

Introducción a las Redes Neuronales y Aprendizaje Profundo

Prof. Wílmer Pereira

Wilmer Efrén Pereira González PhD

[LinkedIn](#)

Formación:

- Doctorado, en la *Université de Rennes I* sobre lógicas para razonamiento automático.
- Maestría en la *École Supérieure d'Electricité* en Redes de Computadoras
- Ingeniería en Computación de la Universidad Simón Bolívar.

Docencia:

- *Université de Rennes I* y *Université de Dauphine* (Francia)
- Universidad Simón Bolívar y Universidad Central de Venezuela
- Instituto Tecnológico Autónomo de México, Tecnológico de Monterrey e IEXE Universidad (Puebla)
- Escuela de Organización Industrial (España).

Área de investigación:

- Seguridad computacional específicamente en usos de la criptografía para mecanismos de autentificación.
- También inteligencia artificial bioinspirada (algoritmos genéticos, *deep learning*, colonia de hormigas, etc)

Dinámica del curso

El curso es teórico-práctico, con sesiones de teoría, ejercicios y casos de estudio

- Necesitan disponer de almacenamiento en la nube que trabaje conjuntamente con una herramienta de desarrollo de aplicaciones en python, en red, conocida como google colab. Pueden utilizar varios proveedores de almacenamiento en la nube como google drive o github.
- Tendrán a disposición los códigos en python, las laminas del curso y un repositorio reducido de imágenes (dog_vs_cat) popular en las competencias que organiza Kaggle en:

<https://github.com/ISG-UAH/elap2022>

- Se desarrollará una estrategia de enseñanza intermedia, considerando tanto a persona con poca experiencia en el área, como a quienes tengan alguna familiaridad con el tema, auspiciando en ambos grupos el autoaprendizaje

Inteligencia Artificial

Proceso cognitivo, sensorial y mecánico que emula el comportamiento humano

- Hay definiciones de IA que usan la propia palabra “inteligencia” lo cual es claramente inadecuado

«La Inteligencia Computacional es el estudio del diseño de agentes **inteligentes**». (Poole et al., 1998)

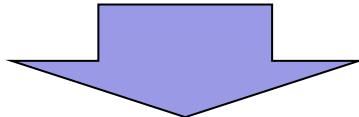
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de **inteligencia**». (Kurzweil, 1990)

- La IA intenta resolver problemas que demandan gran cantidad de recursos computacionales, bien sea en tiempo de CPU o en espacio de almacenamiento. Es decir, pretende resolver problemas cuyas exigencias de recursos crecen exponencialmente en función a la cantidad de datos de entrada. Esto se conoce como la **intratabilidad**.

Aprendizaje automático

Las técnicas clásicas de IA dependen del saber, lo más general posible, que aporta el diseñador (con heurísticas) para a posteriori **deducir** conocimiento. Es decir, las técnicas clásicas de IA no aportan conocimiento, más alla de lo codificado por el diseñador

- En cambio, el **aprendizaje automático**, se apoya en la **inducción** para construir un cuerpo de conocimiento dinámico que aumenta y se modifica. Es decir, el conocimiento crece apoyandose en la Inferencia Estadística o en Estrategias Bioinspiradas.
- Para ello se utilizan grandes bases de conocimientos (*big data*) para extraer reglas o leyes generales utilizando estrategias basadas en el **empirismo**.

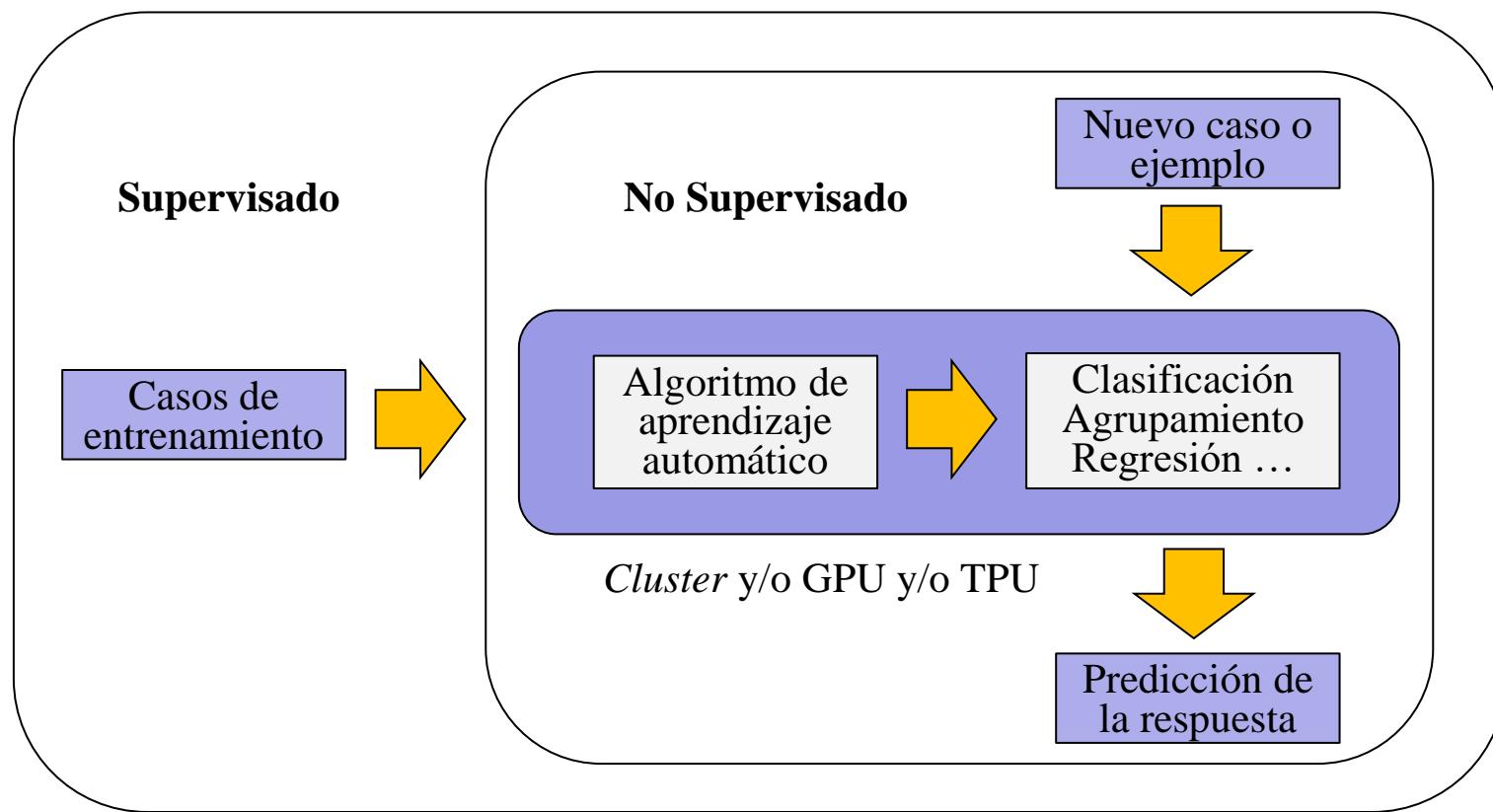


Aprendizaje Automático es una rama de la IA donde el agente mejora su desempeño a través de la **experiencia** mediante el uso de grandes cantidades de datos suministrados en tiempo real o con antelación.

Estrategias de Aprendizaje Automático

○ Las estrategias más generales consisten en:

1. Aprender en tiempo real realizando el mejor esfuerzo dado el conocimiento hasta el momento, es decir, conociendo sólo el pasado (**No Supervisado**).
2. Aprender a priori o por adelantado, con una fase inicial de aprendizaje *off-line*, con datos etiquetados, antes de predecir el nuevo conocimiento (**Supervisado**)



Fases del Aprendizaje Automático Supervisado

El proceso clásico de entrenamiento está conformado por cuatro fases:

- Separación: Dividir los datos disponibles al menos en un conjunto de entrenamiento y un conjunto de prueba.
- Entrenamiento: Se parte de los casos conocidos del conjunto de entrenamiento (x,y) donde x es la entrada y y la salida o la clase que corresponde a ese caso. Se ajustan la técnica para minimizar el error entre la predicción \hat{y} y la salida y
- Prueba: Después del aprendizaje, a posteriori, se verifica el precisión del resultado ante nuevos casos (x,y) perteneciente al conjunto de prueba. Si los resultados no son satisfactorios se entrena de nuevo.
- Operación: Una vez que los resultados de la fase prueba son satisfactorios se coloca el programa en ejecución

Técnicas de Aprendizaje Automático

Hay dos técnicas de aprendizaje automático para predecir resultados: clasificación y regresión.

Clasificación: Se trata de etiquetar una serie de casos o vectores utilizando una o varias categorías o clases. Se vale de conocimiento ya etiquetado para construir un programa que etiquete nuevo conocimiento.

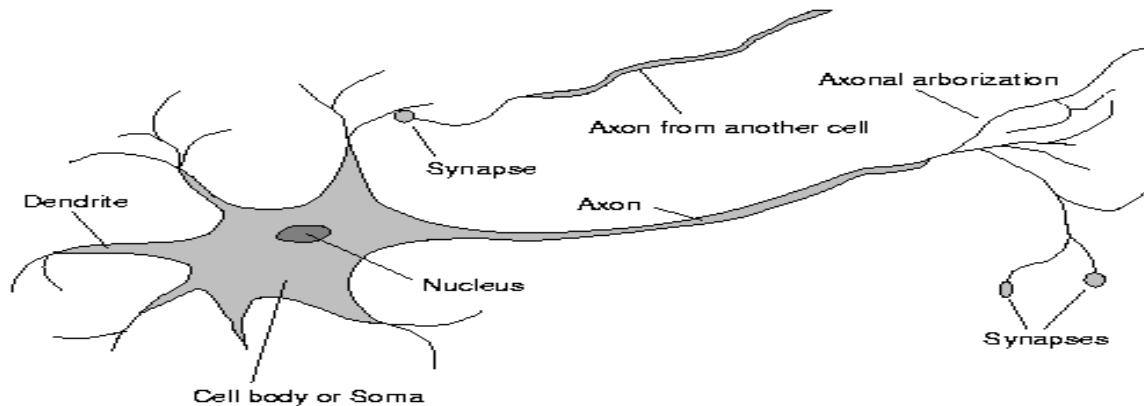
Regresión: Aproxima la relación de dependencia entre una variable de salida y y n variables independiente x_i . El objetivo es aproximar una función para predecir estocásticamente dada una entrada la respuesta y .

El objetivo de ambas técnicas es predecir futuros resultados o generar nuevas soluciones a partir del conocimiento previo

Redes neuronales son una técnica de
aprendizaje automático

Redes Neuronal Natural

Estructura celular del cerebro donde residen las capacidades intelectuales del hombre.
Desde 100×10^9 hasta 10×10^{12} neuronas ...



