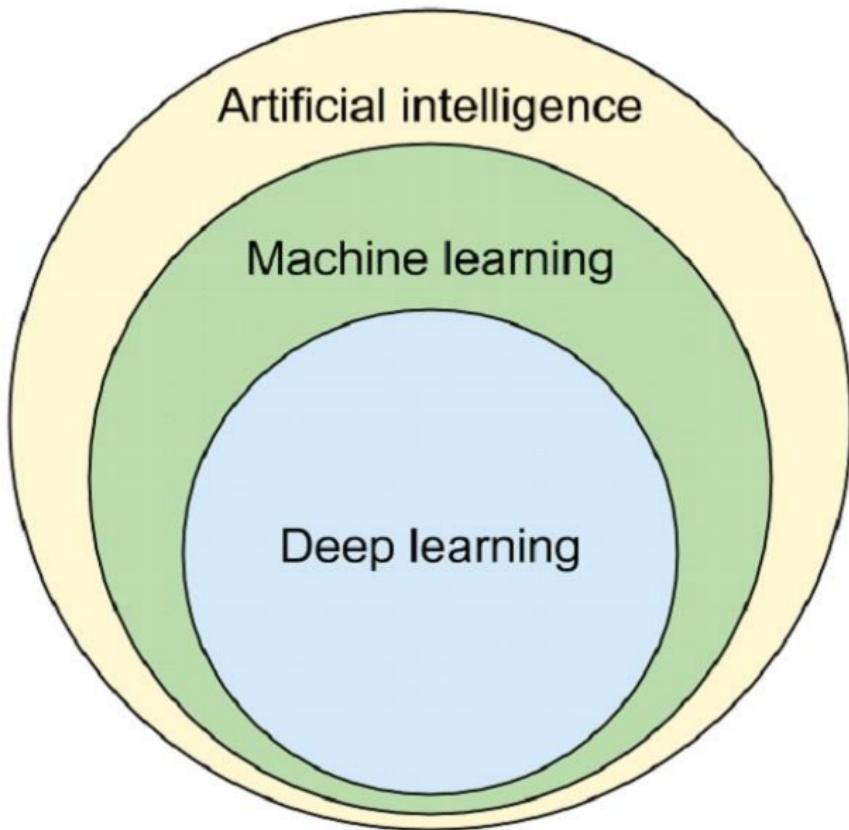
²Investigador posdoctoral en ICF-UNAM

Clase del semestre enero-julio 2022

Contenido

- 1 Bases de Aprendizaje Automático
- 2 Redes de propagación hacia adelante
- 3 Regularización

Aprendizaje profundo



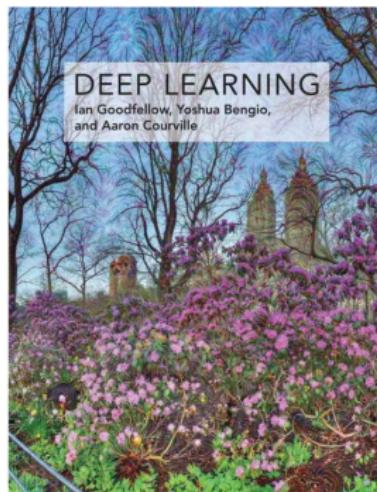
Bibliografía

Bibliografía principal

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Diapositivas:

- https://www.deeplearningbook.org/lecture_slides.html
- <https://github.com/InfolabAI/DeepLearning>



Born	March 5, 1964 (age 57) Paris, France
Citizenship	Canada
Alma mater	McGill University
Known for	Deep learning, neural machine translation, generative adversarial networks, "attention model"®, word embeddings, denoising auto-encoders, neural language models, learning to learn Marie-Victorin Prize (2012) Turing Award (2018) AAAI Fellow (2019)
Awards	
Scientific career	



Born	1985/1986 (age 35–36)
Nationality	American
Alma mater	Stanford University Université de Montréal
Known for	Generative adversarial networks, Adversarial Examples
Scientific career	
Fields	Computer science
Institutions	Apple Inc. Google Brain OpenAI
Thesis	Deep Learning of Representations and its Application to Computer Vision (2014)
Doctoral advisor	Yoshua Bengio
Website	www.iangoodfellow.com



Aaron Courville

Université de Montréal
Dirección de correo verificada de umontreal.ca - Página personal
Machine learning Artificial Intelligence

TÍTULO

CITADO POR

Generative adversarial nets

I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, ...
Advances in neural information processing systems 27

40148

Deep learning

I Goodfellow, Y Bengio, A Courville
Nature

36121

Representation learning: A review and new perspectives

Y Bengio, A Courville, P Vincent
IEEE transactions on pattern analysis and machine intelligence 35 (8), 1770-1828

10333

Temario

- Se abordarán los capítulos 5, 6, 7, 8, 9 y 10.
- Opcionales: capítulos 11 y 12.
- Temas elegidos por el grupo de la parte III del libro (capítulos 13-20).

Temas de la parte 1 del curso

1 Bases de Aprendizaje Automático

- Algoritmos de aprendizaje.
- Capacidad, subajuste y sobreajuste
- Hiperparámetros y conjuntos de validación.
- Estimadores, sesgo y varianza.
- Estimación de Máxima Verosimilitud(MLE).
- Estadística Bayesiana.
- Algoritmos de aprendizaje supervisado.
- Descenso del gradiente estocástico.
- Desafíos actuales Aprendizaje Profundo.

2 Redes de propagación hacia adelante

3 Regularización

Algoritmos de aprendizaje.

“ Un programa de computadora se dice que aprende de la experiencia E con respecto a cierta clase de tareas T y con medida de rendimiento P , si su rendimiento en las tareas en T , medido por P , mejora con la experiencia E . ”

Tom Mitchell, 1997

Glosario

- **Ejemplo.** $x \in \mathbf{R}^n$, colección de n características.
- **Conjunto de datos.** Colección de ejemplos.
- **Características.** Atributos.
- **Matriz de datos.**

Tareas T

- Clasificación
- Regresión
- Transcripción
- Traducción
- Salidas estructuradas
- Detección de anomalías
- Síntesis y muestreo
- Imputing valores perdidos
- Quitar ruido
- Estimación de densidad o probabilidad

La medida de rendimiento P

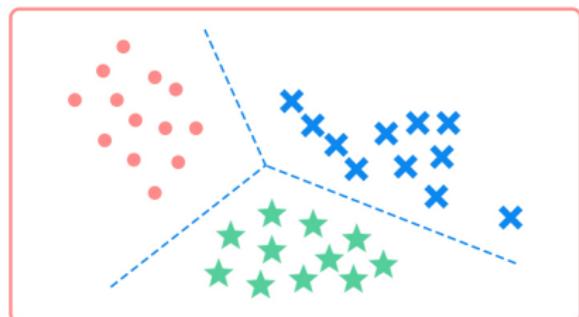
- Exactitud $\frac{TP+TN}{TP+TN+FP+FN}$
- Error cuadrático medio $MSE = \frac{1}{n} \sum_i^n ||\hat{y}_i - y_i||$
- Densidad de probabilidad
- etc

La experiencia, E



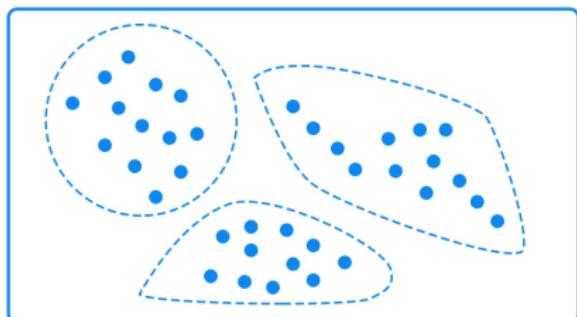
Supervised vs. Unsupervised Learning

Classification



Supervised learning

Clustering

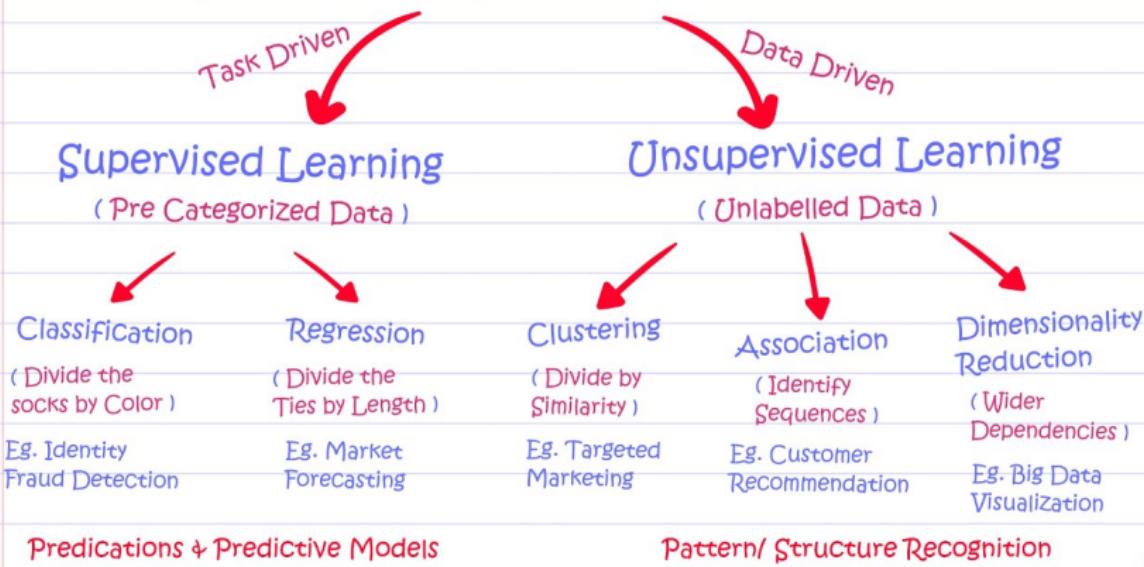


Unsupervised learning

Fuente:

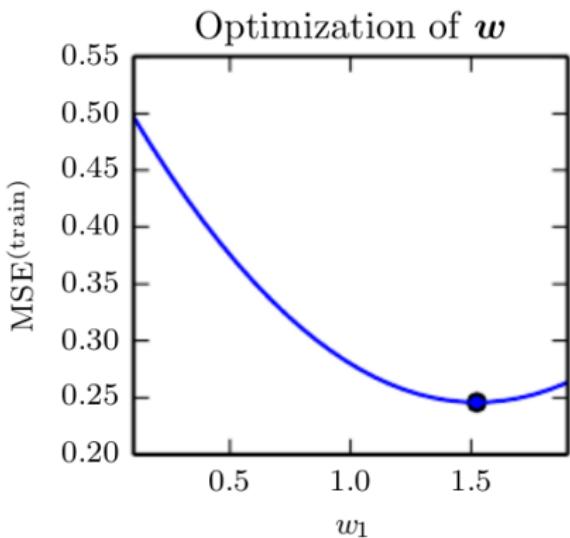
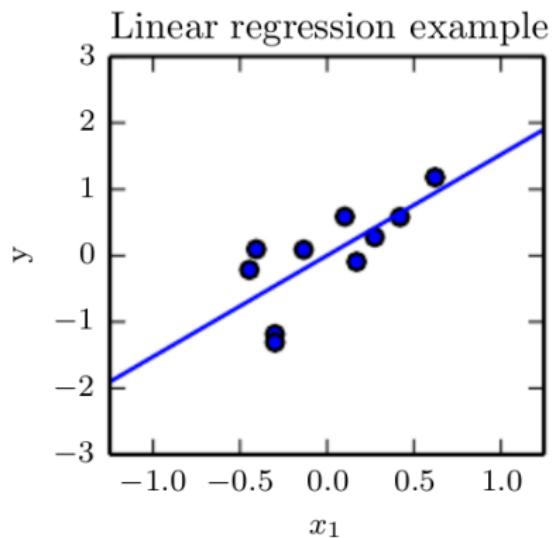
[https://analystprep.com/study-notes/cfa-level-2/quantitative-method/
supervised-machine-learning-unsupervised-machine-learning-deep-learning/
attachment/img_12-4/](https://analystprep.com/study-notes/cfa-level-2/quantitative-method/supervised-machine-learning-unsupervised-machine-learning-deep-learning/attachment/img_12-4/)

Classical Machine Learning



Fuente: <https://medium.com/@recrosoft.io/supervised-vs-unsupervised-learning-key-differences-cdd46206cdcb>

Ejemplo: regresión lineal

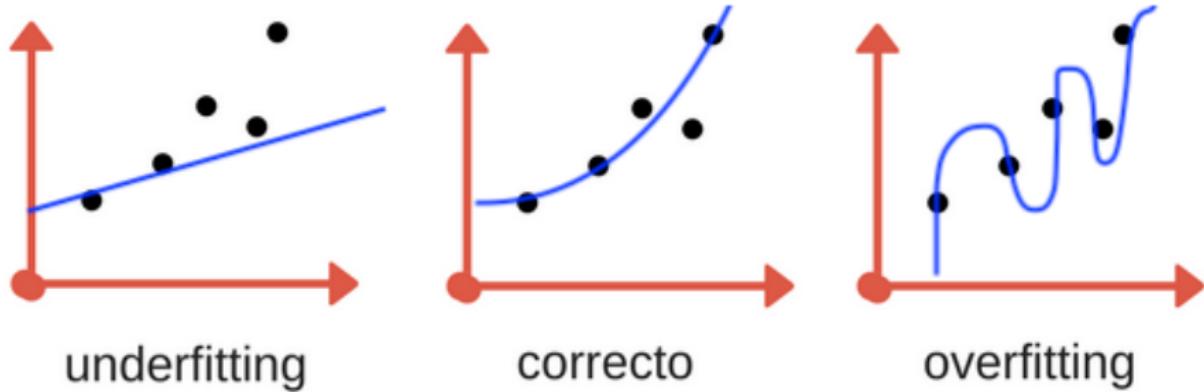


Ver páginas 107-108 del libro.

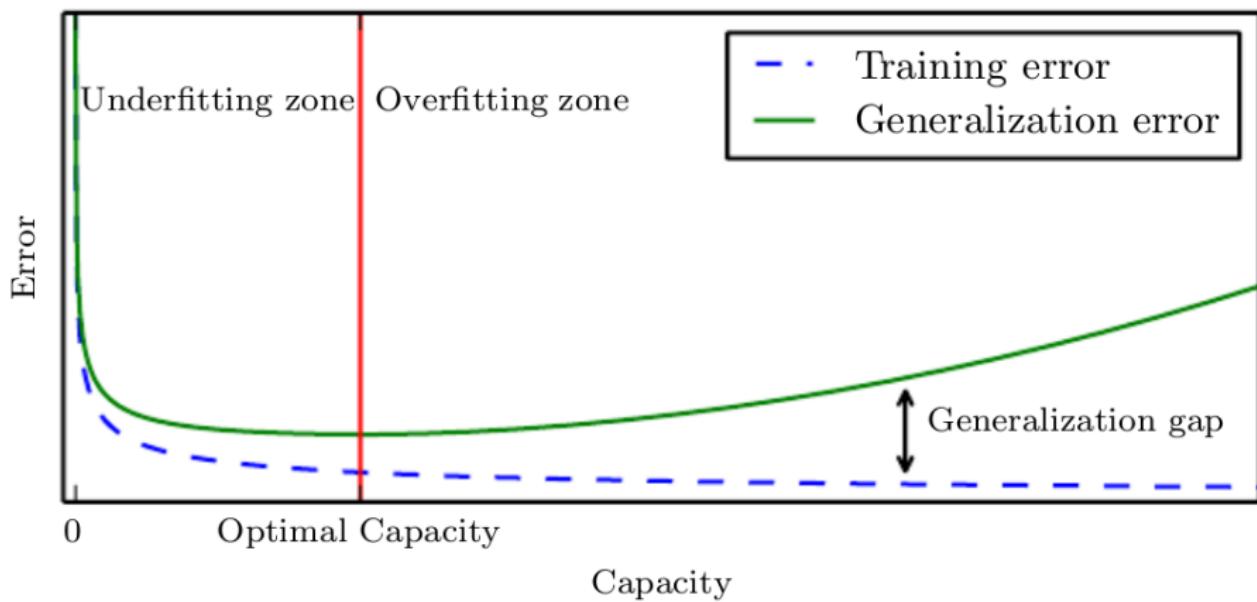
Capacidad, subajuste y sobreajuste

- Algo que separa al aprendizaje automático de la optimización es el uso de un error de generalización o error de prueba (*test error*).
- Hay que minimizar tanto el error de entrenamiento como el error de generalización.
- La brecha entre ambos errores debe ser pequeña.

Capacidad, subajuste y sobreajuste



Capacidad, subajuste y sobreajuste



Capacidad, subajuste y sobreajuste: No free lunch theorem



Capacidad, subajuste y sobreajuste: No free lunch theorem

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality



Simplification is based on assumptions (model bias)



Assumptions fail in certain situations

Roughly speaking:

"No one model works best for all possible situations."