- Interneuronas
- Neuronas motoras
(directo al músculo)
- Neuronas receptoras
(directo desde el órgano sensor)

Neurona:

Célula nerviosa donde se procesan reacciones químicas

Núcleo:

Codifica el funcionamiento de la neurona (gracias al ADN)

Dendritas:

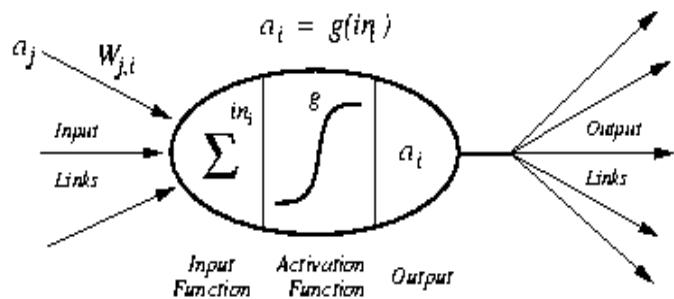
Ramificaciones entre neuronas que representan la entrada

Axón:

Prolongación de la neurona que transmite la activación o salida

Sinapsis:

Punto de unión entre dendritas donde ocurren las reacciones



- Las entradas (a_i) representan las dendritas
- La función de entrada (Σ) sintetiza las reacciones químicas con una suma ponderada
- La función de activación (\int) es el disparador de la neurona
- La salida es el axon que transmite la respuesta

Propiedades de la Red Neuronal Natural



Plasticidad:

Nexos entre neuronas que se fortalecen (temporal o permanentemente) con los patrones de estímulo que generan proteínas en la neurona y la cambian. Esto representa los pesos en la neurona artificial (w_{ij})



Elasticidad:

Capacidad de crecimiento de los nexos entre neuronas (sinapsis). No es dinámico porque una vez definida la red neuronal no cambia su arquitectura, es decir, no agrega más neurona ni capa (nexos ya tiene todos los necesarios ...)



La especialización por capas de una red neuronal artificial en las primeras propuestas (redes neuronales superficiales) no era explícita. En las redes neuronales profundas hay una especialización estratificada ...

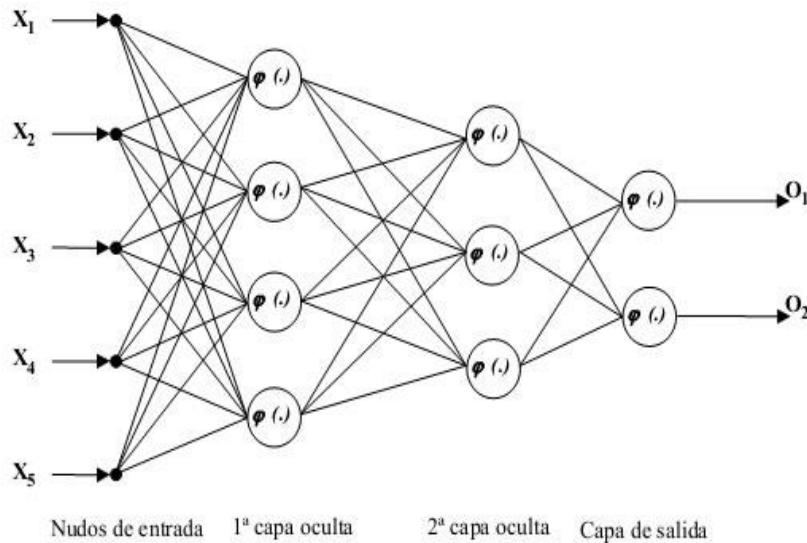
No obstante, sigue siendo un modelo de caja negra pues ninguna red neuronal tiene capacidad de explicación.

Cerebro vs Computadora

- Procesamiento: La maquina con más procesadores tiene 10,649,600 ([TOP500](#)). Un cerebro promedio tiene al menos 100×10^9 (10,000 veces más “procesadores”) ...pero la evolución en cantidad de procesadores crece vertiginosamente ([SCALAC](#) en Latinoamerica)
- Velocidad. Computadora tiempo en η seg y el cerebro tiempo en mseg *pero ... el cerebro es masivamente paralelo. Por ello, en algunas tareas el cerebro 10⁶ veces más rápido*
- Tolerancia a fallas: Una neurona natural dañada afecta de manera marginal el comportamiento del cerebro. En cambio, cualquier error altera todo el procesamiento a nivel de la computadora
- Complejidad de ejecución: El cerebro realiza tareas muy complejas que son sencillas al humano pero difíciles para cualquier computadora (vision)
- Aprendizaje: En el cerebro es *online* mientras que en la computadora frecuentemente es *offline*. Sin embargo, ambos son distribuidos
- Ejecución: Centralizado vs Distribuido
Computadora Cerebro

Redes Neuronales Artificiales

Unidades enlazadas a través de conexiones
cargadas por pesos numéricos w_{ij}

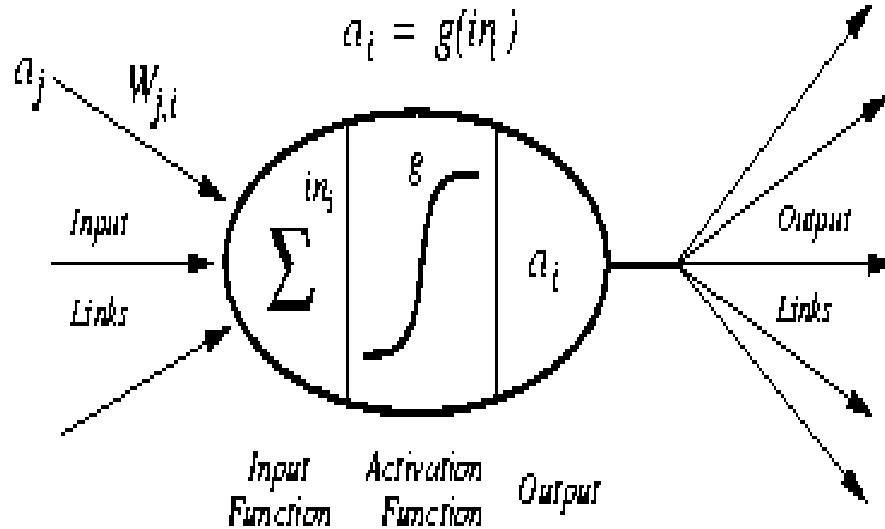


Frecuentemente la unión entre capas es **densa**, es decir, todas de las neuronas de una capa tienen conexión con todas las neuronas de la capa anterior. El conjunto de entradas a cada neurona define la **función de propagación**.

Una red puede tener una o varias salidas dependiendo de si es una clasificador binario o multiclase. También puede tener una salida continua (regresor).

- El **nivel de activación** de la neurona artificial (equivalente al impulso excitatorio) es un cálculo individual en cada neurona, sin control global (distribuido)
- El aprendizaje se basa en la actualización de esos pesos que se inicializan aleatoriamente pero se ajustan en la fase de entrenamiento de la red para acoplarse a los datos de entrada. Este proceso se llama la **retropropagación** (*backpropagation*) y se configura ajustando múltiples hiperparámetros.

Funciones de Propagación



Suma Ponderada sin sesgo

$$in_i = \sum w_{ji}a_j$$

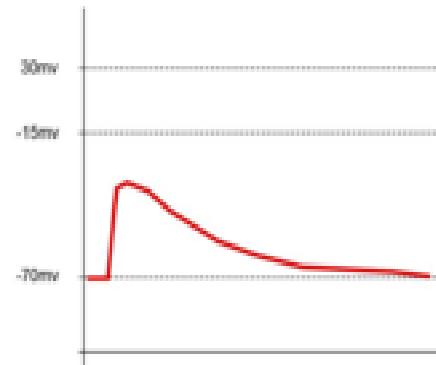
Distancia Euclídea

$$in_i = \sum (a_j - w_{ji})^2$$

Ponderada con sesgo

$$in_i = \sum w_{ji}a_j - b_i$$

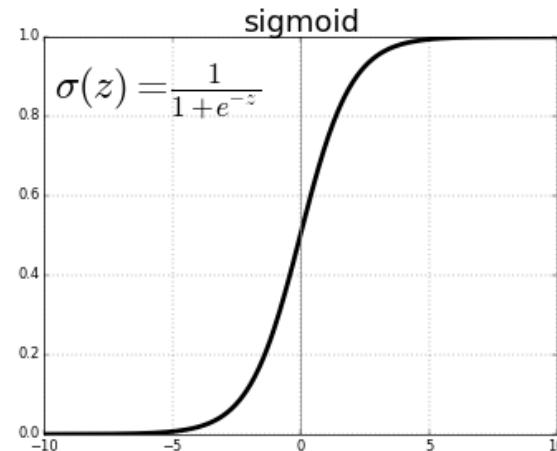
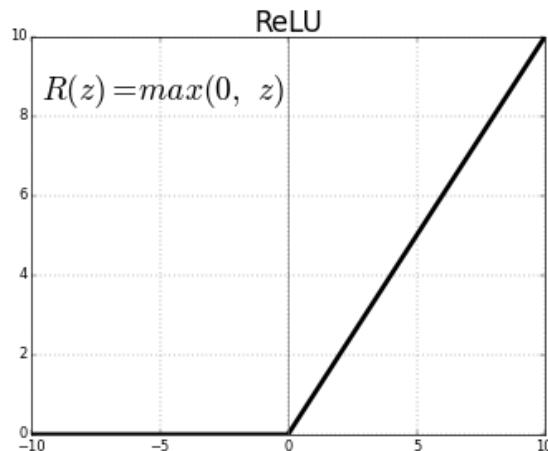
Manhattan, Sigma-Pi, ...



- La función de entrada más común es la suma ponderada que realmente es la ecuación de un hiperplano
- La función de entrada, activación y salida deben estar separadas.
- La función salida normalmente es la identidad $f(x) = x$ aunque el disparo de la neurona puede depender de una probabilidad

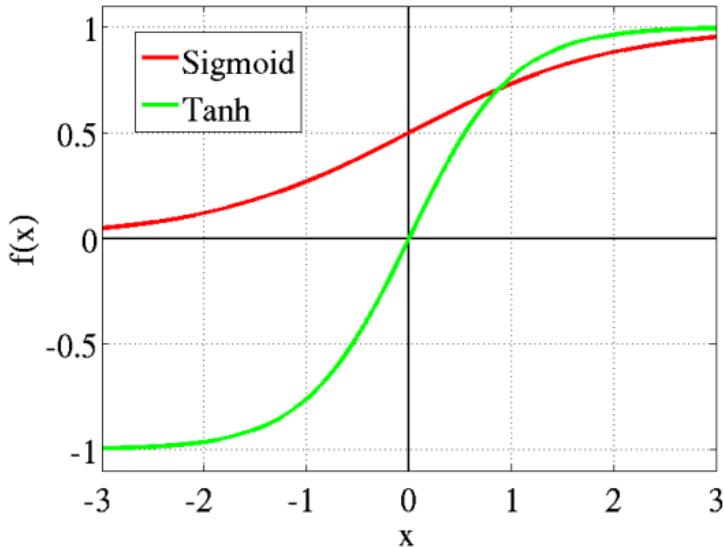
Funciones de Activación (g)

Función que deben tener todas las neuronas artificiales y que determina el valor de salida, por neurona, dado los estímulos de entrada



- La función de activación debe ser monótona, creciente, continua y derivable
- Además de sigmoide y ReLU (con sus variantes) están: softmax, tangente hiperbólica, maxcut, ...
- La función de activación puede ser diferente en cada neurona aunque lo común es que sea la misma función de activación por capa. Frecuentemente en las capas intermedias es ReLU y en la última capa se usa sigmoide (clasificación binaria), softmax (clasificación multiclase) y lineal (regresor).

sigmoide, tanh y softmax



En algunos casos particulares, para redes neuronales recurrente, Tanh es preferida sobre sigmoide. Sin embargo, en la mayoría de los casos, sigmoide se comporta mejor ...

La sigmoide, en general, se utiliza para modelar propagación de epidemias, difusión en redes sociales, ...

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

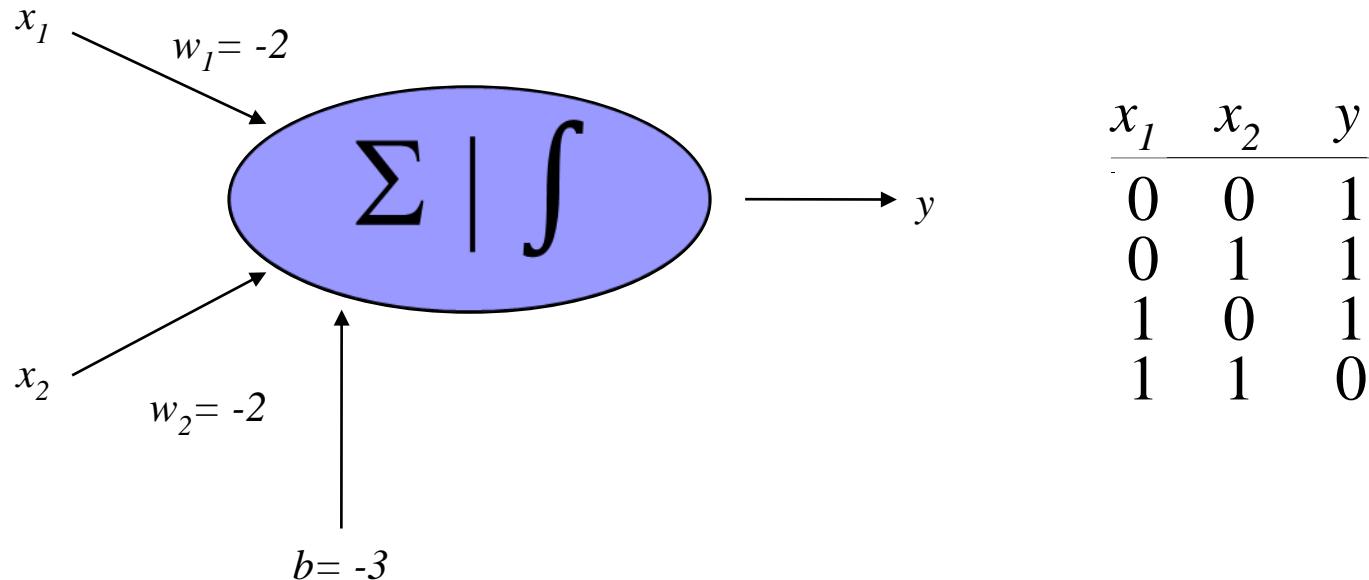
softmax aplica para clasificar cuando las categorías son mutuamente exclusivas por lo que es una generalización de la sigmoide.

No es fácil de graficar porque es una función multidimensional ...

Poder de una Red Neuronal

Como una red neuronal puede representar un NAND entonces, puede a su vez, representar cualquier función booleana, es decir, son capaces de modelar cualquier circuito digital combinatorio (sin retroalimentación)

Sea una red de una sola capa y una sola neurona con función de activación escalón



Sin embargo esto no significa que una red de una capa modele cualquier tipo de función. Por ejemplo, no necesariamente modela un circuito digital secuencial (con memoria)

Otros tipos de Aprendizaje Automático



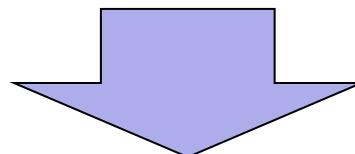
Semisupervisado:

Coexistencia de supervisado y no supervisado. Tiene en cuenta tanto los datos etiquetados como no etiquetados. Es posible en las redes neuronales.



Reforzado:

Se tiene información del error más no de la salida correcta. Con la información del buen o mal comportamiento se ajusta la red. En principio, las redes neuronales no tienen este comportamiento



El objetivo es aprender a mapear acciones para maximizar una cierta función de recompensa por ensayo y error

Red Neuronal Supervisada

El objetivo es construir una función multivariable desconocida $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ a partir de la entrada $x \in \mathbb{R}^n$ y la respuesta $y \in \mathbb{R}^m$.

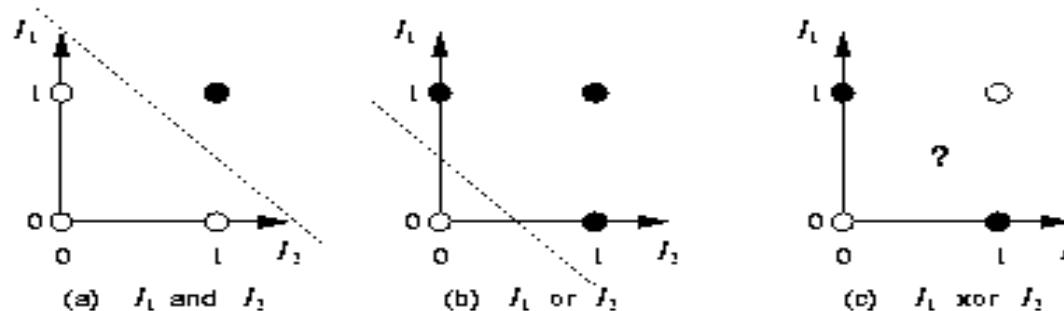
Se minimiza iterativamente el error mediante aproximación estocástica.

El error es una función de comparar y contra \hat{y} (predicción)

- Separación: Dividir los datos de entrada en conjunto de entrenamiento y y conjunto de prueba (frecuentemente se incluye también un conjunto de validación). Se inician los pesos de la red w_j^i aleatorios
- Entrenamiento: Se parte de los casos conocidos del conjunto de entrenamiento (x,y) donde x es la entrada y y la salida o clase que corresponde a ese caso. Se ajustan los pesos w_j^i con retropropagación para minimizar el error entre la predicción \hat{y} y la salida y
- Prueba: Despues del aprendizaje, se verifica el precisión de la red ante nuevos casos (x,y) perteneciente al conjunto de prueba y validación. Mientras la validación y/o las pruebas no sean exitosas se repite el proceso de entrenamiento.
- Operación: Una vez que los resultados de la fase prueba y validación son satisfactorios se coloca la red en ejecución

Ascenso y descenso de las redes neuronales

- La primera propuesta de red neuronal (perceptrón: a una capa, sólo de salida) se demostró que sólo podía modelar problemas linealmente separables (Minsky&Papert 1969). Aún con muchas capas, según ellos, no se solucionaba el problema



... sin embargo ...

- Alrededor de ese mismo año Byron&Ho publicaron el algoritmo para entrenar redes neuronales multicapas con retropropagación usando **descenso de gradiente** ... nadie les prestó atención y sólo 10 años después se redescubrió el algoritmo ☺ ...

... de nuevo ...

- Por el año 2010 el área estaba de nuevo estancándose, hasta que varios investigadores canadienses, franceses y norteamericanos, Hinton, Bengio, LeCun entre otros, reactivaron el área. Estaba el *big data* en pleno crecimiento
- El algoritmo de aprendizaje, clave para la eficiencia del área, es descenso de gradiente, con una gran cantidad de variantes ...

Deep Learning

Esta propuesta se basa en una cantidad considerable de capas, inicialmente inspirado en las redes neuronales superficiales (pocas capas ocultas).

Mientras más grande, sea la red mejor debería ser su capacidad de predicción ...

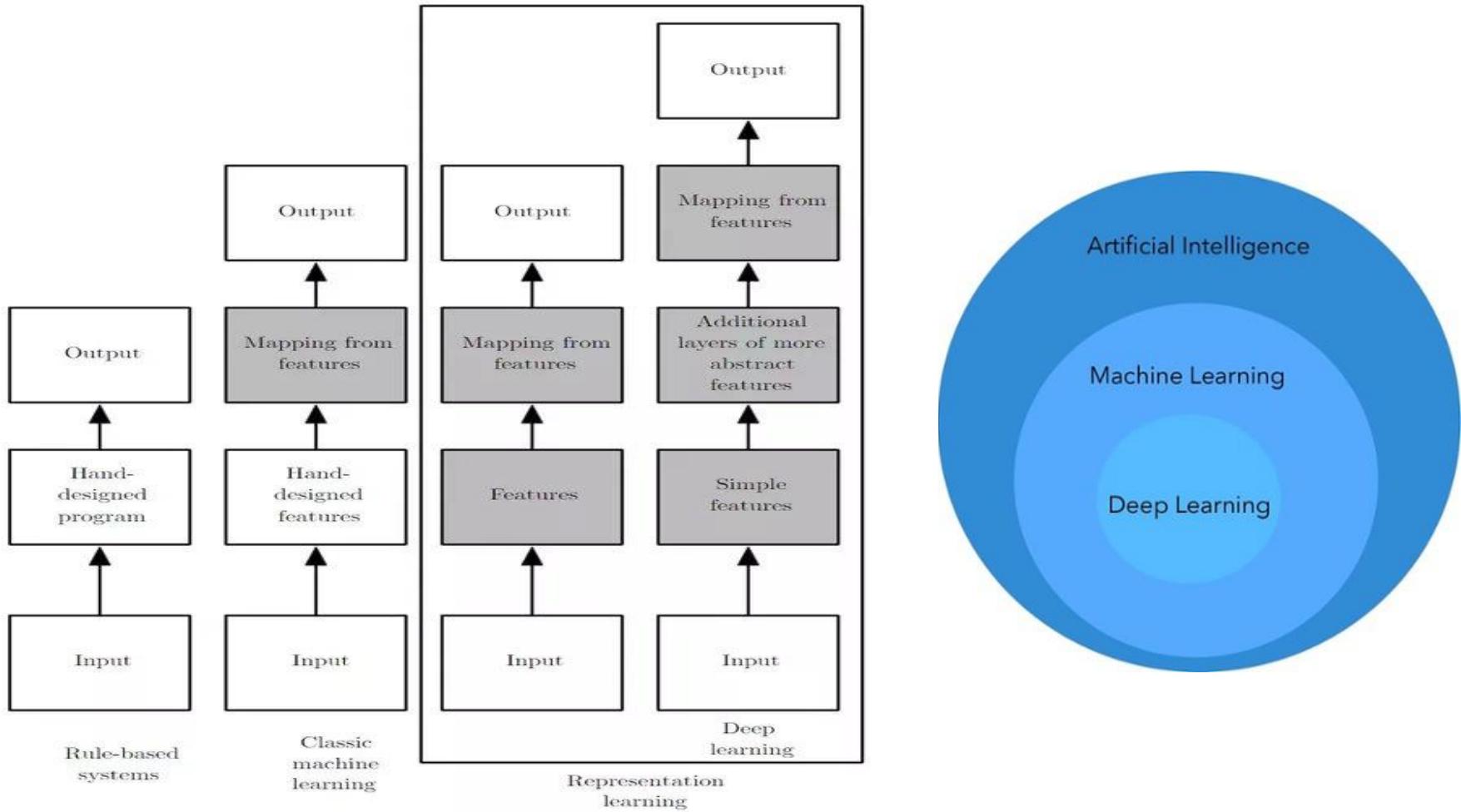
- El énfasis está en colocar sucesivas capas de representación para procesar la información, por lo que el modelo de base es la red neuronal pero *deep learning* no siempre sigue el modelo del cerebro ...
- La gran cantidad de capas eran una limitación hace unos pocos años (*overfitting* y desvanecimiento de gradiente). De hecho alrededor del 2010 muy pocos investigadores trabajaban en redes neuronales
 - ... pero ...
 1. Un grupo de investigadores de IDSIA crearon una inmensa base de datos de imágenes ([ImageNet](#)) y la aparición de las redes neuronales convolucionales, dieron un vuelco al procesamiento digital de imágenes y las competencias de [Kaggle](#)
 2. El enorme crecimiento de número de procesadores que ofrecen las GPU y más recientemente las TPU, así como las plataformas en la nube de procesamiento paralelo, Azure, AWS y Google *cloud* (más específicamente [google colab](#)) ha favorecido el florecimiento del *deep learning*. También está [playground tensorflow](#).
 3. Las redes neuronales superficiales no competían con *random forest* y SVM, pero las mejoras en la retropropagación, propiciaron los avances en *deep learning*...