Fuente: <https://analyticsindiamag.com/what-are-the-no-free-lunch-theorems-in-data-science/>

Capacidad, subajuste y sobreajuste: Regularización

Por ejemplo, decaimiento del peso:

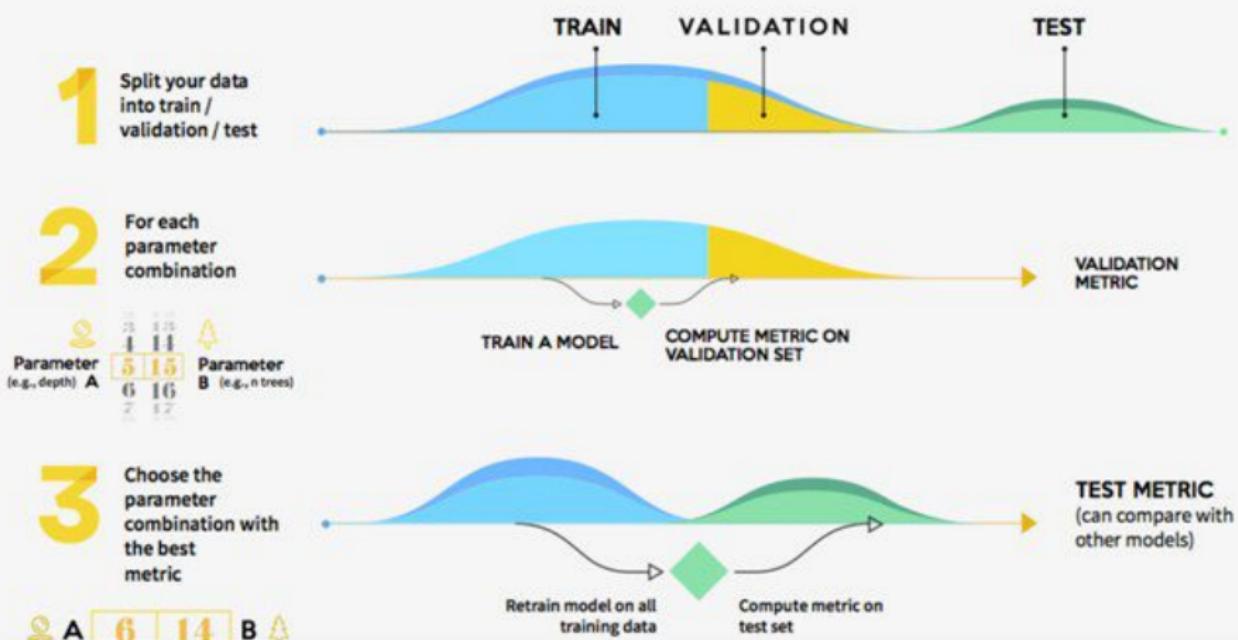
$$J(w) = MSE + \lambda w^T w$$

Def.

Regularización es cualquier mecanismo que ayuda al agoritmo de aprendizaje a recudir su error de generalización.

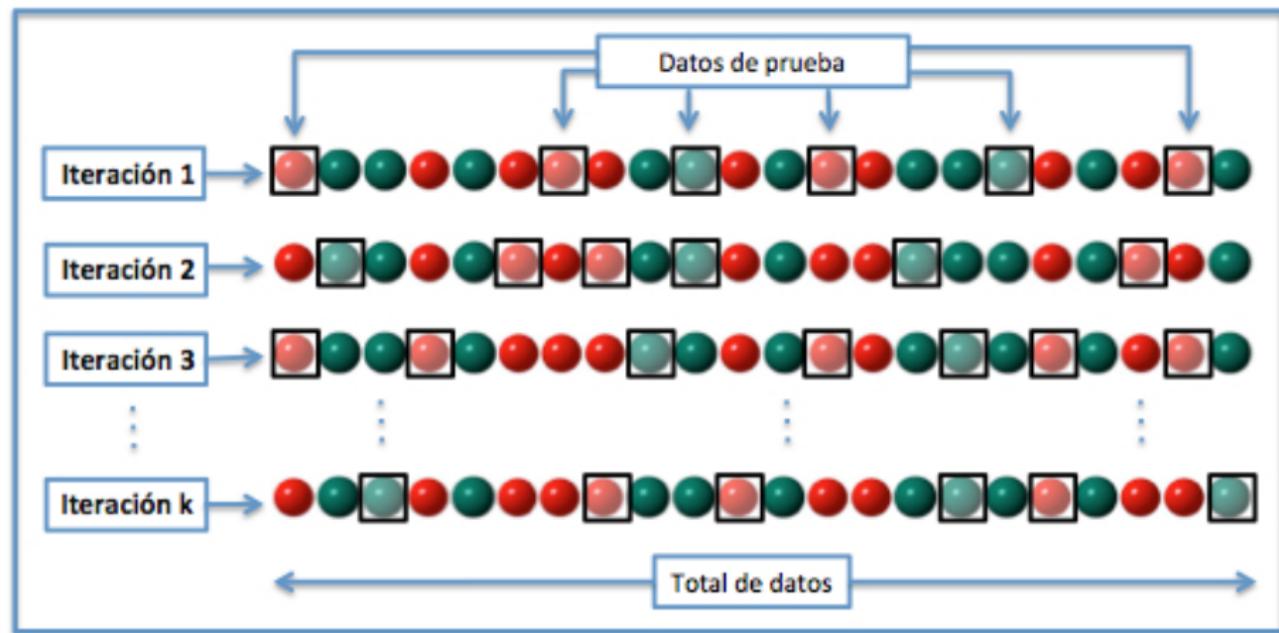
Hiperparámetros y conjuntos de validación.

HOLDOUT STRATEGY



Fuente: [https://www.kdnuggets.com/2017/08/
dataiku-predictive-model-holdout-cross-validation.html](https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html)

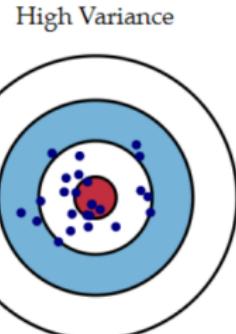
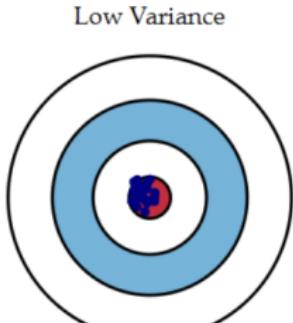
Hiperparámetros y conjuntos de validación: validación cruzada



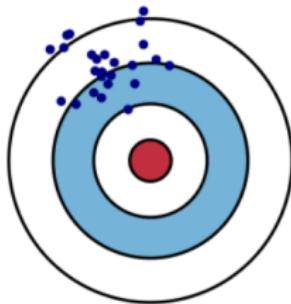
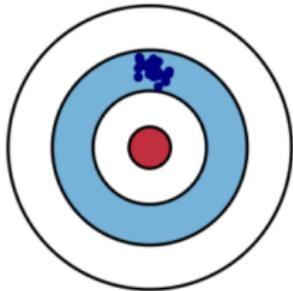
Fuente: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

Estimadores, sesgo y varianza.

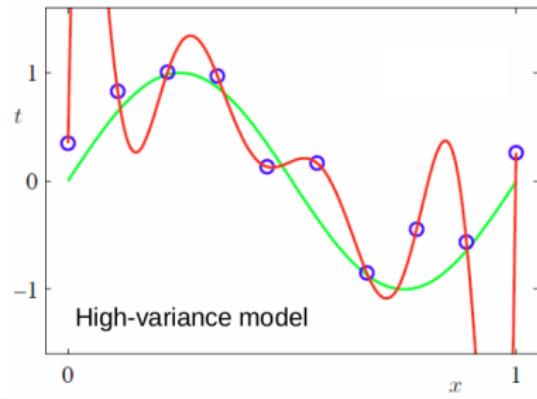
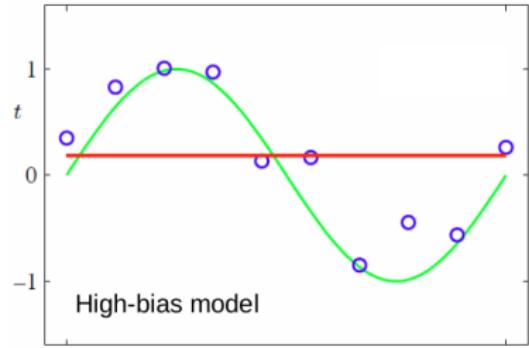
Low Bias



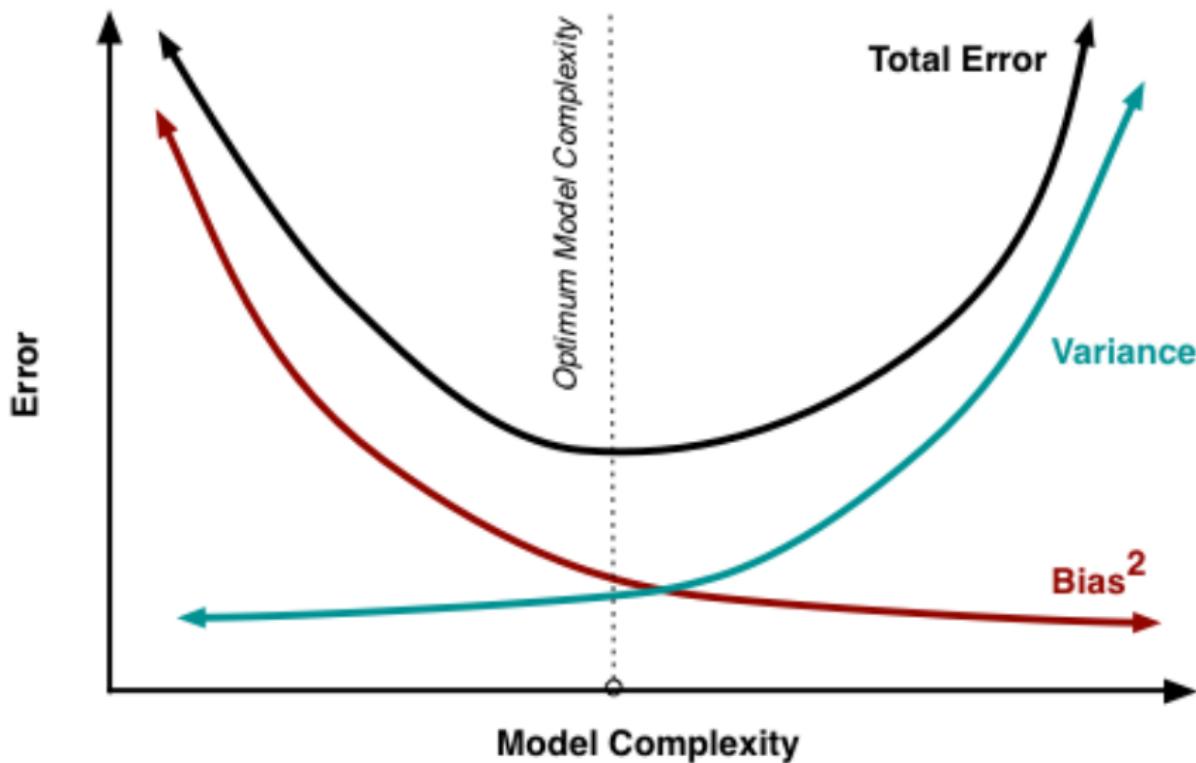
High Bias



Scott Fortmann-Roe, Understanding the Bias-Variance Tradeoff, 2012



Estimadores, sesgo y varianza.



Estimadores, sesgo y varianza.

Visitar los siguientes enlaces:

- <http://scott.fortmann-roe.com/docs/BiasVariance.html>
- <https://ml.berkeley.edu/blog/posts/crash-course/part-4/>
- <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>

Estimación de Máxima Verosimilitud(MLE).

- Estimación del Máximo Likelihood (MLE):

$$\ln \mathcal{L}(D, \theta) = \sum_{i=1}^n \ln f(x_i; \theta),$$

$$\theta_{MLE} = \arg \max(\mathcal{L}(\theta, D))$$

Teorema de Bayes

Considerando funciones de densidad de probabilidad:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}, \quad (1)$$

donde:

$$P(D) = \int_{\mathbb{R}^N} P(D|\theta)P(\theta)d\theta, \quad (2)$$

Estadística Bayesiana.

- Estimación del A Posteriori (MAP) ó estimación de parámetros ó inferencia Bayesiana:

$$\theta_{MAP} = \arg \max(\mathcal{L}(\theta, D)P(\theta))$$

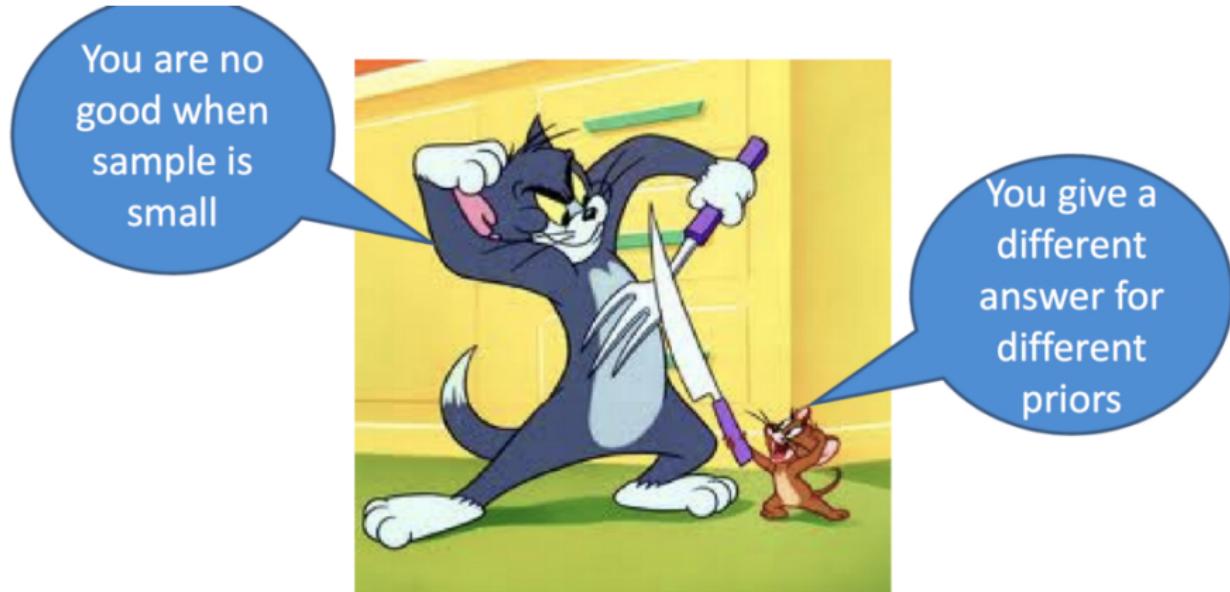
Estadística Bayesiana.

- Estimación del A Posteriori (MAP) ó estimación de parámetros ó inferencia Bayesiana:

$$\theta_{MAP} = \arg \max(\mathcal{L}(\theta, D)P(\theta))$$

- Comparación de modelos (puede ser parte de la inferencia Bayesiana).

Estadística frecuentista vs Bayesiana.



Fuente:<https://laptrinhx.com/maximum-likelihood-estimation-vs-maximum-a-posteriori-2539680111/>

Estadística frecuentista vs Bayesiana.