ImageNet Large Scale Visual Recognition Challenge

YEAR	WINNER	TOP 5 ERROR RATE %
2012	ALEXNET	15.3
2013	ZFNET	11.2
2014	INCEPTION V1 (GoogLeNet) VGG NET (Runner up)	6.67 7.3
2015	ResNet	3.57
2016	ResNeXt	4.1
2017	SENet	2.251
2018	PNASNet-5	3.8

ILSVRC tiene alrededor de 14 millones de imágenes repartidas en un poco más de 21,000 grupos o clases

Diferencias entre IA convencional, Machine Learning y Deep Learning



Red neuronal profunda más grande ...

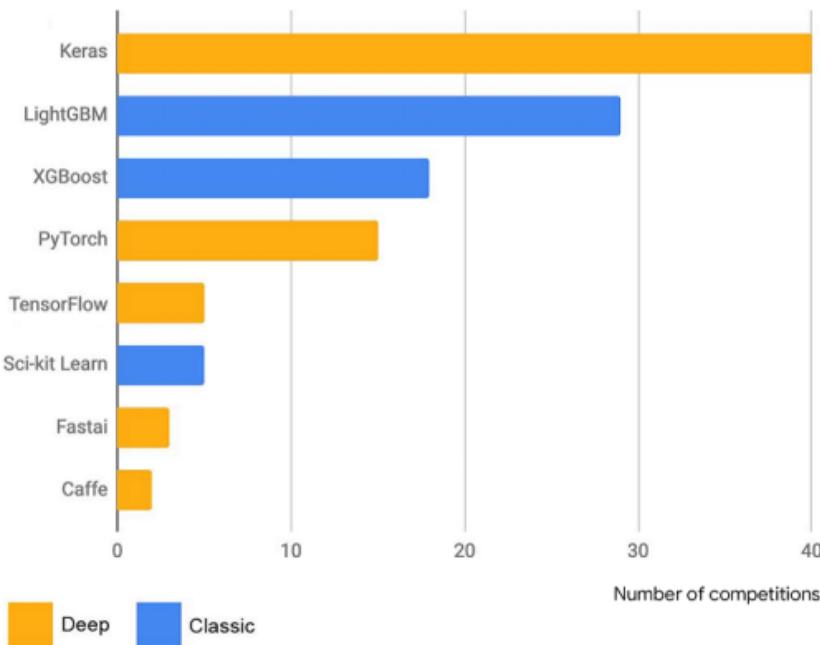
Librerías de deep learning

- Las librerías simples y con capacidad de paralelización automática, han impulsado la democratización del área. La comunidad *deep learning* publica sus resultados en [arXiv](#) antes de presentarlo en conferencias lo que facilita la obtención de información.

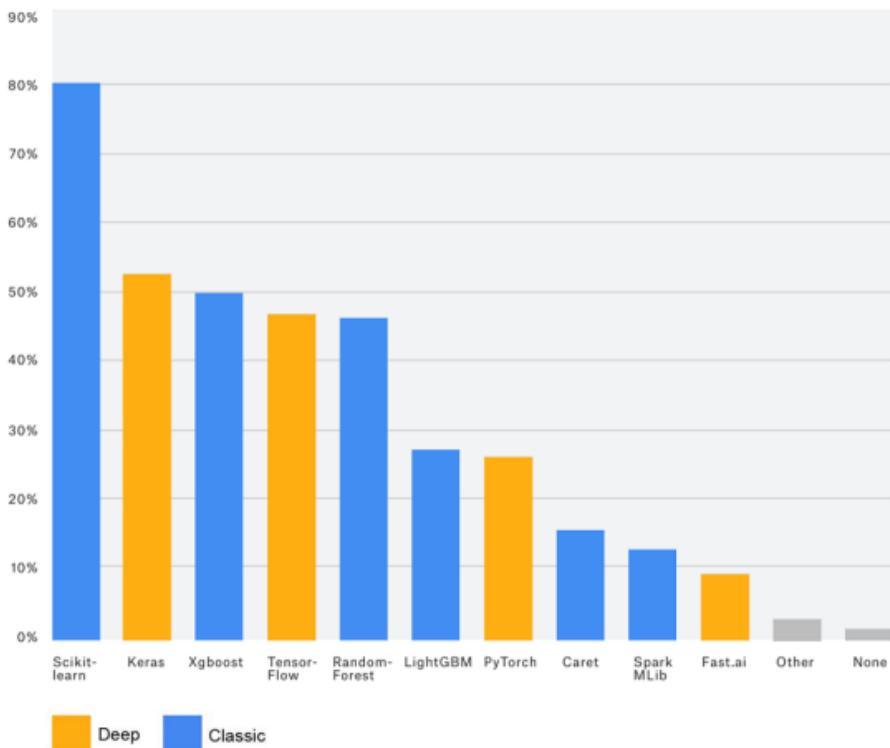
Keras es una librería de alto nivel ([librería MIT](#)) para aprendizaje profundo que puede estar sobre Tensor Flow, Theano o CNTK. Se ejecuta sobre cluster usando [eigen](#) o sobre GPU mediante [cuDNN](#) (librería de CUDA aceleradora de GPU para *deep learning*)

- En *deep learning* , además de las diferentes funciones de activación, también hay varios tipos de funciones de pérdida y una gran variedad de versiones de descenso de gradiente (optimizadores) para parametrizar las implementaciones.

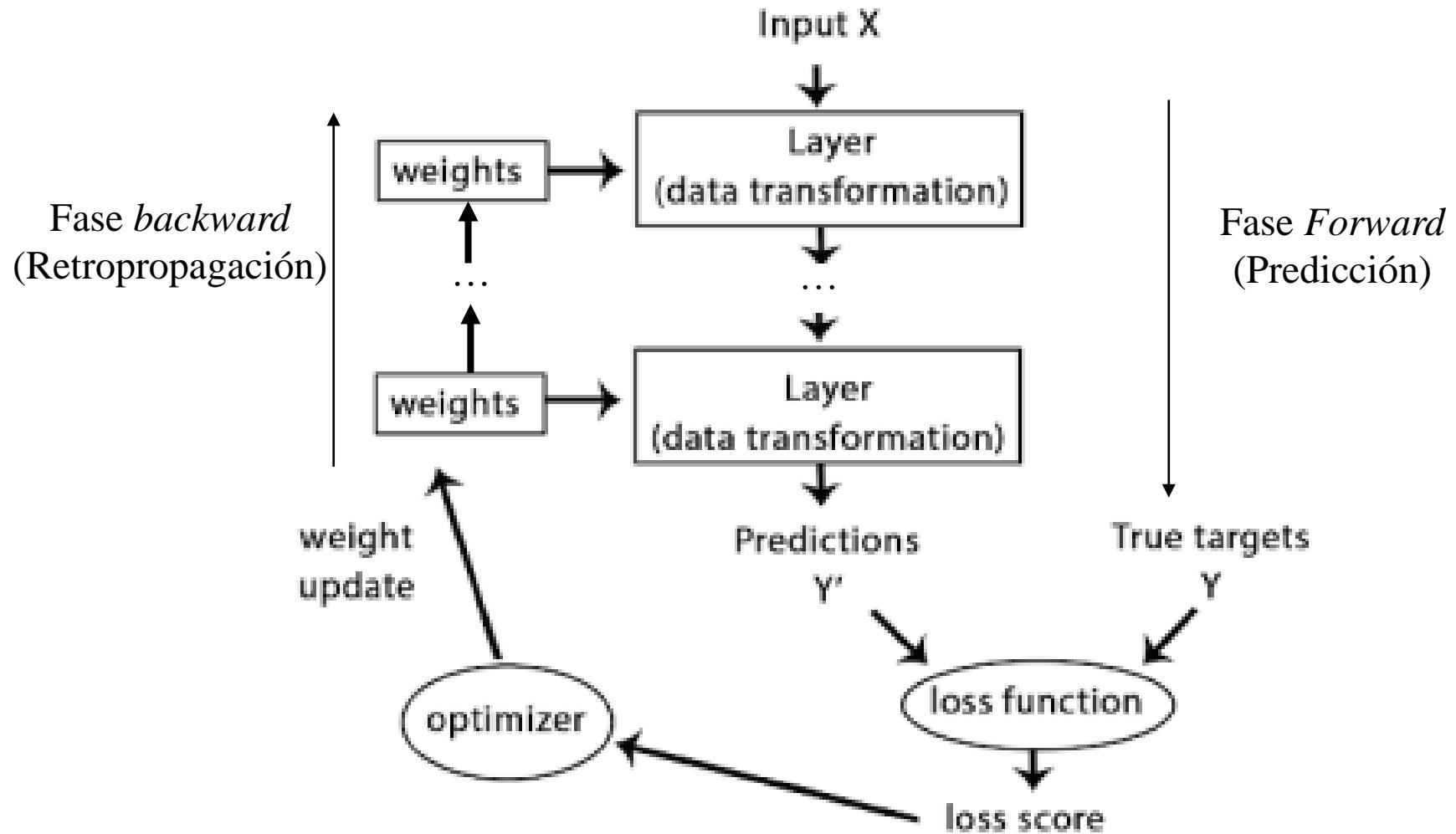
Primary ML tool used by top-5 teams in Kaggle competitions,
2017-2018 (N=120)



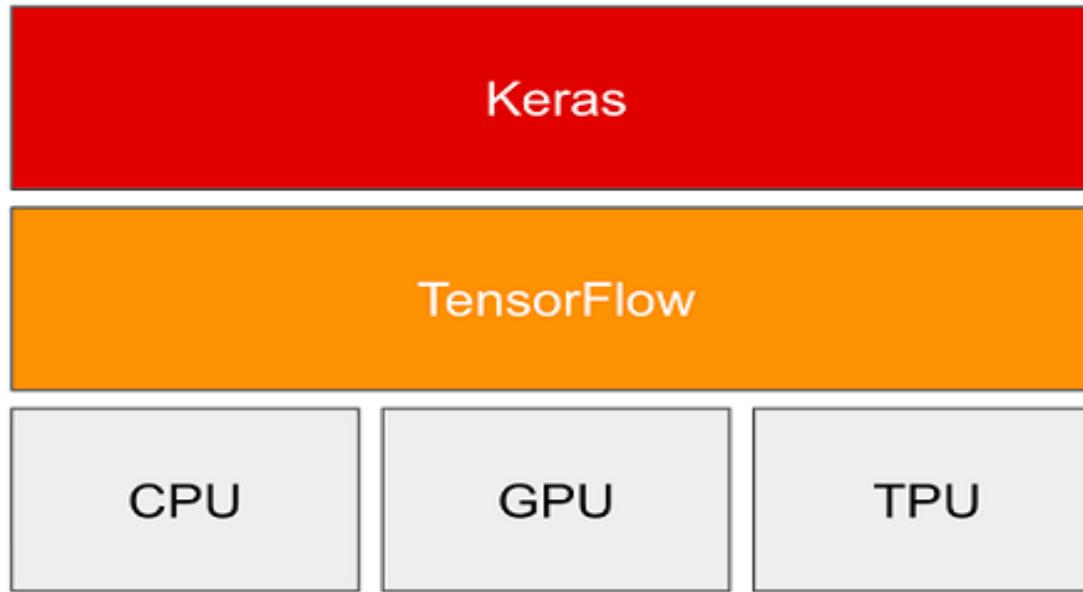
Percentage of machine learning & data science professionals
using each ML software framework, 2019



Principios del proceso de retropropagación



Arquitectura Keras



- A diferencia de TensorFlow, Keras es una librería específica para *Deep Learning*.
No requiere un explícito conocimiento previo de Python.
- Fue desarrollada por François Chollet, ingeniero de Google quien tiene un libro sobre Keras y *Deep Learning*.

Ejemplo simple con Keras

- Una vez importada las clases necesarias y cargada la base de conocimiento para el entrenamiento de la red neuronal, se definen las capas y cantidad de neuronas por capas de la red

```
from tensorflow.keras import models
from tensorflow.keras import layers
model = models.Sequential([
    layers.Dense(512, activation='relu'),
    layers.Dense(10, activation='softmax')
])
```

- A continuación, se fija el algoritmo de retropropagación (`optimizer`) a utilizar, la función de pérdida (`loss`) a minimizar con retropropagación y las métricas para medir el aprendizaje con el conjunto de prueba y el conjunto de validación.

```
model.compile(optimizer='sgd',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy']))
```

- Luego se entrena la red, determinando al cantidad de pasada de entrenamiento con el conjunto de entrenamiento (`epochs`) y la estrategia de actualización de pesos por lotes (`batch_size`)

```
model.fit(train_images, train_labels, epochs=5, batch_size=128)
```

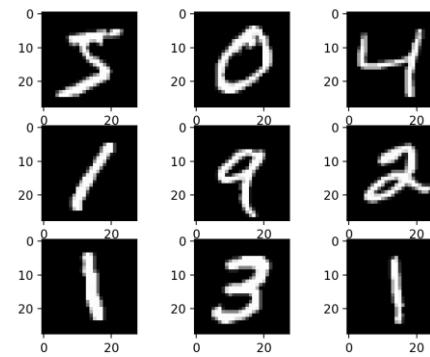
- Por último se verifica el desempeño de la red neuronal con las métricas fijadas para las pruebas validación ([códigos del libro Francois Chollet, 2da edición](#))

```
test_loss, test_acc = model.evaluate(test_images, test_labels)
```

Primera BD a entrenar ...

- Partamos de la conocida BD para reconocimiento de dígitos. [MNIST](#) está formada por 70000 ejemplo etiquetados en 10 clases. Las imágenes son 28x28 píxeles de niveles de grises (256 → 1 byte).

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9



Si aplazamos la entrada de cada dígito, se tienen $28 \times 28 = 784$ neuronas de entrada. Normalmente se deben normalizar a valores entre 0 y 1 (en lugar de 0-255)

Descenso de gradiente

Algoritmo de optimización para encontrar el mínimo de una función diferenciable.

Los hiperplanos que conforman la entrada de cada neurona, son derivables, y lo que se pretende es optimizar los coeficientes w_{ij} tal que se minimice la función del error E (o función de pérdida), es decir, $\min\{E(y, \hat{y})\}$. En principio es intractable ...

- Sea la función de n valores de entrada a una neurona, esta es una función multivariable:

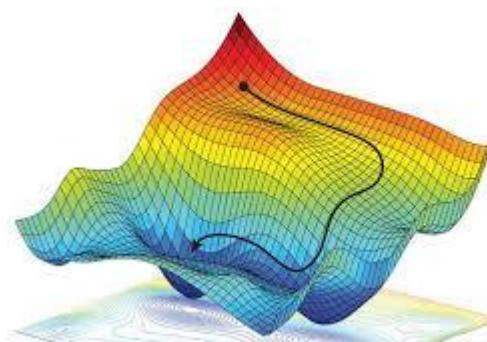
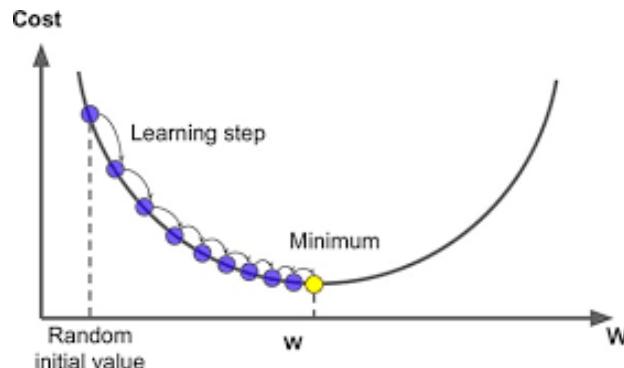
$$y_k = x_1 w_{i_1 j} + x_2 w_{i_2 j} + \dots + x_n w_{i_n j} + b_k$$

La derivada de una función multivariable es el gradiente

$$\nabla y_K = \left(\frac{\partial y_k}{\partial w_{j i_1}}, \frac{\partial y_k}{\partial w_{j i_2}}, \dots, \frac{\partial y_k}{\partial w_{j i_n}} \right)$$

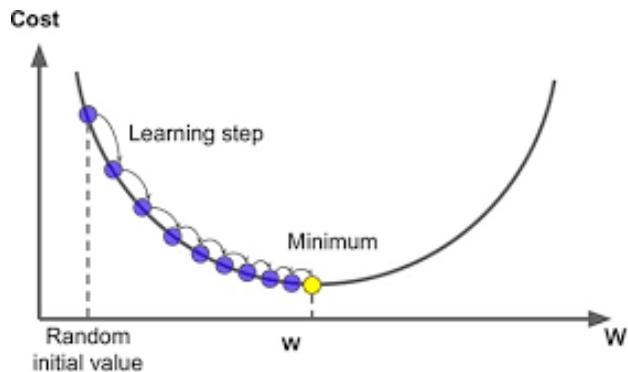
- Entonces el objetivo es minimizar el ajuste de los pesos $w^{k+1} = w^k - \eta \frac{\partial E}{\partial w^k}$

donde η es la tasa de aprendizaje que predetermina la magnitud de los pasos de aproximación al mínimo, al menos local ...



Refinamiento de descenso de gradiente

- La función de error más conocida es el error cuadrático $E = \frac{1}{2}(y - \hat{y})^2$ aunque hay otras funciones de error como la entropía binaria cruzada o el error absoluto medio (regresión)
- Así en la fase forward, se calcula \hat{y} para obtener E y comenzar la fase de retropropagación, por cada peso de la red, hasta la capa de entrada, usando la regla de la cadena. Por ello todos los pesos se ajustarán un pequeño porcentaje, en función a la derivada del error.
- Aunque η no varíe, los pasos de aproximación serán cada vez más pequeños. Lo importante es no tener la tasa de aprendizaje muy grande para evitar la oscilación



Las variantes de descenso de gradiente, conocidas como *optimizer*, permiten una aproximación más eficiente al mínimo

- Hay muchos *optimizer* y entre los más conocidos están: momentum, Nesterov, ADAM, AdamDelta, RMSprop, ... En todos ellos se altera la ecuación de descenso de gradiente

$$w^{k+1} = w^k - \eta \frac{\partial E}{\partial w^k}$$

para lograr más eficiencia en todo el proceso de retropropagación ...

Métodos de optimización para la retropropagación

- El método básico es el descenso estocástico del gradiente (*Stochastic Gradient Descent SGD*) utiliza la fórmula clásica:

$$w_{ij}^{t+1} = w_{ij}^t - \alpha \frac{\partial E}{\partial w_{ij}^t}$$

Sin embargo su convergencia es lenta porque no es posible controlar de manera dinámica la tasa de aprendizaje α .

- La primera variante propuesta ADAGRAD (*Adaptative Gradiente Algorithm*) controla la disminución excesiva de tasa de aprendizaje α e incluye ε para evitar la división por cero

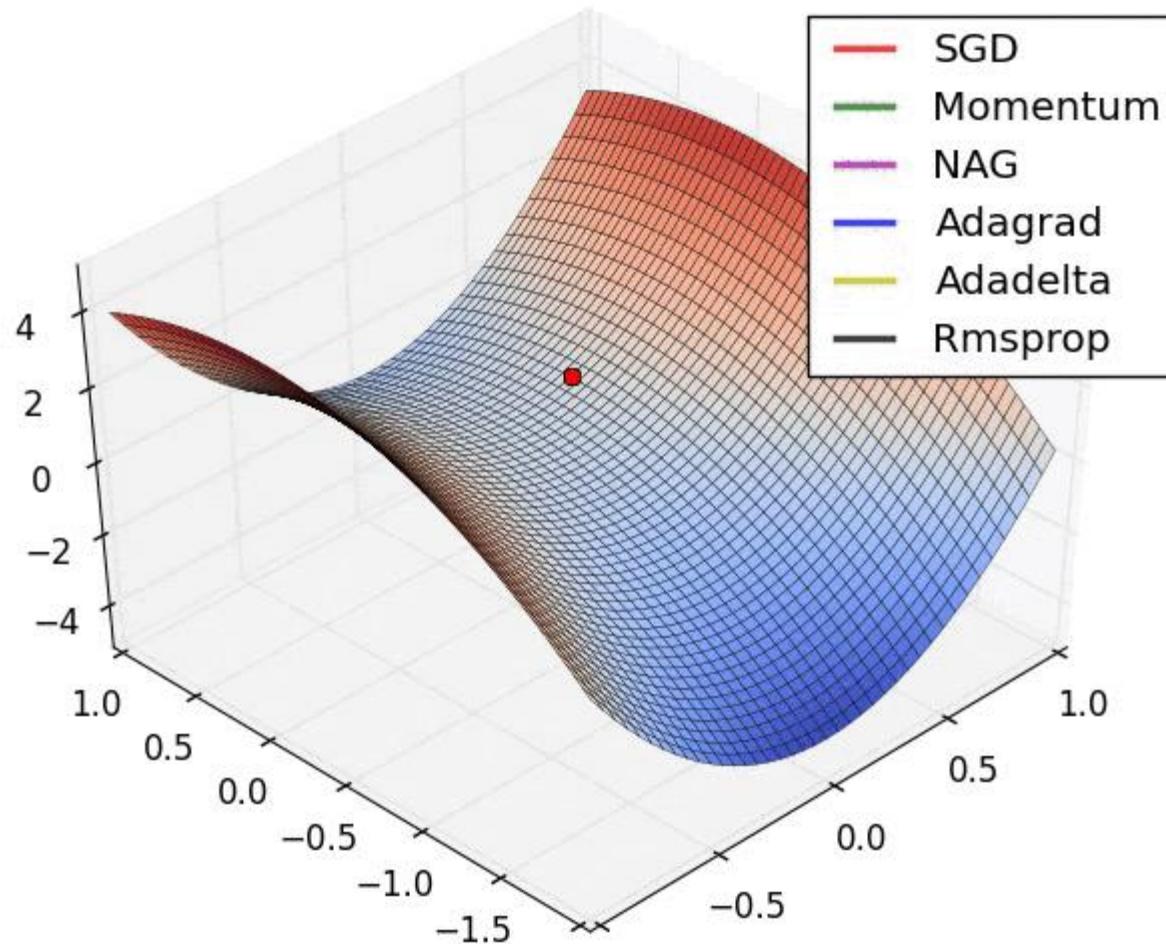
$$C = C + \frac{\partial E}{\partial w_{ij}^t} \quad w_{ij}^{t+1} = w_{ij}^t - \frac{\alpha \frac{\partial E}{\partial w_{ij}^t}}{\sqrt{C + \varepsilon}}$$