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

$$\boldsymbol{\theta}^{\text{MLE}} = \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^N p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^N p(\mathbf{x}^{(i)} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

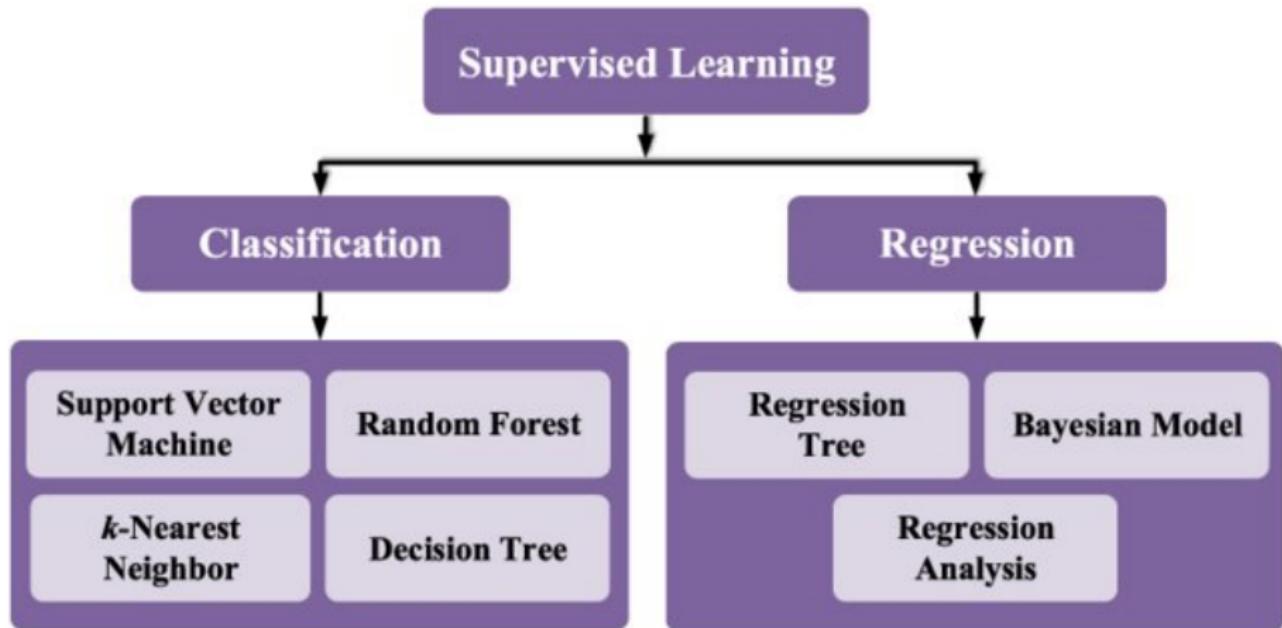


Prior

Fuente: <https://medium.com/@tzjy/>

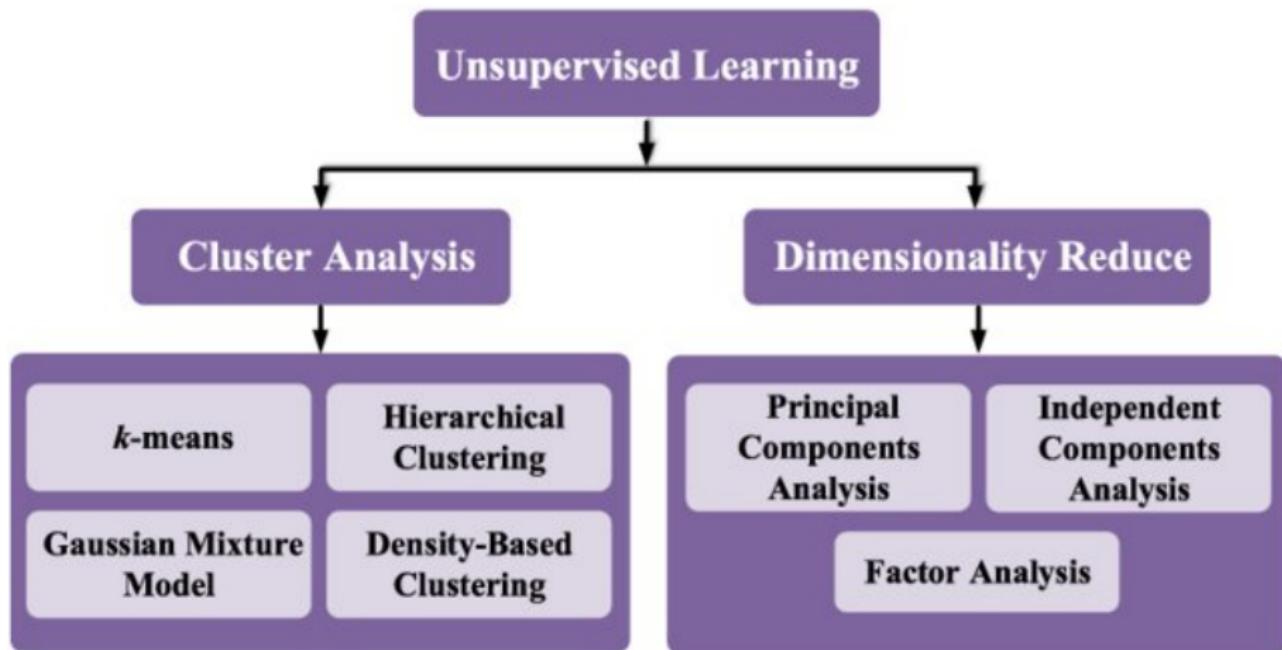
whats-the-difference-between-maximum-likelihood-estimation-mle-and-maximum

Algoritmos de aprendizaje supervisado.



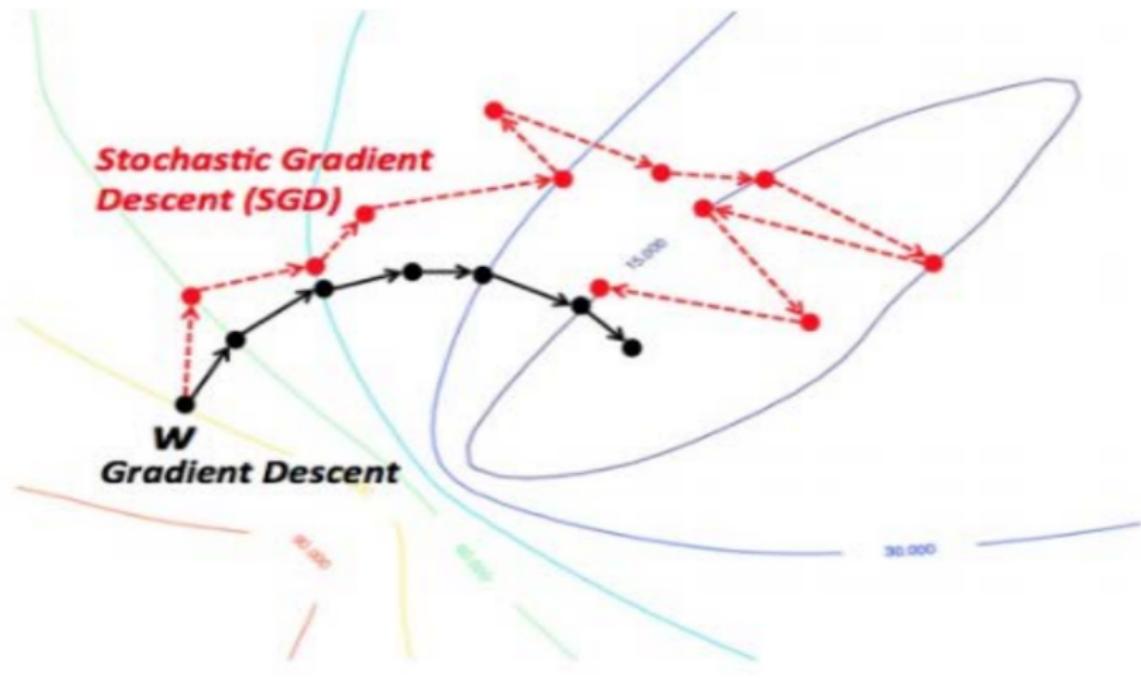
Fuente: arXiv:2003.10146

Algoritmos de aprendizaje no supervisado.



Fuente: arXiv:2003.10146

Descenso del gradiente estocástico.



Fuente: <https://www.slideshare.net/microlife/from-neural-networks-to-deep-learning>

Dra. Lili Guadarrama Bustos Dr. Isidro Gómez

Descenso del gradiente estocástico.

Tarea propuesta

Programar un descenso del gradiente estocástico y minimizar una función con él.

Retos que motivan al Aprendizaje Profundo.

- Curso de la dimensionalidad.
- Local Constancy and Smoothness Regularization.
- Manifold Learning.

Temas de la parte 2 del curso

1 Bases de Aprendizaje Automático

2 Redes de propagación hacia adelante

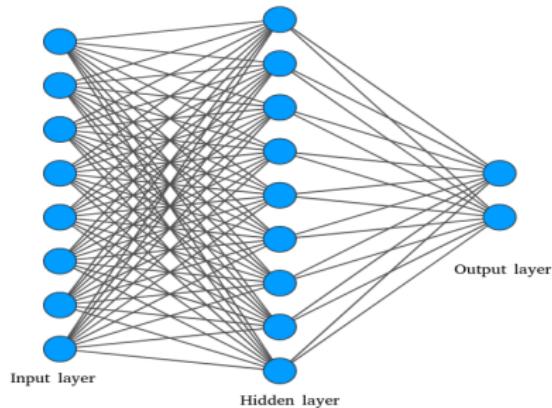
- 6.1 Ejemplo: Aprendiendo XOR
- Gradient-based learning
- Unidades ocultas
- Diseño de arquitectura
- Backpropagation y otros algoritmos de diferenciación

3 Regularización

Introducción al Aprendizaje Profundo

- Inicia la parte II del libro de referencia.
- Nos centraremos en el capítulo 6.

Perceptrón multicapa



Deep feedforward networks /feedforward neural networks

$$f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$$

$f^{(i)}$ es llamada la i -ésima pared de la red. $f(x)$ debe coincidir con $f^*(x)$, cada ejemplo x tiene asociada una etiqueta $y = f^*(x)$

Perceptrón multicapa

Para que obtener un algoritmo no-lineal, hay que emplear una función $\phi(x)$ para modificar las entradas a las funciones.

Truco del kernel (descrito en 5.7.2)

$$w^T x + b = b \sum_{i=1}^m \alpha_i x^T x^{(i)}$$

se puede cambiar $x - > \phi(x)$ y el producto punto por un kernel $k(x, x^{(i)}) = \phi(x) \dot{\phi}(x^{(i)})$. Entonces se pueden hacer predicciones mediante:

$$f(x) = b + \sum_{i=1}^m \alpha_i k(x, x^{(i)})$$

donde la función es no-lineal respecto a x , pero la relación entre $\phi(x)$ y $f(x)$ es lineal. También entre α y $f(x)$ es lineal.

Truco del kernel (descrito en 5.7.2)

La función basada en kernel es exactamente equivalente a preprocesar los datos aplicando $\phi(x)$ a todas las entradas, y luego aprender un modelo lineal en el nuevo espacio transformado.

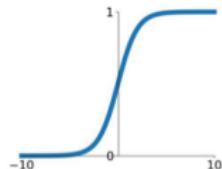
Ventajas:

- ① Permite aprender modelos que son funciones no lineales de x usando técnicas de optimización convexas.
- ② La función kernel, generalmente, permite una implementación que es significativamente más eficiente, en términos computacionales, que construir dos vectores $\phi(x)$ y luego tomar explícitamente sus productos.

Funciones de activación

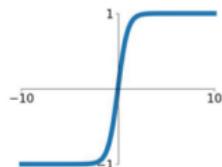
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



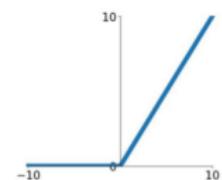
tanh

$$\tanh(x)$$



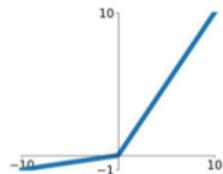
ReLU

$$\max(0, x)$$



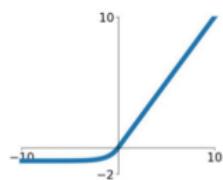
Leaky ReLU

$$\max(0.1x, x)$$



ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Multilayer Feedforward Networks are Universal Approximators

KURT HORNICK

Technische Universität Wien

MAXWELL STINCHCOMBE AND HALBERT WHITE

University of California, San Diego

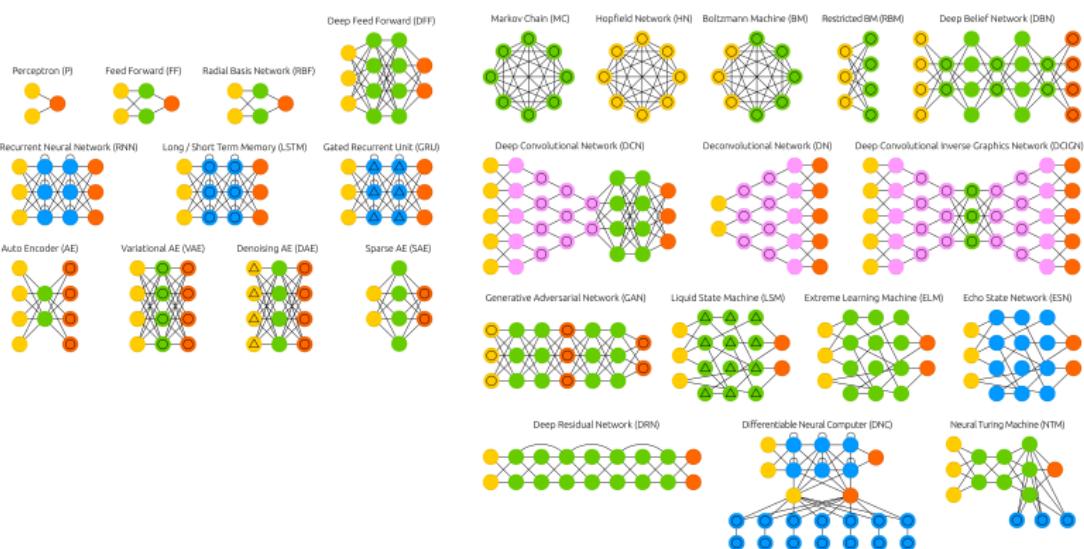
(Received 16 September 1988; revised and accepted 9 March 1989)

Abstract—This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.

Keywords—Feedforward networks, Universal approximation, Mapping networks, Network representation capability, Stone-Weierstrass Theorem, Squashing functions, Sigma-Pi networks, Back-propagation networks.

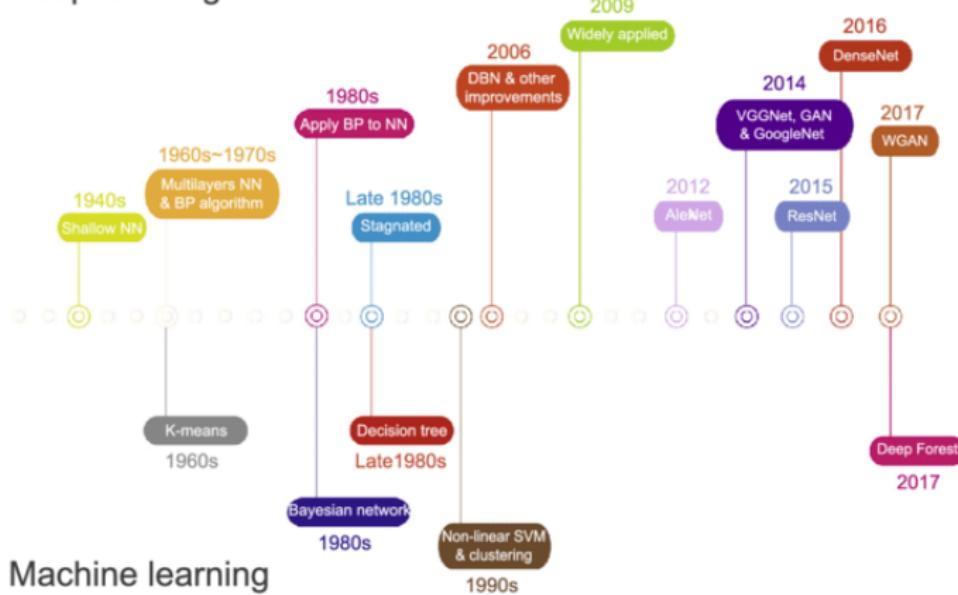
Múltiples arquitecturas

- Input Cell
- Backfied Input Cell
- △ Noisy Input Cell
- Hidden Cell
- Probabilistic Hidden Cell
- ▲ Spiking Hidden Cell
- Capsule Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- △ Gated Memory Cell
- Kernel
- Convolution or Pool



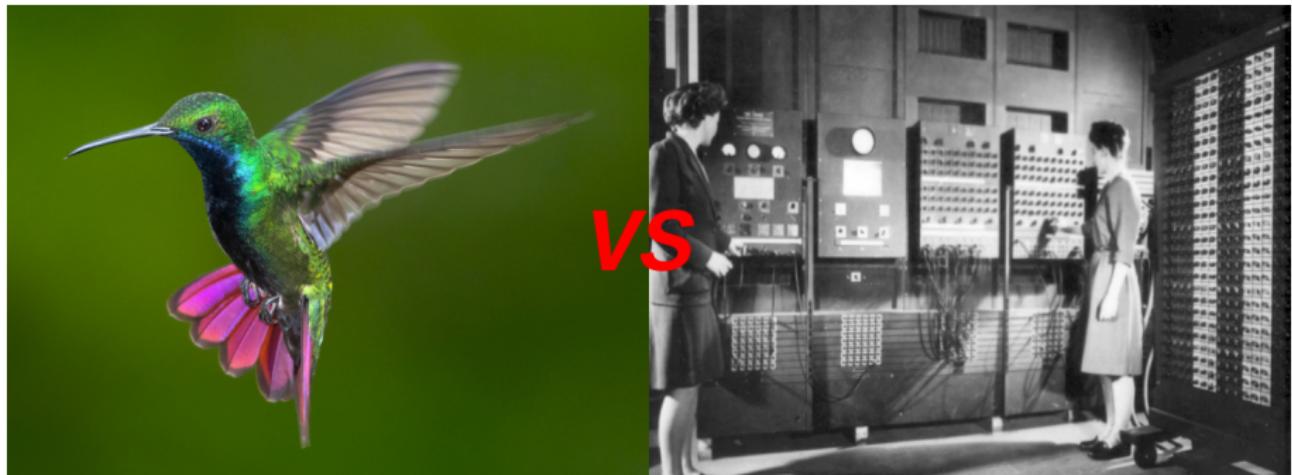
Cronología ML

Deep learning

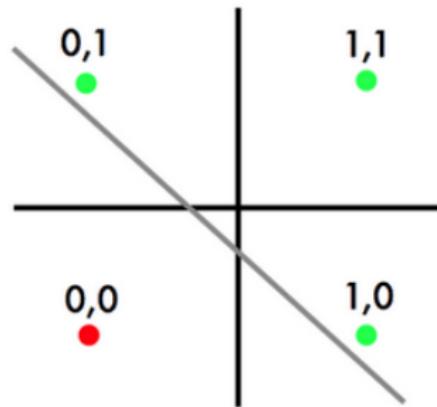


Aprendizaje máquina

1940s



El problema de la compuerta XOR



OR



XOR

Ver notebooks: <https://github.com/igomezv/DLCIMATAGS/tree/main/notebooks/clase%202>

Gradient-based learning

Como preámbulo: repasar secciones 4.3 y 5.9, correspondientes al aprendizaje por medio del descenso del gradiente y al descenso del gradiente estocástico, respectivamente.

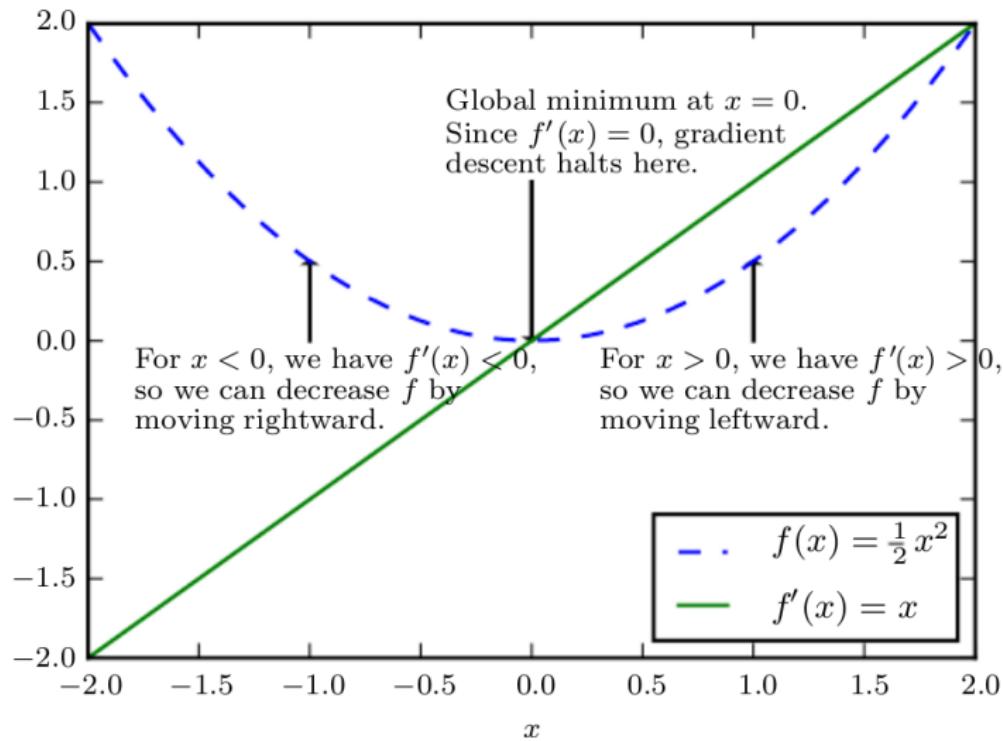
Aprendizaje basado en descenso del gradiente

Optimización:

Maximizar o minimizar una función $f(x)$.

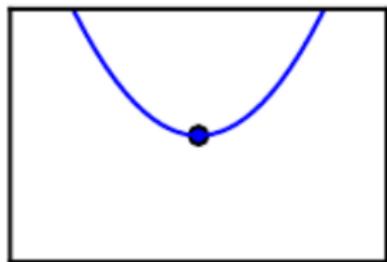
- La función a maximizar o minimizar se conoce como *función objetivo* o *criterio*.
- Cuando esta función se está minimizando, se le suele nombrar *función de costo*, *función de pérdida* o *función de error*.

Aprendizaje basado en descenso del gradiente

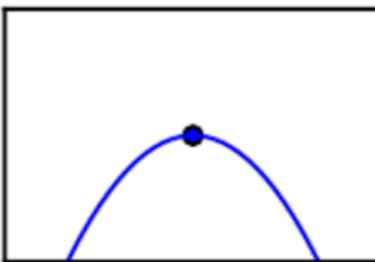


Aprendizaje basado en descenso del gradiente

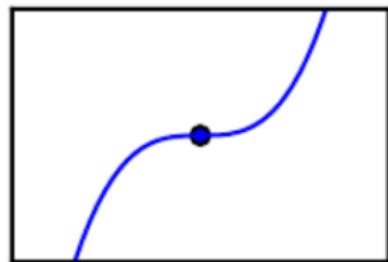
Minimum



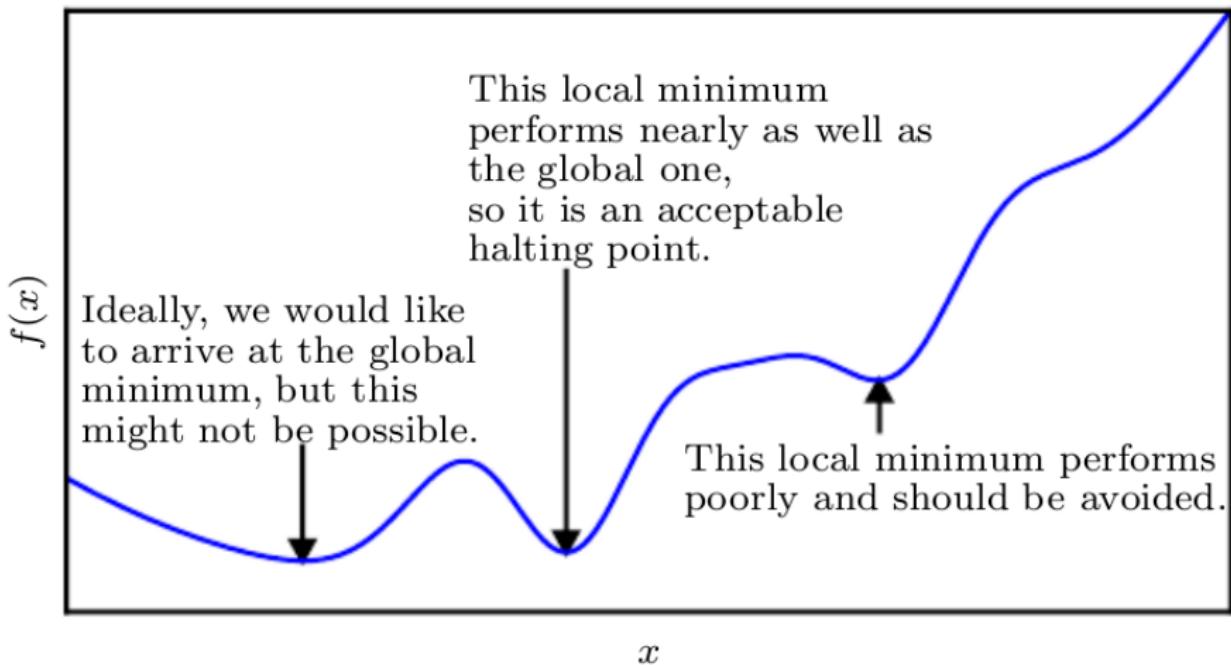
Maximum



Saddle point



Aprendizaje basado en descenso del gradiente



Descenso del gradiente

$$x' = x - \epsilon \nabla_x f(x)$$

Descenso del gradiente estocástico

En el descenso del gradiente, si:

$$\nabla_{\theta} J(\theta) = \frac{1}{m} \sum_{i=1}^m \nabla_{\theta} L(x^{(i)}, y^{(i)}, \theta)$$

entonces el costo computacional del cálculo es $O(m)$.

Descenso del gradiente estocástico

Sea $B = x^{(1)}, \dots, x^{(m')}$ con $x^{(i)}$ extraídos uniformemente del conjunto de entrenamiento. B se conoce como *mini-batch* con tamaño m' , regularmente $1 < m' < \sim 1000$. Entonces:

$$g = \frac{1}{m'} \sum_{i=1}^{m'} \nabla_{\theta} L(x^{(i)}, y^{(i)}, \theta)$$

y

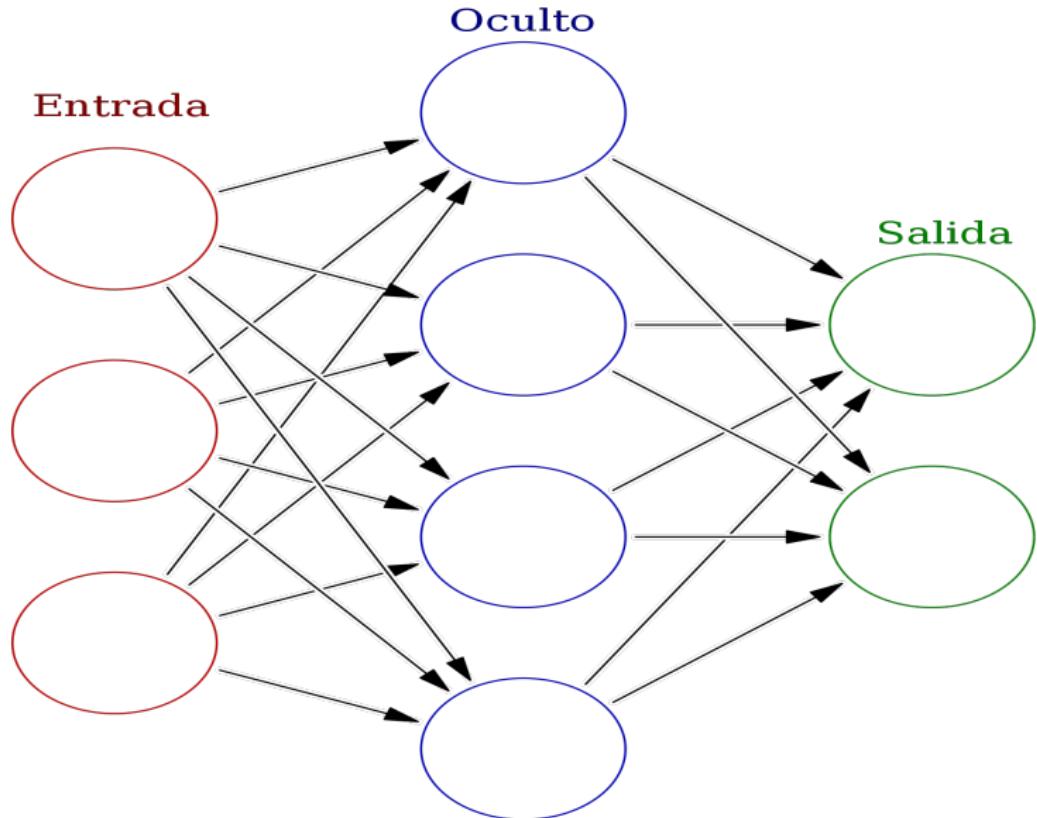
$$\theta \leftarrow \theta - \epsilon g$$

Descenso del gradiente estocástico

En general, el descenso del gradiente se ha considerado lento o poco fiable. En el pasado, la aplicación del descenso de gradiente a problemas de optimización no convexos se consideraba temeraria o sin principios. Hoy en día, sabemos que los modelos de aprendizaje automático funcionan muy bien cuando se entranan con el descenso de gradiente.

El algoritmo de optimización puede no estar garantizado para llegar incluso a un mínimo local en un tiempo razonable, pero a menudo encuentra un valor muy bajo de la función de coste lo suficientemente rápido como para ser útil.

Unidades ocultas



Unidades ocultas

¿Cómo elegir el tipo de unidades ocultas en las capas intermedias?

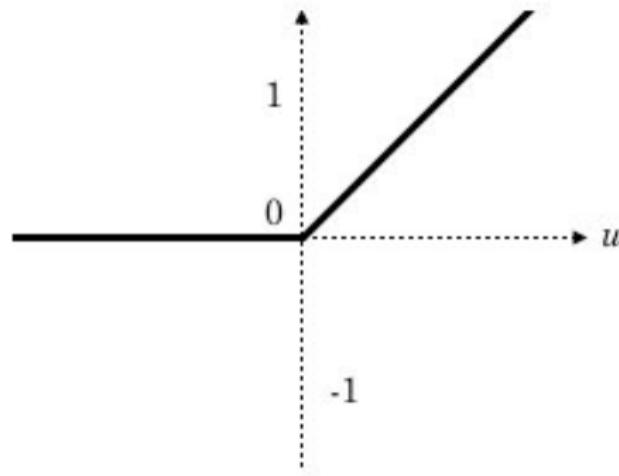
Se trata de un área de investigación muy activa y que todavía no está guiada por principios teóricos.

Recordemos...

$$h = g(W^T x + b)$$

Rectified Linear Unit (ReLU)

$$f(u) = \max(0, u)$$



Generalizaciones de ReLU

$$h_i = g(z, \alpha)_i = \max(0, z_i) + \alpha_i \min(0, z_i)$$

- Absolute value rectification: $\alpha_i = -1 \implies g(z) = |z|$.

Generalizaciones de ReLU

$$h_i = g(z, \alpha)_i = \max(0, z_i) + \alpha_i \min(0, z_i)$$

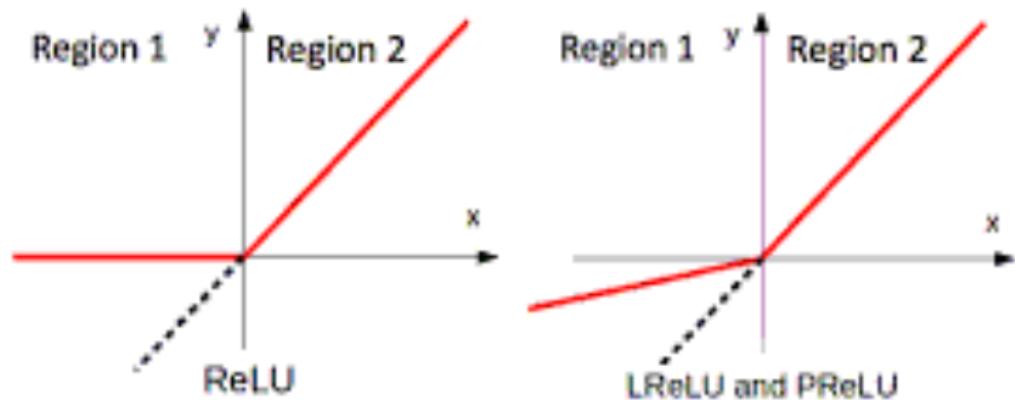
- Absolute value rectification: $\alpha_i = -1 \implies g(z) = |z|$.
- Leaky ReLU: α_i tiene valores cercanos a 0,01.

Generalizaciones de ReLU

$$h_i = g(z, \alpha)_i = \max(0, z_i) + \alpha_i \min(0, z_i)$$

- Absolute value rectification: $\alpha_i = -1 \implies g(z) = |z|$.
- Leaky ReLU: α_i tiene valores cercanos a 0,01.
- Parametric ReLU: α es un parámetro a aprender.

Generalizaciones de ReLU



Unidades Maxout

$$g(z)_i = \max z_j, j \in G^{(i)}$$

donde $G^{(i)}$ es el conjunto de índices dentro de las entradas para el grupo i , $(i-1)k+1, \dots, i_k$. A mayor valor k , unidades maxout pueden aprender a aproximar cualquier función convexa con fidelidad arbitraria.

Unidades Maxout

$$g(z)_i = \max z_j, j \in G^{(i)}$$

donde $G^{(i)}$ es el conjunto de índices dentro de las entradas para el grupo i , $(i-1)k+1, \dots, i_k$. A mayor valor k , unidades maxout pueden aprender a aproximar cualquier función convexa con fidelidad arbitraria.

Ayudan a...

Evitar el olvido catastrófico.

Unidades Maxout

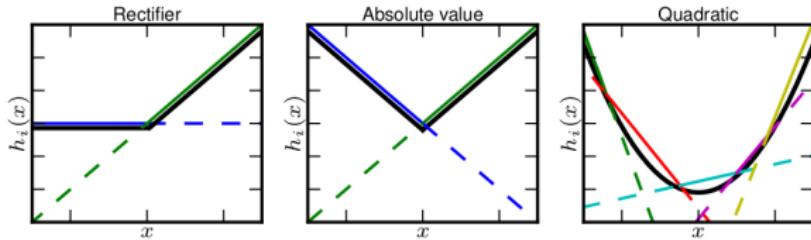


Figure 1. Graphical depiction of how the maxout activation function can implement the rectified linear, absolute value rectifier, and approximate the quadratic activation function. This diagram is 2D and only shows how maxout behaves with a 1D input, but in multiple dimensions a maxout unit can approximate arbitrary convex functions.