- Otra variante muy popular es RMSProp (*Root Mean Square Propagation*) que fue concebida por Geoffrey Hinton pero finalmente no publicó porque lo rechazaron en una conferencia

$$C = C + d.C.(1 - d) \frac{\partial E}{\partial w_{ij}^t} \quad w_{ij}^{t+1} = w_{ij}^t - \frac{\alpha \frac{\partial E}{\partial w_{ij}^t}}{\sqrt{C + \varepsilon}}$$

donde d es la tasa de decaimiento.

Optimizadores



Funciones de pérdida

- La función de pérdida es aquella que se debe minimizar y, para ello, se calcula su gradiente por ser una función multidimensional
- Para las regresiones se usa o el error medio cuadrático (MSE) o el error medio absoluto MAE. Sus representaciones son:

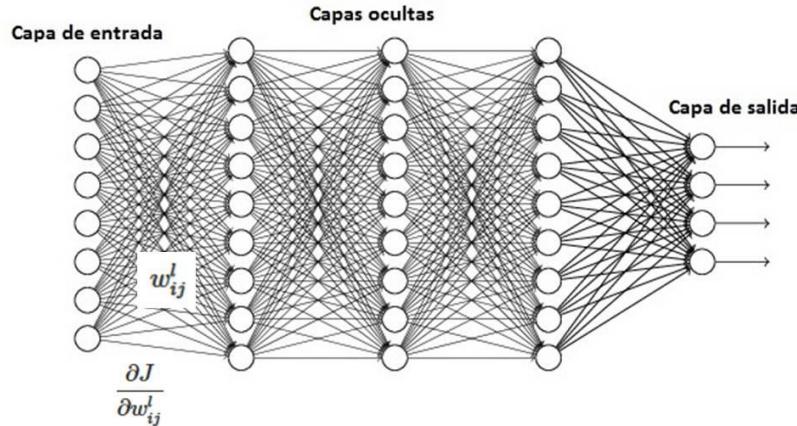
$$MSE = \frac{1}{n} \sum_{j=1}^n (y_i - \hat{y}_i)^2$$

$$MAE = \frac{\sum_{j=1}^n |e_j|}{n}$$

- Para las clasificaciones se usan funciones dependientes de la entropía. La más conocida es entropía cruzada. Será categórica si se trata de un clasificador multiclas o binaria si es un clasificador a sólo dos clases (categorical_crossentropy o binary_crossentropy) ...

Desvanecimiento y explosión de gradiente

- Cuando los pesos tienen valores menores a uno (1) y muchas capas, el gradiente irá disminuyendo a valores muy pequeños lo que hace ineficaz el ajuste de los pesos (desvanecimiento). Por otro lado, cuando la derivada de la función de activación toma valores muy grandes, ocurre el efecto contrario, los pesos aumentarán exponencialmente y la red no se estabilizará (explosión). Ambos problemas ocurren en redes con muchas capas..



Entre las soluciones están: introducir ruido en las funciones de activación de las capas intermedias y asegurarse de que sean no lineales como la activación Leaky ReLU. Para la explosión específicamente se aplica un umbral para evitar el crecimiento del valor en los pesos.

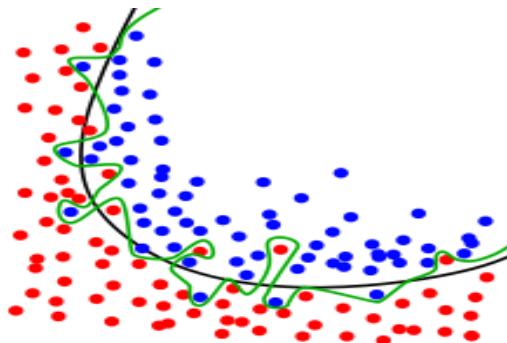
- Otro problema de tener muchas capas y muchos datos de entrenamiento, es la gran cantidad de veces que se deben ajustar de pesos. Una idea es hacer el cálculo de los nuevos pesos para un lote. Es decir, por cada época, definir un grupo de ejemplo, generalmente potencia de 2, para los cuales se hace retropropagación. Un ejemplo en Keras sería:

```
model.fit(train_images, train_labels, epochs=5, batch_size=128)
```

Si se tiene 60.000 ejemplos de entrada, para lotes de 128 por época, entonces se calcularían, sólo 469 corridas de retropropagación por época → 2345 fases de retropropagación en total

Sobreajuste u Overfitting

- Al igual que en el desvanecimiento de gradiente, el *overfitting* se acrecienta mientras más neuronas y capas tenga la red. Lo sufren muchas técnicas de *machine learning* pero especialmente las redes neuronales profundas



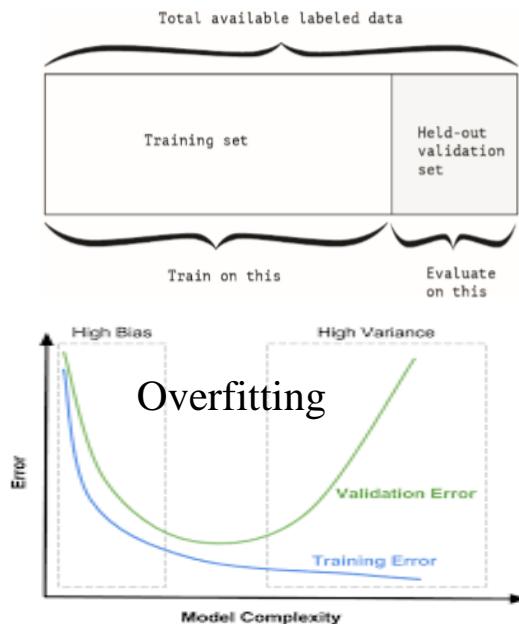
En este caso, clasificar con la línea verde discrimina exactamente los casos de entrenamiento pero generaliza mal porque a la llegada de nuevos casos tendrá mayor probabilidad de ser mal clasificados

- En primer lugar, hay que detectar *overfitting* y si, desafortunadamente, la red neuronal sufre de este problema, hay tres estrategias para eliminarlo:
 1. Reducir el tamaño de la red y así disminuir la capacidad de memoria
 2. Agregar factores de regularización al ajuste de los pesos con L-normas
 3. Llevar a cero, aleatoriamente, algunos neuronas de la red, durante la fase *forward*.
- Con respecto, a la detección del *overfitting*, existen métricas para medir la exactitud de la predicción con el conjunto de prueba. Sin embargo, sería más certero con el **conjunto de validación** ya que da un mejor grado de **generalización** porque se evalúa en tiempo de aprendizaje.

Detección de overfitting: conjunto de validación

- Inicialmente se debe tener una aproximación del número adecuado de capas y de neuronas, tasa de aprendizaje, número de épocas, ... los cuales se definen de antemano. Estas variables son conocidas como los **hiperparámetros** (ver [TensorFlow Playground](#)) ...
- **El conjunto de validación** es un prueba diferente a la comprobación con el conjunto de prueba. Permite ver que tan bien generaliza la red al momento del aprendizaje. Los porcentajes entrenamiento/validación/prueba pueden ser 80/10/10. La validación tiene tres estrategias:

Simple hold-out validation



```
num_validation_samples = 10000  
  
# Shuffling the data is usually appropriate  
np.random.shuffle(data)  
  
# Define the validation set  
validation_data = data[:num_validation_samples]  
data = [num_validation_samples:]  
  
# Define the training set  
training_data = data[:]  
  
# Train a model on the training data  
# and evaluate it on the validation data  
model = get_model()  
model.train(training_data)  
validation_score = model.evaluate(validation_data)  
  
# At this point you can tune your model,  
# retrain it, evaluate it, tune it again...  
  
# Once you have tuned your hyperparameters,  
# is it common to train your final model from scratch  
# on all non-test data available.  
model = get_model()  
model.train(np.concatenate([training_data,  
                           validation_data]))  
test_score = model.evaluate(test_data)
```

Dividir

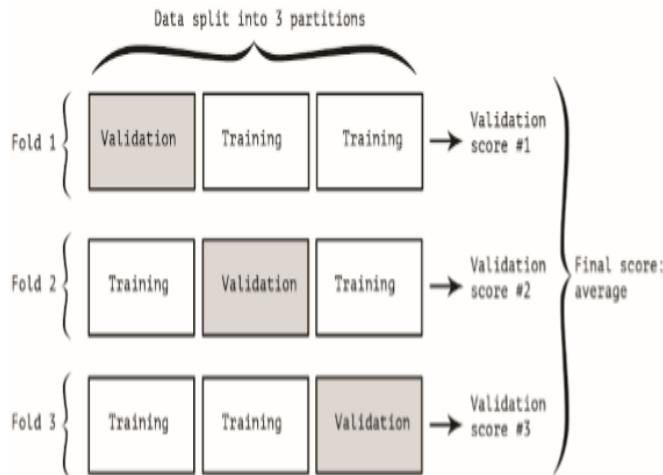
Entrenar y evaluar

Pruebas

Conjunto de validación

- Cuando se tienen pocos datos de entrenamiento, el método *simple hold-out validation*, genera mucho sesgo. El conjunto de validación se puede dividir en k particiones y evaluar sobre cada i -esima porción.