Sigmoide logístico y tangente hiperbólica

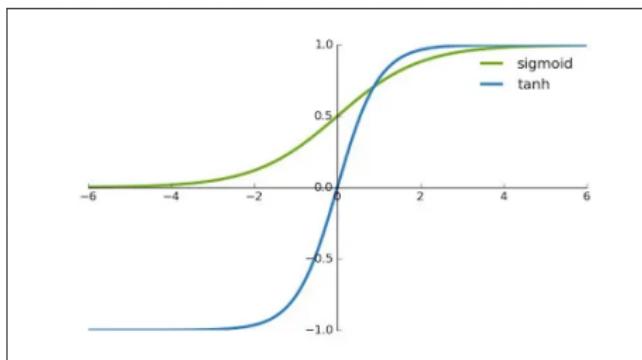
Antes de existir ReLu, la mayoría de las redes neuronales usaban la función sigmoide logístico como función de activación

$$g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

o la función tangente hiperbólica

$$g(z) = \tanh(z)$$

Sigmoide logístico y tangente hiperbólica



Otras unidades ocultas: ventajas de función lineal

Consideremos una red neuronal con algunas paredes con n entradas y p salidas. $h = g(W^T x + b)$. Sean U, V son dos matrices de pesos, si la primera no tiene función de activación se puede sustituir $h = g(V^T U^T x + b)$.

Otras unidades ocultas: ventajas de función lineal

Consideremos una red neuronal con algunas paredes con n entradas y p salidas. $h = g(W^T x + b)$. Sean U, V son dos matrices de pesos, si la primera no tiene función de activación se puede sustituir $h = g(V^T U^T x + b)$. Si U produce q salidas, entonces juntas contienen $(n + p)q$ parámetros, mientras que W tendría np .

Otro tipo de unidades ocultas

- Función de base radial: $h_i = \exp\left(-\frac{1}{\sigma_i^2} \|W_i \cdot i - x\|^2\right)$. Más activa cuando x se acerca a un modelo $W_i \cdot i$. Satura en 0 para la mayoría de los casos, es difícil de optimizar.
- Softplus: $g(a) = \log(1 + e^a)$ (versión suavizada de un rectificador). Recomendada en la última capa, es mejor ReLU en capas ocultas.
- Hard tanh: $g(a) = \max((-1, \min(1, a)))$, similar a tanh y ReLU, pero a diferencia de ReLU, está acotada.

Otro tipo de unidades ocultas

- Función de base radial: $h_i = \exp\left(-\frac{1}{\sigma_i^2} \|W_i(i) - x\|^2\right)$. Más activa cuando x se acerca a un modelo $W_i(i)$. Satura en 0 para la mayoría de los casos, es difícil de optimizar.
- Softplus: $g(a) = \log(1 + e^a)$ (versión suavizada de un rectificador). Recomendada en la última capa, es mejor ReLU en capas ocultas.
- Hard tanh: $g(a) = \max((-1, \min(1, a)))$, similar a tanh y ReLU, pero a diferencia de ReLU, está acotada.

¿Nuevos tipos de unidades ocultas?

Hay gran variedad comparables con los tipos conocidos y son tan comunes que no resultan interesantes.

Diseño de arquitectura

Primera capa:

$$h^{(1)} = g^{(1)}(W^{(1)T}x + b^{(1)}),$$

Diseño de arquitectura

Primera capa:

$$h^{(1)} = g^{(1)}(W^{(1)T}x + b^{(1)}),$$

la segunda capa:

$$h^{(2)} = g^{(2)}(W^{(2)T}x + b^{(2)}),$$

Diseño de arquitectura

Primera capa:

$$h^{(1)} = g^{(1)}(W^{(1)T}x + b^{(1)}),$$

la segunda capa:

$$h^{(2)} = g^{(2)}(W^{(2)T}x + b^{(2)}),$$

y así, sucesivamente.

Diseño de arquitectura

Las redes más profundas suelen ser capaces de utilizar muchas menos unidades por capa y muchos menos parámetros y suelen generalizar en el conjunto de pruebas, pero también suelen ser más difíciles de optimizar. La arquitectura de red ideal para una tarea debe encontrarse mediante la experimentación guiada por el control del error del conjunto de validación.

Diseño de arquitectura

El teorema de aproximación universal establece que una red feedforward con una capa de salida lineal capa de salida lineal y al menos una capa oculta con cualquier función de activación (como la función de activación sigmoide logística) puede aproximar cualquier función de Borel medible de un espacio finito a otro con cualquier cantidad de error no nula deseada, siempre que la red tenga suficientes unidades ocultas.

Diseño de arquitectura

El teorema de aproximación universal establece que una red feedforward con una capa de salida lineal capa de salida lineal y al menos una capa oculta con cualquier función de activación (como la función de activación sigmoide logística) puede aproximar cualquier función de Borel medible de un espacio finito a otro con cualquier cantidad de error no nula deseada, siempre que la red tenga suficientes unidades ocultas.

Función de Borel

Cualquier función continua sobre un subconjunto cerrado y acotado de R^n es Borel medible y, por tanto, puede ser aproximada por una red neuronal.

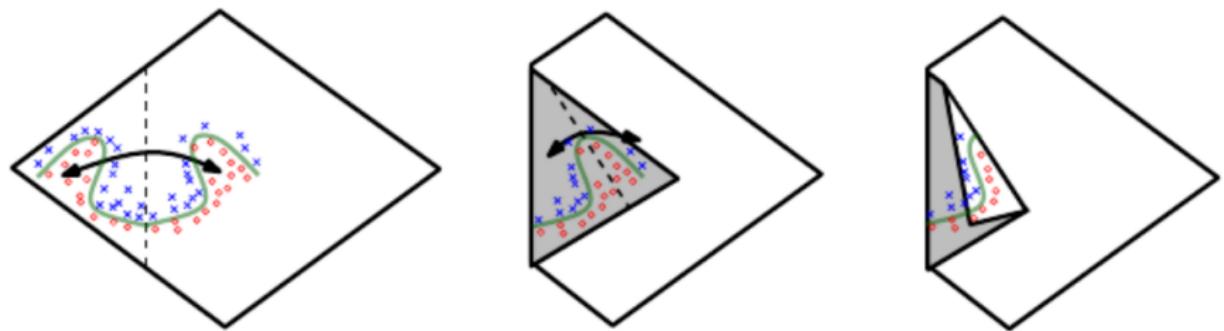
Diseño de arquitectura

Las redes lineales a trozos (que pueden obtenerse de ReLU y sus variantes o las unidades maxout) pueden representar funciones con un número de regiones que es exponencial en la profundidad de la red.

Diseño de arquitectura

Las redes lineales a trozos (que pueden obtenerse de ReLU y sus variantes o las unidades maxout) pueden representar funciones con un número de regiones que es exponencial en la profundidad de la red.

Diseño de arquitectura: más profundo, mejor



Diseño de arquitectura

La elección de un modelo profundo codifica un modelo muy general de que la función que queremos aprender debe implicar la composición de varias funciones más simples.

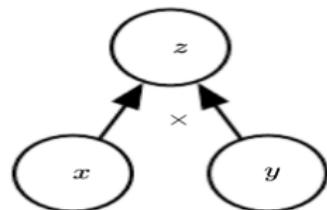
Diseño de arquitectura

<https://playground.tensorflow.org>

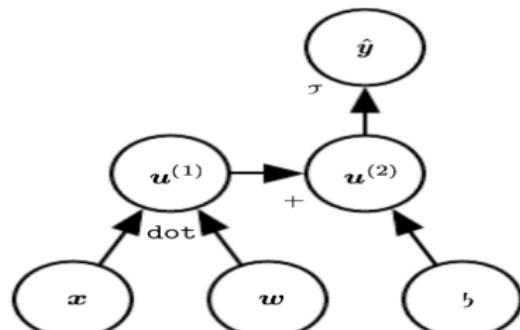
Backpropagation y otros algoritmos de diferenciación

Backpropagation solo se refiere a calcular el gradiente, mientras que otro algoritmo, que podría ser del tipo del descenso del gradiente, se usa para llevar a cabo el aprendizaje de este gradiente.

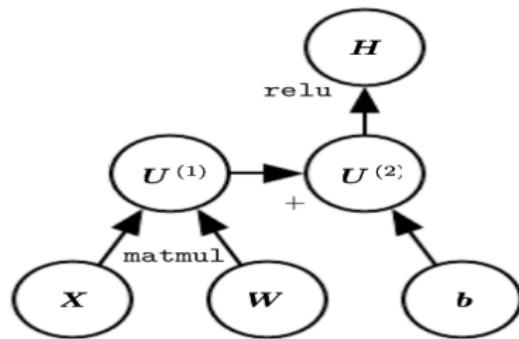
Grafos computacionales



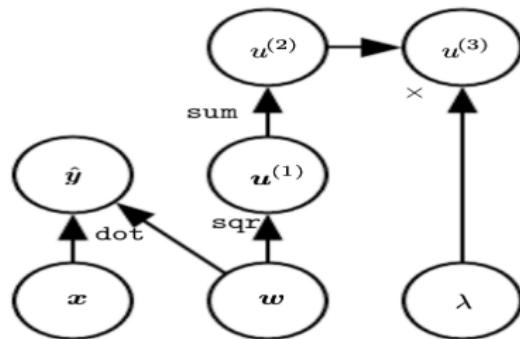
(a)



(b)



(c)



(d)

Backpropagation y regla de la cadena

Sea $y = g(x)$, $z = f(g(x)) = f(y)$

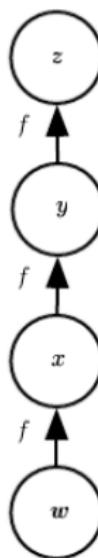
$$\frac{dz}{dx} = \frac{dz}{dy}$$

Sean $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$, si $y = g(x)$ y $z = f(y)$:

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_i}$$

En forma vectorial: $\nabla_x z = (\frac{\partial y}{\partial x})^T \nabla_y z$, donde $(\frac{\partial y}{\partial x}) = J_y \in \mathbb{M}^{n \times m}$.

Subexpresiones repetidas



$$\begin{aligned}\frac{\partial z}{\partial w} &= \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \frac{\partial x}{\partial w} \\ &= f'(y) f'(x) f'(w) \\ &= f'(f(f(w))) f'(f(w)) f'(w)\end{aligned}$$

Propagación hacia adelante general

Algorithm 6.1 A procedure that performs the computations mapping n_i inputs $u^{(1)}$ to $u^{(n_i)}$ to an output $u^{(n)}$. This defines a computational graph where each node computes numerical value $u^{(i)}$ by applying a function $f^{(i)}$ to the set of arguments $\mathbb{A}^{(i)}$ that comprises the values of previous nodes $u^{(j)}$, $j < i$, with $j \in Pa(u^{(i)})$. The input to the computational graph is the vector \mathbf{x} , and is set into the first n_i nodes $u^{(1)}$ to $u^{(n_i)}$. The output of the computational graph is read off the last (output) node $u^{(n)}$.

```
for  $i = 1, \dots, n_i$  do
     $u^{(i)} \leftarrow x_i$ 
end for
for  $i = n_i + 1, \dots, n$  do
     $\mathbb{A}^{(i)} \leftarrow \{u^{(j)} \mid j \in Pa(u^{(i)})\}$ 
     $u^{(i)} \leftarrow f^{(i)}(\mathbb{A}^{(i)})$ 
end for
return  $u^{(n)}$ 
```

Backpropagation simple

Algorithm 6.2 Simplified version of the back-propagation algorithm for computing the derivatives of $u^{(n)}$ with respect to the variables in the graph. This example is intended to further understanding by showing a simplified case where all variables are scalars, and we wish to compute the derivatives with respect to $u^{(1)}, \dots, u^{(n)}$.

Run forward propagation (algorithm 6.1 for this example) to obtain the activations of the network

Initialize `grad_table`, a data structure that will store the derivatives that have been computed. The entry `grad_table[u(i)]` will store the computed value of $\frac{\partial u^{(n)}}{\partial u^{(i)}}$.

```
grad_table[u(n)] ← 1  
for  $j = n - 1$  down to 1 do
```

The next line computes $\frac{\partial u^{(n)}}{\partial u^{(j)}} = \sum_{i:j \in Pa(u^{(i)})} \frac{\partial u^{(n)}}{\partial u^{(i)}} \frac{\partial u^{(i)}}{\partial u^{(j)}}$ using stored values:

```
grad_table[u(j)] ←  $\sum_{i:j \in Pa(u^{(i)})}$  grad_table[u(i)]  $\frac{\partial u^{(i)}}{\partial u^{(j)}}$ 
```

end for

```
return {grad_table[u(i)] |  $i = 1, \dots, n$ }
```

Propagación hacia adelante en un Perceptrón Multicapa

Algorithm 6.3 Forward propagation through a typical deep neural network and the computation of the cost function. The loss $L(\hat{\mathbf{y}}, \mathbf{y})$ depends on the output $\hat{\mathbf{y}}$ and on the target \mathbf{y} (see section 6.2.1.1 for examples of loss functions). To obtain the total cost J , the loss may be added to a regularizer $\Omega(\theta)$, where θ contains all the parameters (weights and biases). Algorithm 6.4 shows how to compute gradients of J with respect to parameters \mathbf{W} and \mathbf{b} . For simplicity, this demonstration uses only a single input example \mathbf{x} . Practical applications should use a minibatch. See section 6.5.7 for a more realistic demonstration.

Require: Network depth, l

Require: $\mathbf{W}^{(i)}, i \in \{1, \dots, l\}$, the weight matrices of the model

Require: $\mathbf{b}^{(i)}, i \in \{1, \dots, l\}$, the bias parameters of the model

Require: \mathbf{x} , the input to process

Require: \mathbf{y} , the target output

$$\mathbf{h}^{(0)} = \mathbf{x}$$

for $k = 1, \dots, l$ **do**

$$\mathbf{a}^{(k)} = \mathbf{b}^{(k)} + \mathbf{W}^{(k)} \mathbf{h}^{(k-1)}$$

$$\mathbf{h}^{(k)} = f(\mathbf{a}^{(k)})$$

end for

$$\hat{\mathbf{y}} = \mathbf{h}^{(l)}$$

$$J = L(\hat{\mathbf{y}}, \mathbf{y}) + \lambda \Omega(\theta)$$

Backpropagation en un Perceptrón Multicapa

Algorithm 6.4 Backward computation for the deep neural network of algorithm 6.3, which uses in addition to the input \mathbf{x} a target \mathbf{y} . This computation yields the gradients on the activations $\mathbf{a}^{(k)}$ for each layer k , starting from the output layer and going backwards to the first hidden layer. From these gradients, which can be interpreted as an indication of how each layer's output should change to reduce error, one can obtain the gradient on the parameters of each layer. The gradients on weights and biases can be immediately used as part of a stochastic gradient update (performing the update right after the gradients have been computed) or used with other gradient-based optimization methods.

After the forward computation, compute the gradient on the output layer:

$\mathbf{g} \leftarrow \nabla_{\hat{\mathbf{y}}} J = \nabla_{\hat{\mathbf{y}}} L(\hat{\mathbf{y}}, \mathbf{y})$
for $k = l, l - 1, \dots, 1$ **do**

Convert the gradient on the layer's output into the gradient into the pre-nonlinearity activation (element-wise multiplication if f is element-wise):

$$\mathbf{g} \leftarrow \nabla_{\mathbf{a}^{(k)}} J = \mathbf{g} \odot f'(\mathbf{a}^{(k)})$$

Compute gradients on weights and biases (including the regularization term, where needed):

$$\nabla_{\mathbf{b}^{(k)}} J = \mathbf{g} + \lambda \nabla_{\mathbf{b}^{(k)}} \Omega(\theta)$$

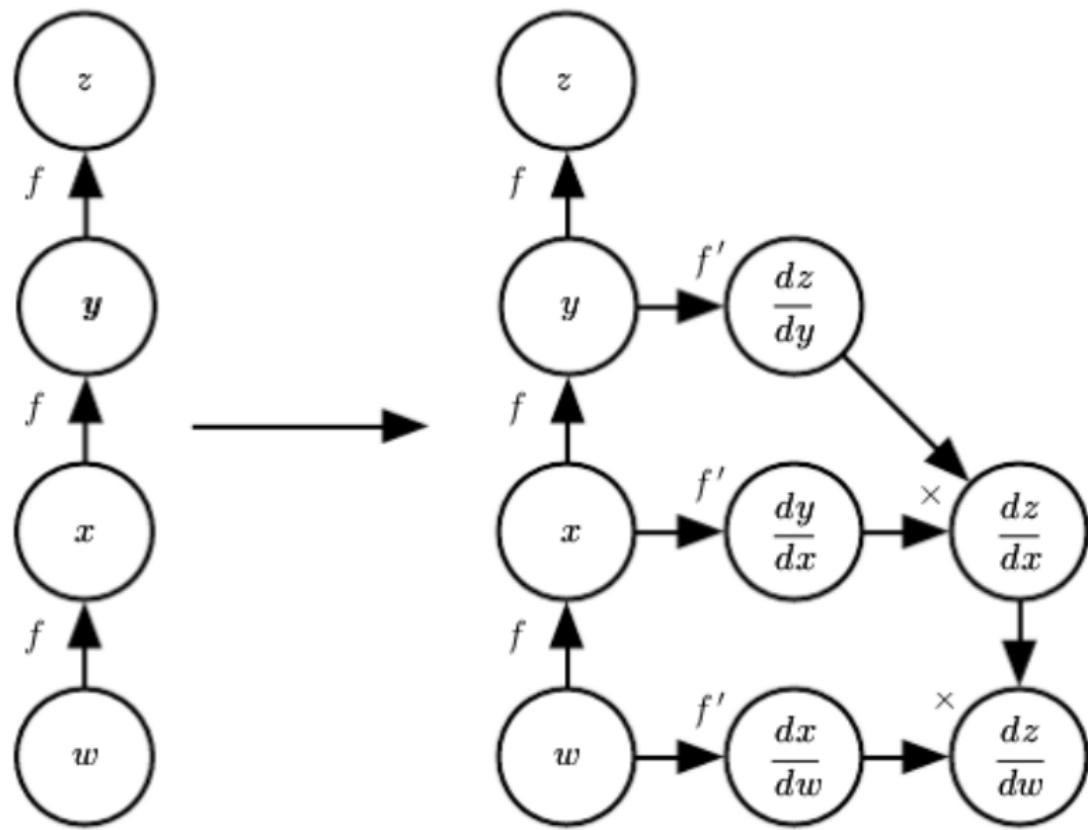
$$\nabla_{\mathbf{W}^{(k)}} J = \mathbf{g} \mathbf{h}^{(k-1)\top} + \lambda \nabla_{\mathbf{W}^{(k)}} \Omega(\theta)$$

Propagate the gradients w.r.t. the next lower-level hidden layer's activations:

$$\mathbf{g} \leftarrow \nabla_{\mathbf{h}^{(k-1)}} J = \mathbf{W}^{(k)\top} \mathbf{g}$$

end for

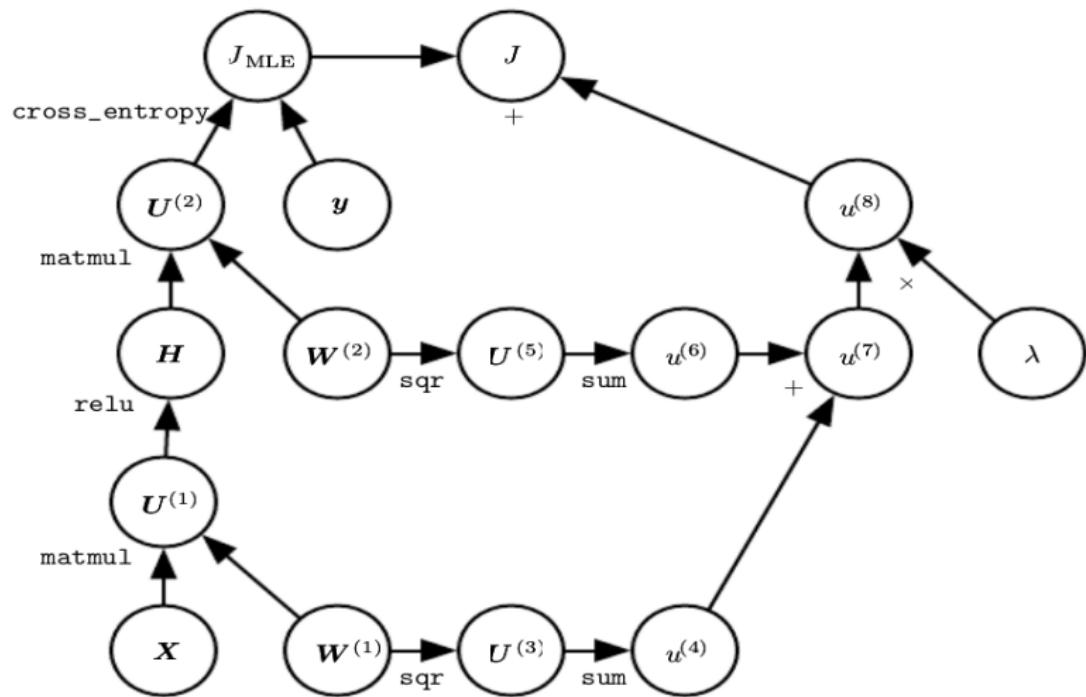
Símbolo a símbolo / símbolo a número



Ejemplo: Representación de un MLP

1 capa oculta, relu, weight decay, cross entropy:

$$J = J_{\text{MLE}} + \lambda \left(\sum_{i,j} (W_{i,j}^{(1)})^2 + \sum_{i,j} (W_{i,j}^{(2)})^2 \right)$$



Temas de la parte 3 del curso

- 1 Bases de Aprendizaje Automático
- 2 Redes de propagación hacia adelante
- 3 Regularización
 - Regularización
 - Parameter norm penalties
 - Inferencia bayesiana
 - 7.3-7.4
 - Robustez del ruido (noise robustness)
 - Aprendizaje semi-supervisado
 - Aprendizaje multi-tarea
 - Detención temprana (early stopping)

Regularización para AP

La regularización es cualquier modificación que se realiza en el algoritmo de aprendizaje para reducir el error de generalización.

mayor sesgo \leftrightarrow menor varianza

Idea:

El mejor modelo ajuste es aquel modelo general regularizado
adecuadamente

Parameter norm penalties

Sea $J = J(\theta; X, y)$ la función objetivo, denotamos como \tilde{J} a la función objetivo regularizada definida como

$$\tilde{J}((\theta; X, y) = J(\theta; X, y) + \alpha \Gamma(\theta)$$

- $\alpha \geq 0$
- Γ norm penalty

En AP se utilizaran normas que penalicen solamente los pesos en cada capa dejando los sesgos sin regularizar (evitando que aumente la varianza & underfitting)

Parámetro de regularización L^2 (weight decay)

Sea

$$\Omega(\theta) = \frac{1}{2} \|\omega\|_2^2,$$

de esta forma

$$\tilde{J}((\omega; X, y) = J(\omega; X, y) + \frac{\alpha}{2} \omega^T \omega.$$

Como gradiente con respecto a los pesos es

$$\nabla_{\omega} \tilde{J}(\omega; X, y) = \nabla_{\omega} J(\omega; X, y) + \alpha \omega,$$

entonces actualizamos los pesos de la siguiente forma,

$$\omega \leftarrow (1 - \epsilon \alpha) \omega - \epsilon \nabla_{\omega} J(\omega; X, y)$$

Inferencia bayesiana

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8699938/>

7.2-7.4

7.5. Robustez del ruido (noise robustness)

- Se analizó previamente la adición de ruido como estrategia para aumentar datos.

7.5. Robustez del ruido (noise robustness)

- Se analizó previamente la adición de ruido como estrategia para aumentar datos.
- La adición de ruido con una varianza infinitesimal en las entradas de los modelos es equivalente a imponer una penalización sobre las normas de los pesos (Bishop 1995a, Bishop 1995b)

7.5. Robustez del ruido (noise robustness)

- Se analizó previamente la adición de ruido como estrategia para aumentar datos.
- La adición de ruido con una varianza infinitesimal en las entradas de los modelos es equivalente a imponer una penalización sobre las normas de los pesos (Bishop 1995a, Bishop 1995b)
- La adición de ruido en las capas ocultas puede ser mucho más poderoso que simplemente compactar los parámetros. El caso más general de esto es el *dropout*.

7.5. Robustez del ruido (noise robustness)

También se suele agregar ruido a los pesos para regularizar modelos.

7.5. Robustez del ruido (noise robustness)

También se suele agregar ruido a los pesos para regularizar modelos.

- Se usa principalmente en redes recurrentes.

7.5. Robustez del ruido (noise robustness)

También se suele agregar ruido a los pesos para regularizar modelos.

- Se usa principalmente en redes recurrentes.
- Esto se puede interpretar como una implementación estocástica de inferencia Bayesiana sobre los pesos.

7.5. Robustez del ruido (noise robustness)

También se suele agregar ruido a los pesos para regularizar modelos.

- Se usa principalmente en redes recurrentes.
- Esto se puede interpretar como una implementación estocástica de inferencia Bayesiana sobre los pesos.
- El tratamiento Bayesiano del aprendizaje considera que los pesos tienen incertidumbre mediante una distribución de probabilidad.

7.5. Robustez del ruido (noise robustness)

También se suele agregar ruido a los pesos para regularizar modelos.

- Se usa principalmente en redes recurrentes.
- Esto se puede interpretar como una implementación estocástica de inferencia Bayesiana sobre los pesos.
- El tratamiento Bayesiano del aprendizaje considera que los pesos tienen incertidumbre mediante una distribución de probabilidad.
- Añadir ruido a los pesos es una forma estocástica de manifestar dicha incertidumbre.

7.5. Ruido como interpretación Bayesiana

1424

IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 7, NO. 6, NOVEMBER 1996

An Analysis of Noise in Recurrent Neural Networks: Convergence and Generalization

Kam-Chuen Jim, *Member, IEEE*, C. Lee Giles, *Senior Member, IEEE*, and Bill G. Horne, *Member, IEEE*

Practical Variational Inference for Neural Networks

Alex Graves

Department of Computer Science
University of Toronto, Canada
graves@cs.toronto.edu

7.5. Robustez del ruido (noise robustness)

El ruido aplicado a los pesos también puede ser interpretado como una forma de regularización tradicional (bajo ciertas suposiciones), garantizando estabilidad a la función objetivo.

7.5. Robustez del ruido (noise robustness)

El ruido aplicado a los pesos también puede ser interpretado como una forma de regularización tradicional (bajo ciertas suposiciones), garantizando estabilidad a la función objetivo.

Sean \hat{y} predicciones y y valores reales:

$$J = \mathbb{E}_{p(x,y)}[(\hat{y}(x) - y)^2]$$

con m elementos etiquetados (x_i, y_i) del conjunto de entrenamiento.

7.5. Robustez del ruido (noise robustness)

El ruido aplicado a los pesos también puede ser interpretado como una forma de regularización tradicional (bajo ciertas suposiciones), garantizando estabilidad a la función objetivo.

Sean \hat{y} predicciones y y valores reales:

$$J = \mathbb{E}_{p(x,y)}[(\hat{y}(x) - y)^2]$$

con m elementos etiquetados (x_i, y_i) del conjunto de entrenamiento.

Se puede asumir que cada entrada incluye una ruido (perturbación) aleatorio en los pesos: $\epsilon_W \sim N(\epsilon; 0, \eta I)$. En un MLP, consideremos el modelo perturbado como $\hat{y}_{\epsilon_W}(x)$.

7.5. Robustez del ruido (noise robustness)

$$\tilde{J}_W = \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}(x) - y)^2]$$

7.5. Robustez del ruido (noise robustness)

$$\begin{aligned}\tilde{J}_W &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}(x) - y)^2] \\ &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}^2(x) - 2y\hat{y}_{\epsilon_W}(x) + y^2)]\end{aligned}$$

7.5. Robustez del ruido (noise robustness)

$$\begin{aligned}\tilde{J}_W &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}(x) - y)^2] \\ &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}^2(x) - 2y\hat{y}_{\epsilon_W}(x) + y^2)]\end{aligned}$$

para pequeñas varianzas η , la minimización de J con ruido en los pesos (covarianza ηI) es equivalente a minimizar J con un término adicional de regularización:

$$\eta \mathbb{E}_{p(x,y)} [\|\nabla_w \hat{y}(x)\|^2]$$

7.5. Robustez del ruido (noise robustness)

$$\begin{aligned}\tilde{J}_W &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}(x) - y)^2] \\ &= \mathbb{E}_{p(x,y,\epsilon_W)} [(\hat{y}_{\epsilon_W}^2(x) - 2y\hat{y}_{\epsilon_W}(x) + y^2)]\end{aligned}$$

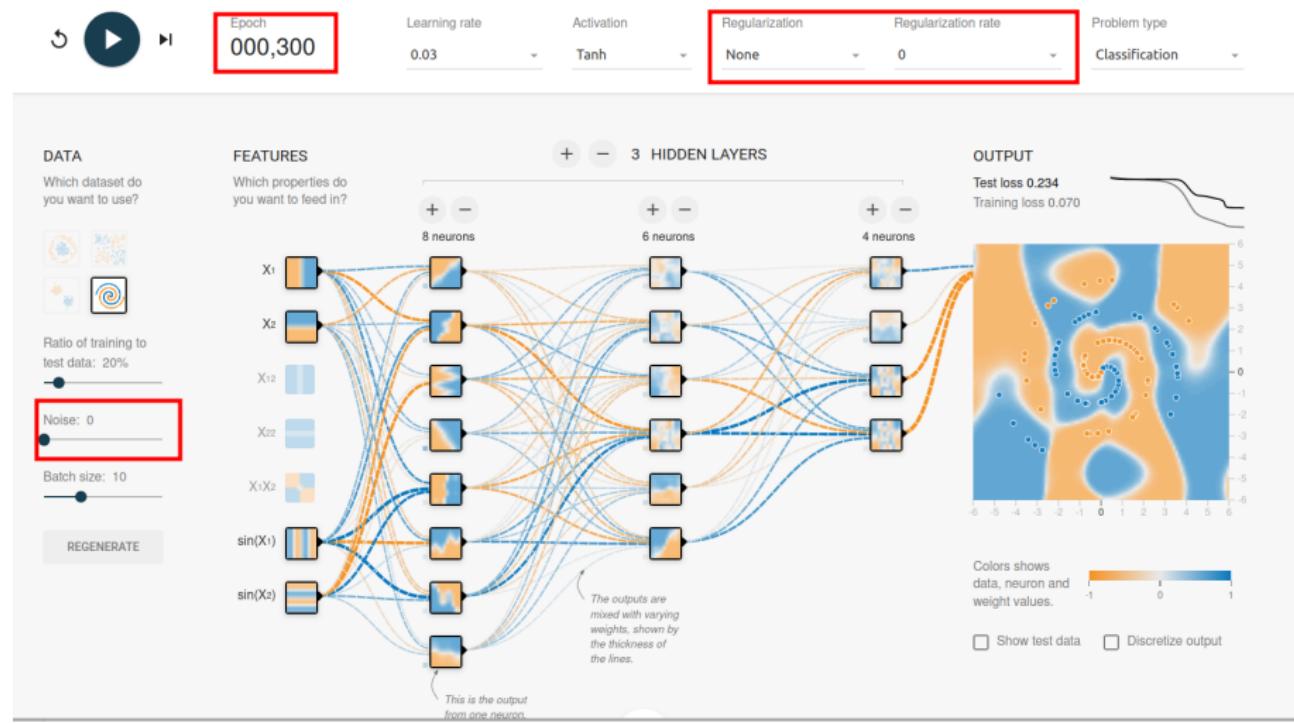
para pequeñas varianzas η , la minimización de J con ruido en los pesos (covarianza ηI) es equivalente a minimizar J con un término adicional de regularización:

$$\eta \mathbb{E}_{p(x,y)} [\|\nabla_w \hat{y}(x)\|^2]$$

Esta forma de regularización empuja al modelo hacia regiones donde es relativamente insensible a pequeñas variaciones en los pesos, encontrando puntos que no son simplemente mínimos, sino mínimos rodeados de regiones planas.

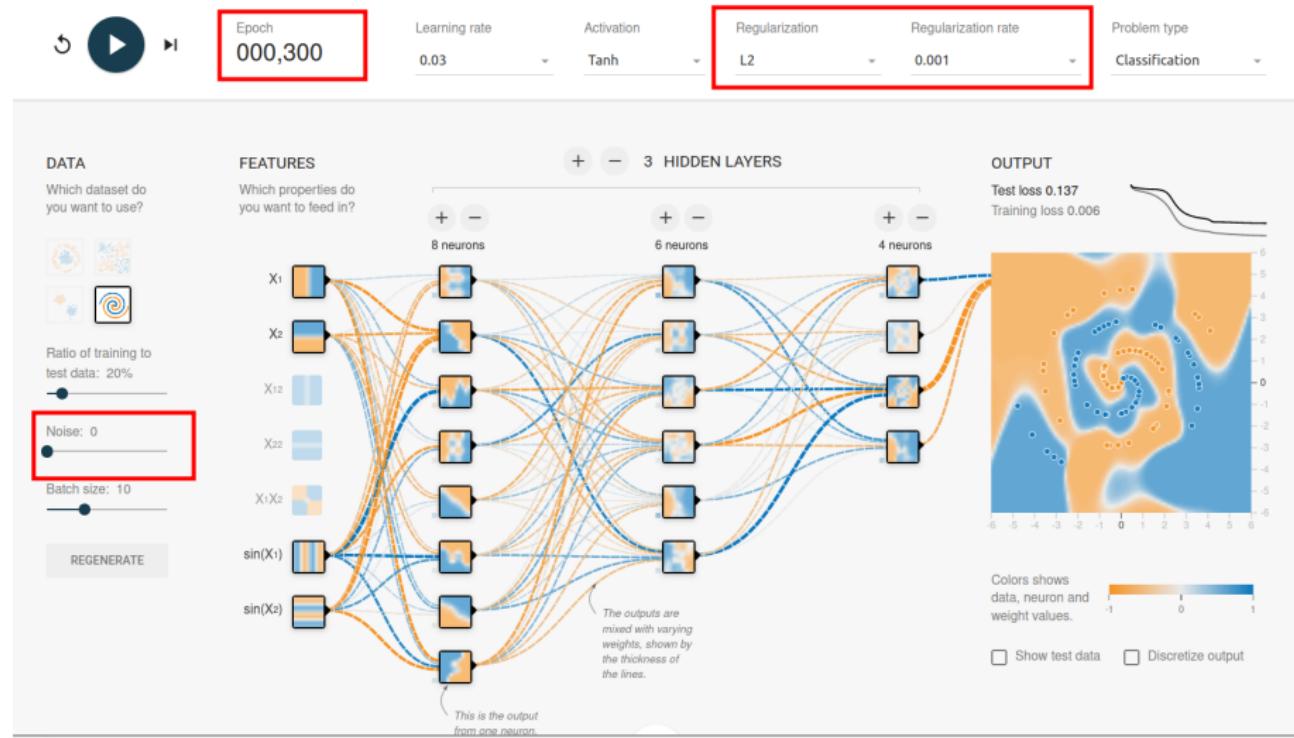
Ejemplo simple en playground.tensorflow

Sin regularización, ni ruido en las entradas.