K-fold validation



```
k = 4
num_validation_samples = len(data) // k

np.random.shuffle(data)

validation_scores = []
for fold in range(k):
    # Select the validation data partition
    validation_data = data[num_validation_samples * fold: num_validation_samples * (fold + 1)]
    # The remainder of the data is used as training data.
    # Note that the "+" operator below is list concatenation, not summation
    training_data = data[:num_validation_samples * fold] + data[num_validation_samples * (fold + 1):]

    # Create a brand new instance of our model (untrained)
    model = get_model()
    model.train(training_data)
    validation_score = model.evaluate(validation_data)
    validation_scores.append(validation_score)

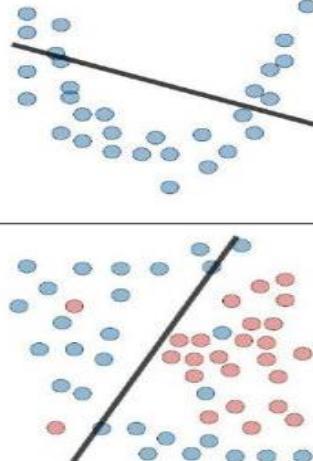
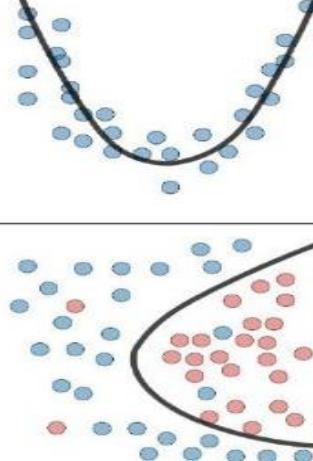
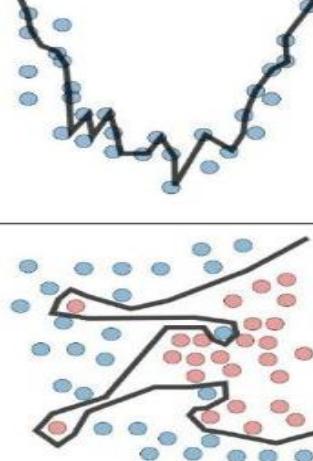
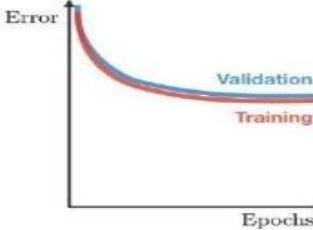
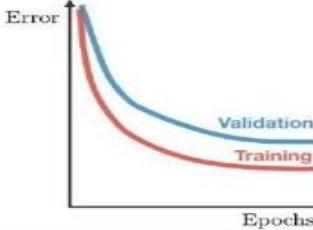
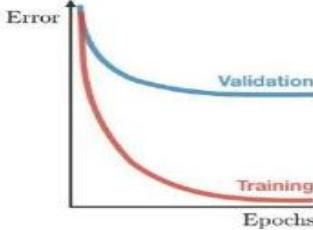
# This is our validation score:
# the average of the validation scores of our k folds
validation_score = np.average(validation_scores)

# We train our final model on all non-test data available
model = get_model()
model.train(data)
test_score = model.evaluate(test_data)
```

- La tercera estrategia se basa en la anterior pero iterando varias veces para tener una idea del desajuste los hiperparámetros ... Esta estrategia es conocida como: *Iterated K-fold validation*. Es evidentemente más costosa en tiempo de aprendizaje ...
- Cualquiera de estos tres protocolos de comprobación por validación debe asegurarse de la representatividad y aleatoriedad de cada uno de los conjuntos (entrenamiento, validación y prueba), evitando los datos repetidos ...

Overfitting vs Underfitting

- Así como se puede sobreaprender también puede ocurrir subaprendizaje (el desvanecimiento de gradiente puede generar *underfitting*). Ambos se pueden detectar con el conjunto de validación:

Symptoms	Underfitting	Just right	Overfitting
Regression	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

Segundo BD: IMDB

IMDB es una base de datos de 50,000 registros de texto sobre opiniones de películas donde 50% son reportes positivos y 50% negativos.

- Se trata de textos cortos formados de 10,000 palabras más frecuentes. Adjunto vienen las etiquetas que discriminan en sólo dos clases: 0 y 1 (crítica positiva o negativa). Se reservarán 15000 textos para entrenamiento, 10000 textos para validación y 25000 textos para pruebas.
- Cada texto está construido como una secuencia de enteros. Cada entero corresponde a una palabra para así representar el texto como un vector de enteros
- Las reseñas se transforman en una matriz dispersa de 0 y 1 donde las filas representan una crítica y las columnas los enteros asociados a las 10,000 palabras posibles. Si una reseña tiene una palabra tendrá un 1 representando la presencia de una palabra en esa opinión. Si la palabra aparece varias veces en el texto sólo habrá un 1 por columna.

Corregir el overfitting: reducción

- Una vez que se ha detectado el *overfitting*, una primera solución es reducir la talla de la red

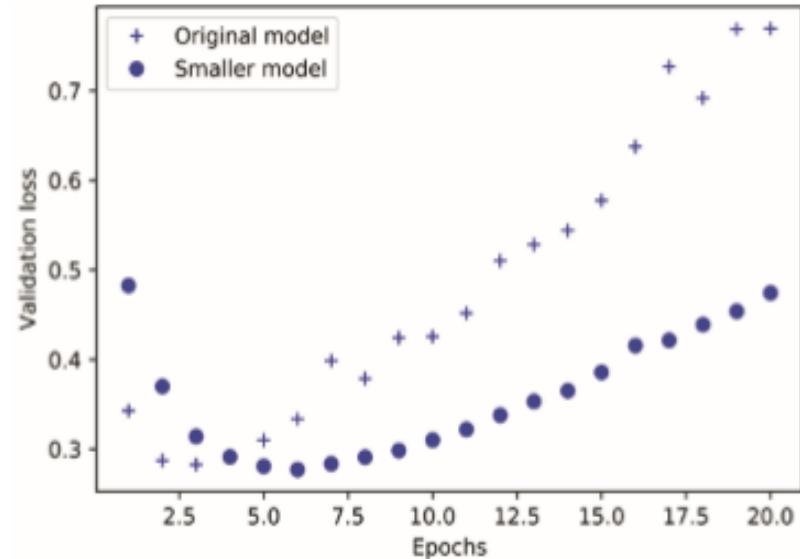
Modelo inicial (+)

```
from keras import models
from keras import layers

model = models.Sequential()
model.add(layers.Dense(16, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(16, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```

Modelo reducido (•)

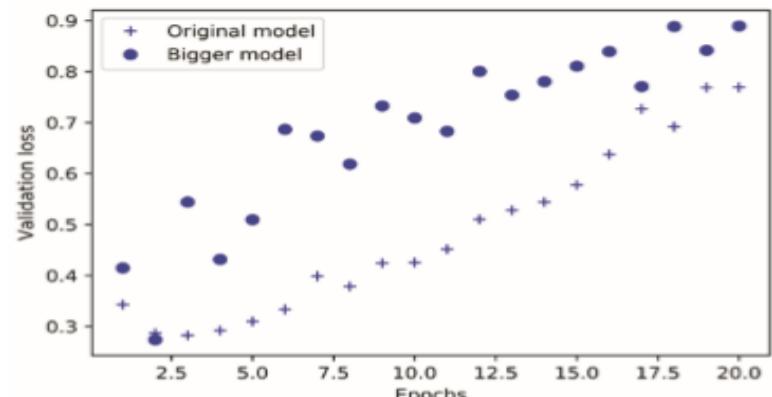
```
model = models.Sequential()
model.add(layers.Dense(4, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(4, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```



- Si se agranda el modelo, el *overfitting* empeora y la pérdida es más errática ...

Modelo agrandado (•)

```
model = models.Sequential()
model.add(layers.Dense(512, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(512, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```



Corregir overfitting: regularización

- El objetivo es ajustar el error con un factor de castigo para que se altere el cálculo de los pesos.
Por ejemplo, sea el error cuadrático medio para ajustar una red:

$$E = \frac{1}{2}(y - \hat{y})^2 \quad \text{Se agrega el castigo} \quad E = \frac{1}{2}(y - \hat{y})^2 + L_Norma$$

Lo común es agregar la L1-Norma y L2-Norma para las regularizaciones L1 y L2 respectivamente

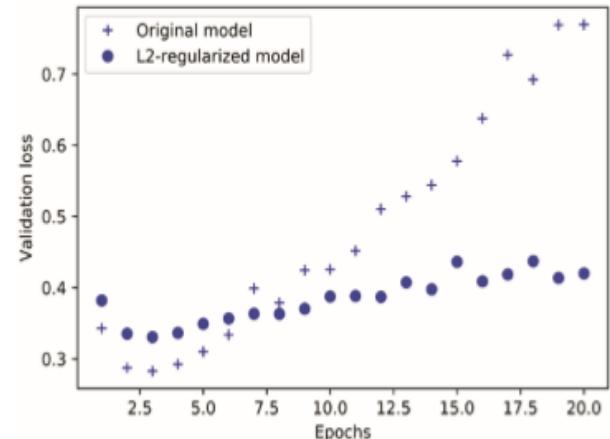
$$L1 = \lambda \sum_{i=1}^n |w_i| \quad L2 = \lambda \sum_{i=1}^n w_i^2 \quad \text{donde } \lambda \text{ es el parámetro de regularización}$$

Visto más general: $E' = \frac{1}{2}(y - \hat{y})^2 + \lambda \sum_{i=1}^n |w_i|$ $E' = \frac{1}{2}(y - \hat{y})^2 + \lambda \sum_{i=1}^n w_i^2$

- En consecuencia al calcular los nuevos pesos con $w^{k+1} = w^k + \eta \frac{\partial E'}{\partial w^k}$ será sobre el error regularizado ...

```
from keras import regularizers

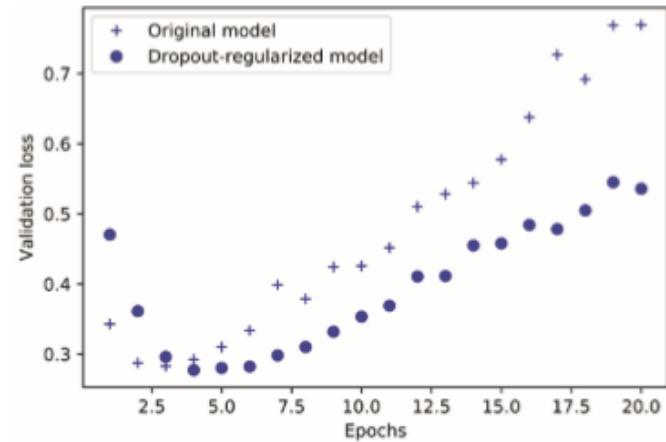
model = models.Sequential()
model.add(layers.Dense(16, kernel_regularizer=regularizers.l2(0.001),
                      activation='relu', input_shape=(10000,)))
model.add(layers.Dense(16, kernel_regularizer=regularizers.l2(0.001),
                      activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```



Corregir Overfitting: dropout

- Al igual que la regularización, se aplica durante la fase de entrenamiento. El objetivo es deshabilitar, aleatoriamente, algunas neuronas, en una o varias capas, durante las fases *forward* y *backward*. La tasa de *dropout*, para anular pesos, más común es entre 0.2 y 0.5, es decir, entre 20% a 50% de los pesos de la capa siguiente se llevan a cero.

```
model = models.Sequential()  
model.add(layers.Dense(16, activation='relu', input_shape=(10000,)))  
model.add(layers.Dropout(0.5))  
model.add(layers.Dense(16, activation='relu'))  
model.add(layers.Dropout(0.5))  
model.add(layers.Dense(1, activation='sigmoid'))
```



- Recapitulando, las tres maneras de corregir el *overfitting* son:

1. Reducir el tamaño de la red neuronal, con ajuste, gracias al conjunto de validación
2. Modificar el cálculo del error de la predicción, mediante L-Normas, para ajustar los pesos durante el *backpropagation*, de tal manera que eviten el *overfitting*.
3. Eliminar neuronas, aleatoriamente, por lo que los pesos en la siguiente capa se desactivan y se reduce la capacidad de memorización de la red neuronal.