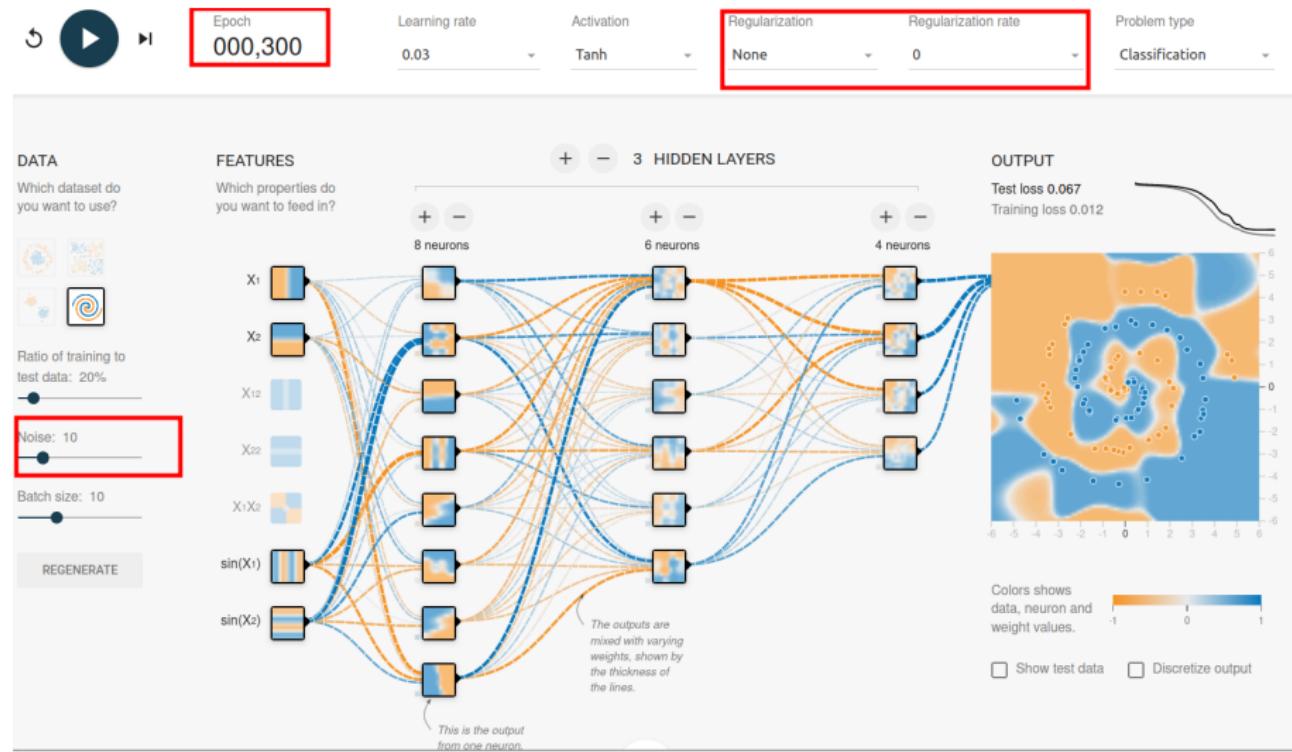
Ejemplo simple en playground.tensorflow

Con regularización, sin ruido en las entradas.



Ejemplo simple en playground.tensorflow

Sin regularización, con ruido en las entradas.



7.5.1 Agregando ruido a las salidas objetivo

Consideraciones:

- La mayoría de los conjuntos de datos tienen algunos errores en las etiquetas y .
- Puede ser perjudicial maximizar $\log p(y|x)$ cuando y es un error.
- Una forma de evitarlo es modelar explícitamente modelar explícitamente el ruido en las etiquetas.

7.5.1 Agregando ruido a las salidas objetivo

Se puede asumir que...

para una pequeña constante ϵ , la etiqueta y del conjunto de entrenamiento es correcta con una probabilidad $1 - \epsilon$, en caso contrario cualquier otra posibilidad podría ser correcta.

7.5.1 Agregando ruido a las salidas objetivo

Se puede asumir que...

para una pequeña constante ϵ , la etiqueta y del conjunto de entrenamiento es correcta con una probabilidad $1 - \epsilon$, en caso contrario cualquier otra posibilidad podría ser correcta.

Esta suposición es más fácil de incorporar analíticamente a la función objetivo, que explícitamente muestrear ruido.

7.5.1 Agregando ruido a las salidas objetivo

Ejemplo: Label smoothing (suavizado de etiquetas)

Regulariza un modelo basado en softmax con k clases, remplazando las clasificaciones 0 y 1 con $\frac{\epsilon}{k-1}$ y $1 - \epsilon$ respectivamente. Se usa desde la década de 1980 hasta la fecha.

7.5.1 Agregando ruido a las salidas objetivo

Ejemplo: Label smoothing (suavizado de etiquetas)

Regulariza un modelo basado en softmax con k clases, remplazando las clasificaciones 0 y 1 con $\frac{\epsilon}{k-1}$ y $1 - \epsilon$ respectivamente. Se usa desde la década de 1980 hasta la fecha.

Se puede usar la función de pérdida entropía-cruzada para estos objetivos suavizados. Aprendizaje MLE con clasificador softmax puede nunca converger, al no poder predecir 0 y 1 de manera exacta e intentar ser cada vez más exacto.

7.5.1 Agregando ruido a las salidas objetivo

Ejemplo: Label smoothing (suavizado de etiquetas)

Regulariza un modelo basado en softmax con k clases, remplazando las clasificaciones 0 y 1 con $\frac{\epsilon}{k-1}$ y $1 - \epsilon$ respectivamente. Se usa desde la década de 1980 hasta la fecha.

Se puede usar la función de pérdida entropía-cruzada para estos objetivos suavizados. Aprendizaje MLE con clasificador softmax puede nunca converger, al no poder predecir 0 y 1 de manera exacta e intentar ser cada vez más exacto.

Ventaja

Previene la búsqueda de probabilidades rígidas sin desalentar la clasificación correcta.

7.6 Aprendizaje semi-supervisado

Utiliza tanto ejemplos no etiquetados de $P(x)$ como ejemplos etiquetados de $P(x, y)$ para estimar $P(y|x)$ o predecir y a partir de x .

7.6 Aprendizaje semi-supervisado

En el contexto del aprendizaje profundo, el aprendizaje semi-supervisado, por lo regular se refiere a una representación $h = f(x)$.

7.6 Aprendizaje semi-supervisado

En el contexto del aprendizaje profundo, el aprendizaje semi-supervisado, por lo regular se refiere a una representación $h = f(x)$.

Objetivo

Aprender una representación en la cual los ejemplos de una misma clase tengan representaciones similares.

7.6 Aprendizaje semi-supervisado

En el contexto del aprendizaje profundo, el aprendizaje semi-supervisado, por lo regular se refiere a una representación $h = f(x)$.

Objetivo

Aprender una representación en la cual los ejemplos de una misma clase tengan representaciones similares.

Ejemplo

Usar PCA como preprocessamiento antes de aplicar un clasificador.

7.6 Aprendizaje semi-supervisado

Es pueden combinar los componentes supervisados y no supervisados en el modelo, de suerte que se tenga un modelo generativo de $P(x)$ o $P(x, y)$ compartan parámetros con un modelo discriminante de $P(y|x)$.

7.6 Aprendizaje semi-supervisado

Es pueden combinar los componentes supervisados y no supervisados en el modelo, de suerte que se tenga un modelo generativo de $P(x)$ o $P(x, y)$ compartan parámetros con un modelo discriminante de $P(y|x)$.

El uso de ejemplos no etiquetados para modelar $P(x)$ mejora significativamente $P(y|x)$.

7.6 Aprendizaje semi-supervisado

Journal Club

Comentar Capítulo 1 de Chapelle et. al (2006).

7.6 Aprendizaje semi-supervisado

Journal Club

Comentar Capítulo 1 de Chapelle et. al (2006).

Algunas conclusiones de la lectura con relación a regularización

Utilizar aprendizaje semi-supervisado permite generar modelos más robustos para los datos, pues en lugar de solo apuntalar esfuerzos a $p(y|x)$ también se considera la estructura intrínseca de los datos mediante $p(x)$.

7.6 Aprendizaje semi-supervisado

Journal Club

Comentar Capítulo 1 de Chapelle et. al (2006).

Algunas conclusiones de la lectura con relación a regularización

Utilizar aprendizaje semi-supervisado permite generar modelos más robustos para los datos, pues en lugar de solo apuntalar esfuerzos a $p(y|x)$ también se considera la estructura intrínseca de los datos mediante $p(x)$. Dado que la regularización tiene como objetivo evitar malos modelos (sobreajustados o subajustados) se podría considerar el aprendizaje semi-supervisado como un enfoque para ello.

7.7 Aprendizaje multi-tarea

Def.

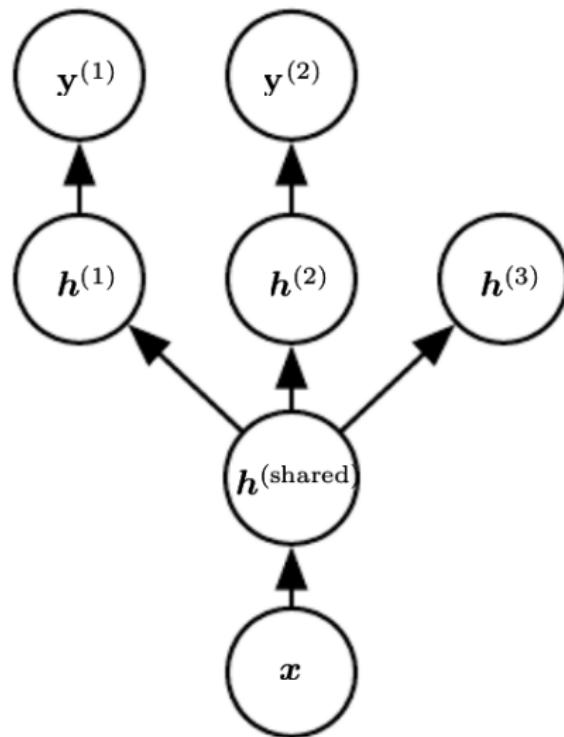
Es una forma de mejorar la generalización mediante la puesta en común (pooling) de ejemplos (que pueden verse como restricciones blandas impuestas a los parámetros) que surgen de varias tareas.

7.7 Aprendizaje multi-tarea

Se puede plantear como dos tipos de tareas con sus asociados parámetros.

- ① Parámetros de tareas específicas, que solo se benefician de los ejemplos de su tarea para lograr una buena generalización.
- ② Parámetros genéricos compartidos a lo largo de todas las tareas, que se benefician de los datos agrupados de todas las tareas.

7.7 Aprendizaje multi-tarea



7.7 Aprendizaje multi-tarea

Premisa desde el punto de vista del aprendizaje profundo

Entre los factores que explican las variaciones observadas en los datos asociados a las distintas tareas, algunos son compartidos entre dos o más tareas.

7.7 Aprendizaje multi-tarea

Premisa desde el punto de vista del aprendizaje profundo

Entre los factores que explican las variaciones observadas en los datos asociados a las distintas tareas, algunos son compartidos entre dos o más tareas.

Ejemplo

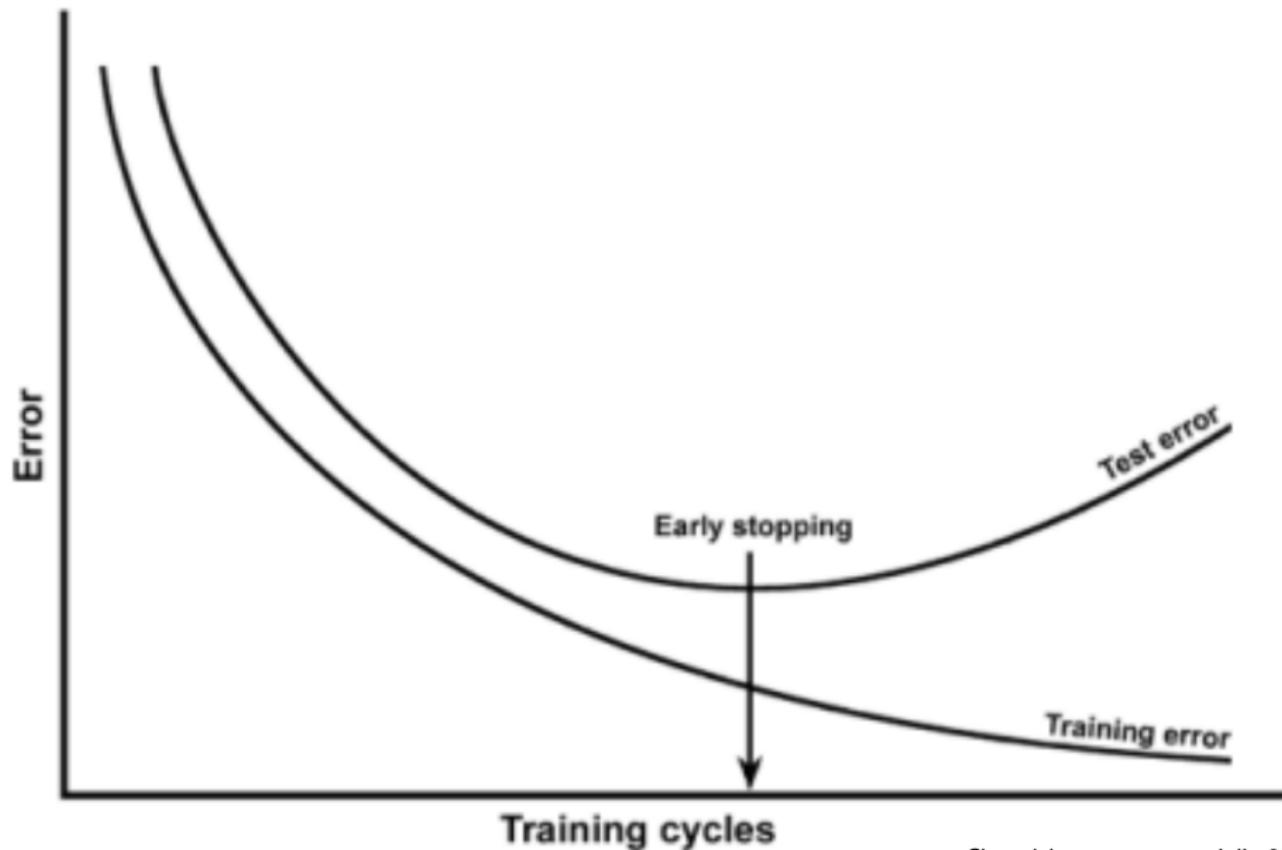
Clasificador softmax incluye neuronas de tareas específicas.

Detención temprana (early stopping)

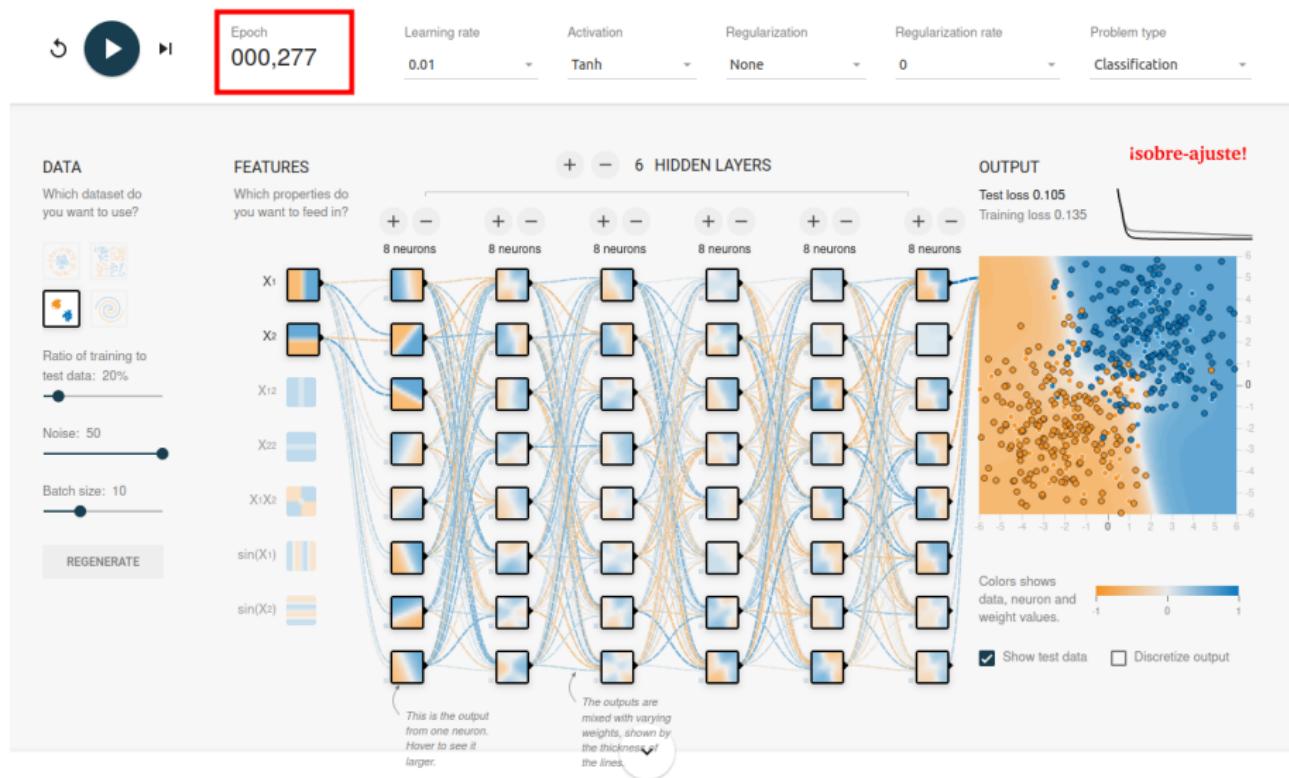
Motivación:

- A menudo hay modelos con la capacidad suficiente para provocar sobreajuste.
- El error de validación empieza a aumentar a medida que el de entrenamiento disminuye.
- Se pueden guardar los errores, pesos y etapas del entrenamiento para cada época y regresar a la mejor configuración, en lugar de a la última.