Ejemplo en Keras

IMDB es una base de datos de 50,000 registros de texto sobre opiniones de películas donde 50% son reportes positivos y 50% negativos. El texto está transformado en secuencia de enteros que se asocian a palabras

- El primer paso es cargar la base de datos en una matriz de enteros con sólo las 10,000 palabras más frecuentes. Las etiquetas discriminan sólo dos clases: 0 y 1 (crítica positiva o negativa)

```
from tensorflow.keras.datasets import mnist
(train_images, train_labels), (test_images, test_labels) = mnist.load_data()
```

- Los datos se transforman en una matriz dispersa de 0 y 1 donde las filas representan una crítica y se tienen 10,000 columnas representando la ausencia o presencia de una palabra en esa opinión. Además se transforman las etiquetas

```
def vectorize_sequences(sequences, dimension=10000):
    results = np.zeros((len(sequences), dimension))
    for i, sequence in enumerate(sequences):
        for j in sequence:
            results[i, j] = 1.
    return results
x_train = vectorize_sequences(train_data)
x_test = vectorize_sequences(test_data)
```

```
y_train = np.asarray(train_labels).astype("float32")
y_test = np.asarray(test_labels).astype("float32")
```

Entrenamiento IMDB

- Se define la red con sigmoide como función de activación de la capa de salida por tratarse de un clasificador binario. Las capas intermedias, como es habitual, tiene activación ReLU.

```
model = keras.Sequential([
    layers.Dense(16, activation="relu"),
    layers.Dense(16, activation="relu"),
    layers.Dense(1, activation="sigmoid")
])
```

- La función de pérdida o error debe ser la entropía binaria cruzada, por ser un clasificador a dos clases y se medirá la precisión clásica (*accuracy*)

```
model.compile(optimizer="rmsprop",
              loss="binary_crossentropy",
              metrics=["accuracy"])
```

- Se define un conjunto de validación para detectar sobreajuste y las condiciones del entrenamiento fijando el procesamiento por lotes (grupo de ejemplos por pasada de descenso de gradiente) y la cantidad de épocas o generaciones de aprendizaje

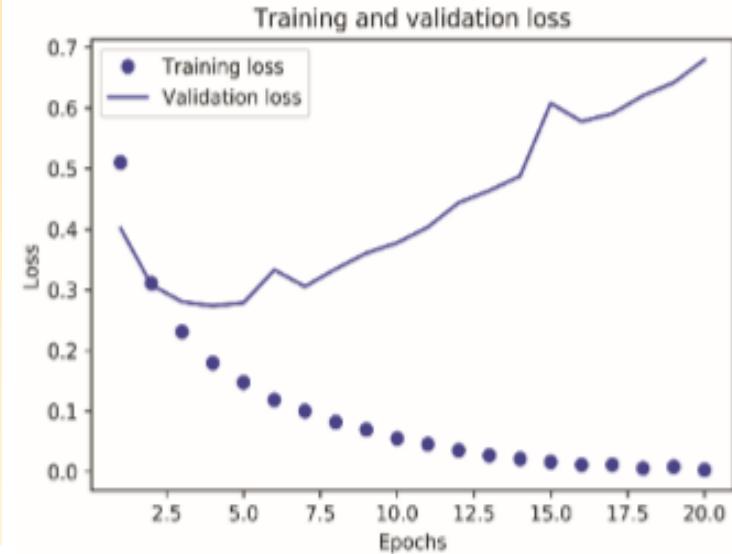
```
x_val = x_train[:10000]
partial_x_train = x_train[10000:]
y_val = y_train[:10000]
partial_y_train = y_train[10000:]
```

```
history = model.fit(partial_x_train,
                     partial_y_train,
                     epochs=20,
                     batch_size=512,
                     validation_data=(x_val, y_val))
```

Evaluación IMDB

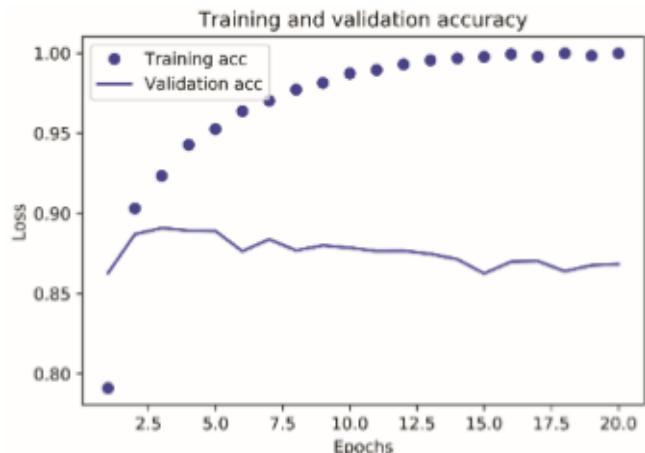
- Primero se visualiza la pérdida entrenando y validando y ... si hay *overfitting* !!!

```
import matplotlib.pyplot as plt
history_dict = history.history
loss_values = history_dict["loss"]
val_loss_values = history_dict["val_loss"]
epochs = range(1, len(loss_values) + 1)
plt.plot(epochs, loss_values, "bo",
         label="Training loss")
plt.plot(epochs, val_loss_values, "b",
         label="Validation loss")
plt.title("Training and validation loss")
plt.xlabel("Epochs")
plt.ylabel("Loss")
plt.legend()
plt.show()
```



- Luego la exactitud entre el conjunto de entrenamiento y validación – Mediocre !!!

```
plt.clf()
acc = history_dict["accuracy"]
val_acc = history_dict["val_accuracy"]
plt.plot(epochs, acc, "bo",
         label="Training acc")
plt.plot(epochs, val_acc, "b",
         label="Validation acc")
plt.title("Training and validation accuracy")
plt.xlabel("Epochs")
plt.ylabel("Accuracy")
plt.legend()
plt.show()
```



IMDB contra el conjunto de prueba

- Se evalúa, después del entrenamiento, con el conjunto de prueba. La exactitud resulta de apenas **88%** ... Los mejores resultados se han logrado con arquitecturas que consiguen un **95%** de *accuracy* ... Justamente con redes neuronales convolucionales ...

```
model = keras.Sequential([
    layers.Dense(16, activation="relu"),
    layers.Dense(16, activation="relu"),
    layers.Dense(1, activation="sigmoid")
])

model.compile(optimizer="rmsprop",
              loss="binary_crossentropy",
              metrics=["accuracy"])

model.fit(x_train, y_train, epochs=4, batch_size=512)
results = model.evaluate(x_test, y_test)
```

- Claramente esta es una solución que debe trabajarse más detalladamente con regularización y/o *dropout* para corregir el problema de *overfitting*. Otro [ejemplo simple](#).
- Además para análisis de sentimientos hay otras arquitecturas de redes neuronales que procesan mejor este tipo de información como las redes neurales recurrente, específicamente, LSTM y/o redes neuronales convolucionadas 1D.

Pasos previos al entrenamiento

- Preprocesar los datos para cambiar o eliminar atributos sin valor (NaN) mediante un valor por defecto o asignar la media o eliminar el registro, ... También está el problema de los atributos con valores atípicos (*outliers*) que se podría solucionar con normalización o también mediante eliminación del registro ...
- Vectorizar o pasar a matrices multidimensionales, también conocidos como **tensores**. Los datos deben ser valores pequeños entre [0,1] o [-1,1] y homogeneizada
- Una vez que los datos están adaptados a las exigencias de las librerías, se selecciona el tipo de red neuronal y se fijan los hiperparámetros a entonar manualmente (capas y neuronas ...)
- Por último se define el optimizador, la función de activación y pérdida. Estas dos últimas, se deben seleccionar en función al problema que se pretende resolver

Tipo de problema	Activación última capa	Función de perdida o error
Clasificación binaria	Sigmoide	binary_crossentropy
Clasificación multiclase exclusiva	Softmax	categorical_crossentropy
Clasificación multiclase no exclusiva	Sigmoide	binary_crossentropy
Regresión para valores arbitrarios	Lineal	Error cuadrático medio (mse)
Regresión para valores entre 0 y 1	Sigmoide	Error cuadrático medio (mse) o binary_crossentropy

Potencia de Cálculo de una Red Neural Artificial



Funciones Booleanas Combinatorias:

Pueden ser representadas por una red neural a una sola capa intermedia ya que basta con el NAND para obtener cualquier circuito digital combinatorio.



Funciones Continuas:

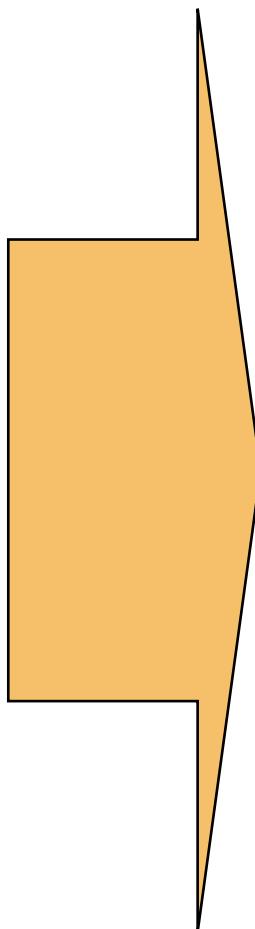
También pueden ser representadas por una red neuronal con al menos dos capas, con precisión arbitraria, a condición que la función de activación, en las capas intermedias sea una función no lineal como ReLU y en la capa de salida lineal sin umbral.



Funciones Arbitrarias n -dimensional:

Cualquier función puede ser aproximada, con precisión arbitraria, con suficientes capas ocultas. Demostrado por el Teorema de Cybenko en 1988. Un teorema más reciente prueba que una red con activación ReLU y al menos $n+1$ capas aproxima cualquier función continua n -dimensional

Consideraciones en una Red Neuronal Artificial



¿ Cuantas unidades o neuronas artificiales ?

Configurable valiéndose del conjunto de validación

¿ Tipo de neurona ?

Depende de la capa (intermedia o salida) y el problema

¿ Topología de la red ?

Tantas capas intermedias como sea posible sin overfitting

¿ Inicialización de los pesos ?

Aleatorio

¿ Número de ejemplos para el entrenamiento ?

Determinado por la cantidad de pesos y cuidando el overfitting

¿ Cómo codificar los datos de entrada y salida ?

Binario es lo común aunque se pueden usar matrices dispersas con valores reales o enteros (aún si se trata texto)

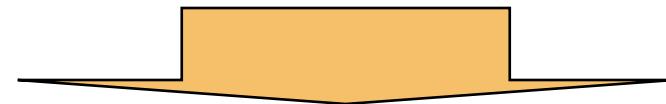
Limitaciones Generales de las Redes Neuronales

- ¿ Cual tipo de red conviene para resolver el problema ?



Se hace empíricamente, en función a problemas parecidos lo cual es muy cuestionable desde el punto de vista formal

- El tiempo de aprendizaje crece exponencialmente por lo que se debe seleccionar adecuadamente el optimizador y considerar indispensable el procesamiento paralelo
- No hay transparencia pues usa enfoque caja negra que impiden saber con certeza como trabaja la red neuronal una vez entrenada



No tienen capacidad de explicación

- Conocimiento a priori no puede ser bien aprovechado en las redes neuronales superficiales. En las redes neuronales profundas es posible dependiendo del tipo de red.