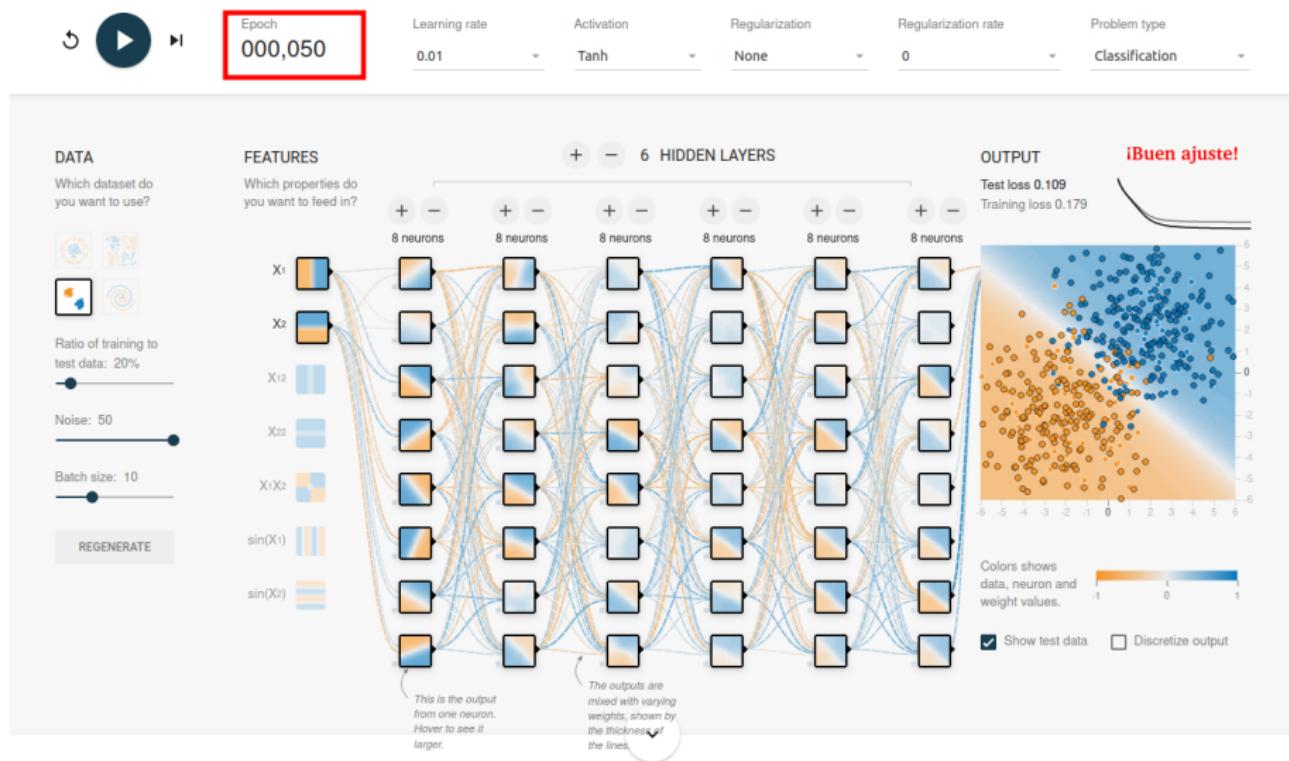
Detención temprana



Detención temprana - Ejemplo gráfico



Detención temprana - Ejemplo gráfico



Detención temprana - Algoritmo 7.1, primera parte

Algorithm 7.1 The early stopping meta-algorithm for determining the best amount of time to train. This meta-algorithm is a general strategy that works well with a variety of training algorithms and ways of quantifying error on the validation set.

Let n be the number of steps between evaluations.

Let p be the “patience,” the number of times to observe worsening validation set error before giving up.

Let θ_o be the initial parameters.

$$\theta \leftarrow \theta_o$$

$$i \leftarrow 0$$

$$j \leftarrow 0$$

$$v \leftarrow \infty$$

$$\theta^* \leftarrow \theta$$

$$i^* \leftarrow i$$

Detención temprana - Algoritmo 7.1, segunda parte

while $j < p$ **do**

 Update θ by running the training algorithm for n steps.

$i \leftarrow i + n$

$v' \leftarrow \text{ValidationSetError}(\theta)$

if $v' < v$ **then**

$j \leftarrow 0$

$\theta^* \leftarrow \theta$

$i^* \leftarrow i$

$v \leftarrow v'$

else

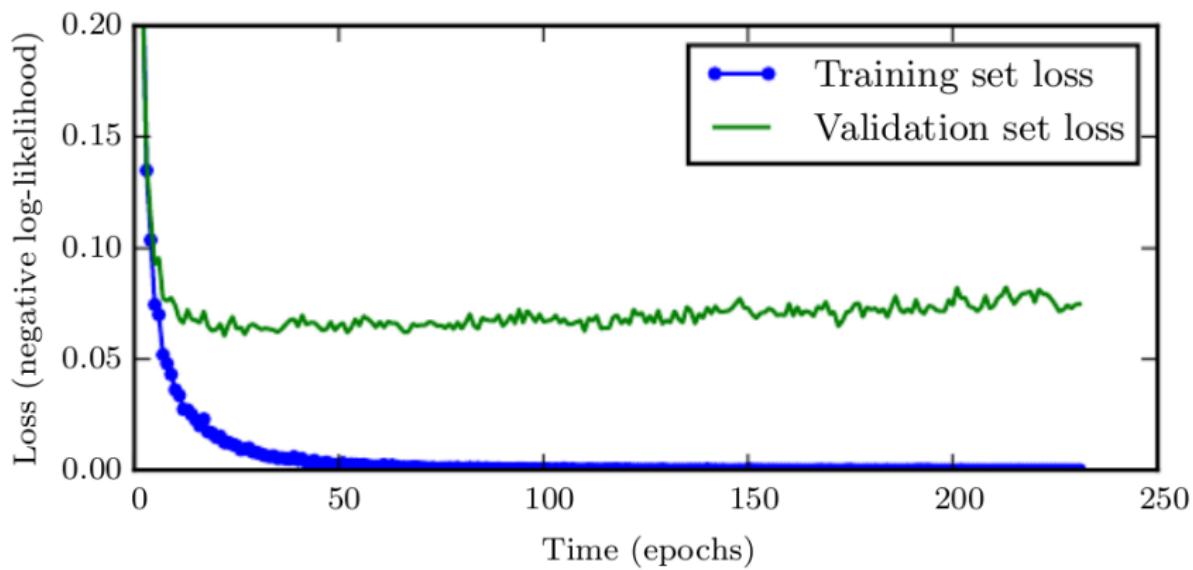
$j \leftarrow j + 1$

end if

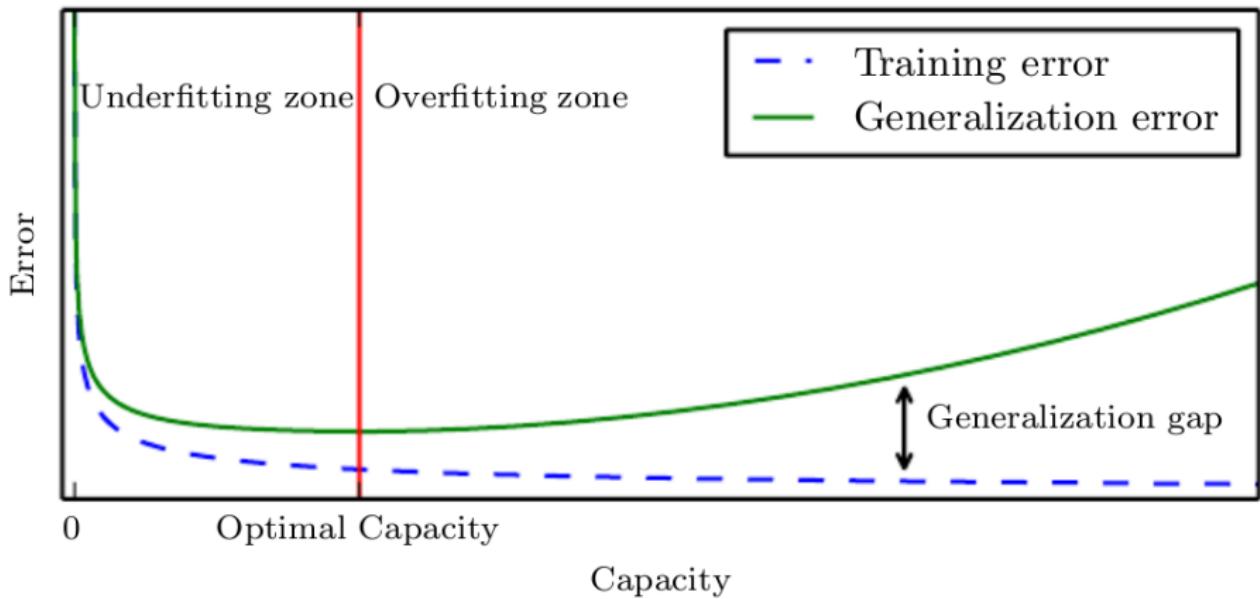
end while

Best parameters are θ^* , best number of training steps is i^*

Detención temprana (early stopping)



Detención temprana (early stopping)



Detención temprana

Generalidades:

- Es el método más común de regularización debido a su simplicidad.

Detención temprana

Generalidades:

- Es el método más común de regularización debido a su simplicidad.
- Los pasos de entrenamiento (épocas) se pueden ver como otro hiper-parámetro.

Detención temprana

Costos:

- Se debe evaluar también la función de costo durante el entrenamiento, aunque sea de manera periódica.

Detención temprana

Costos:

- Se debe evaluar también la función de costo durante el entrenamiento, aunque sea de manera periódica.
- Se debe guardar una copia de los mejores hiper-parámetros.

Detención temprana

Algunas ventajas:

- Es una forma de regularización poco trasgresora, ya que no requiere cambiar el procedimiento de aprendizaje, función objetivo o el valor de los parámetros disponibles.

Detención temprana

Algunas ventajas:

- Es una forma de regularización poco trasgresora, ya que no requiere cambiar el procedimiento de aprendizaje, función objetivo o el valor de los parámetros disponibles.
- Por ejemplo, en el decaimiento de pesos, se debe ser muy cuidadoso de no generar una red que caiga en un mínimo local malo en el cual haya, de manera patológica, pequeños pesos.

Detención temprana

Algunas ventajas:

- Es una forma de regularización poco trasgresora, ya que no requiere cambiar el procedimiento de aprendizaje, función objetivo o el valor de los parámetros disponibles.
- Por ejemplo, en el decaimiento de pesos, se debe ser muy cuidadoso de no generar una red que caiga en un mínimo local malo en el cual haya, de manera patológica, pequeños pesos.
- Se puede usar junto con otras estrategias de regularización (**esto no lo habían destacado los autores en otros métodos**).

Detención temprana - Algoritmo 7.2 - Estrategia 1

Ver algoritmo de la página 117.

Algorithm 7.2 A meta-algorithm for using early stopping to determine how long to train, then retraining on all the data.

Let $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ be the training set.

Split $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ into $(\mathbf{X}^{(\text{subtrain})}, \mathbf{X}^{(\text{valid})})$ and $(\mathbf{y}^{(\text{subtrain})}, \mathbf{y}^{(\text{valid})})$ respectively.

Run early stopping (algorithm 7.1) starting from random $\boldsymbol{\theta}$ using $\mathbf{X}^{(\text{subtrain})}$ and $\mathbf{y}^{(\text{subtrain})}$ for training data and $\mathbf{X}^{(\text{valid})}$ and $\mathbf{y}^{(\text{valid})}$ for validation data. This returns i^* , the optimal number of steps.

Set $\boldsymbol{\theta}$ to random values again.

Train on $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ for i^* steps.

Detención temprana - Algoritmo 7.3 - Estrategia 2

Algorithm 7.3 Meta-algorithm using early stopping to determine at what objective value we start to overfit, then continue training until that value is reached.

Let $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ be the training set.

Split $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ into $(\mathbf{X}^{(\text{subtrain})}, \mathbf{X}^{(\text{valid})})$ and $(\mathbf{y}^{(\text{subtrain})}, \mathbf{y}^{(\text{valid})})$ respectively.

Run early stopping (algorithm 7.1) starting from random $\boldsymbol{\theta}$ using $\mathbf{X}^{(\text{subtrain})}$ and $\mathbf{y}^{(\text{subtrain})}$ for training data and $\mathbf{X}^{(\text{valid})}$ and $\mathbf{y}^{(\text{valid})}$ for validation data. This updates $\boldsymbol{\theta}$.

$\epsilon \leftarrow J(\boldsymbol{\theta}, \mathbf{X}^{(\text{subtrain})}, \mathbf{y}^{(\text{subtrain})})$

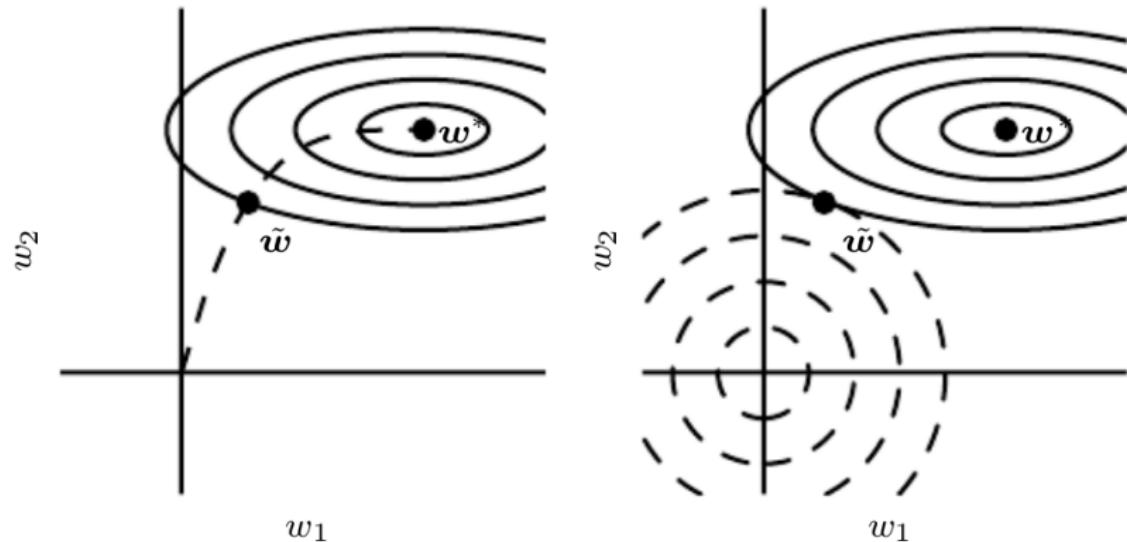
while $J(\boldsymbol{\theta}, \mathbf{X}^{(\text{valid})}, \mathbf{y}^{(\text{valid})}) > \epsilon$ **do**

 Train on $\mathbf{X}^{(\text{train})}$ and $\mathbf{y}^{(\text{train})}$ for n steps.

end while

Ver algoritmo de la página 117.

Regularización con detención temprana



Izquierda.- Las líneas de contorno sólidas indican los contornos de la función de costo. La línea punteada indica la trayectoria tomada por el SGD a partir del origen. En lugar de detenerse en el punto w^* que minimiza el coste, la parada anticipada hace que la trayectoria se detenga en un punto \tilde{w} anterior. **Derecha.-** Regularización L2.