03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

1 DEEP LEARNING – Funciones de Activación

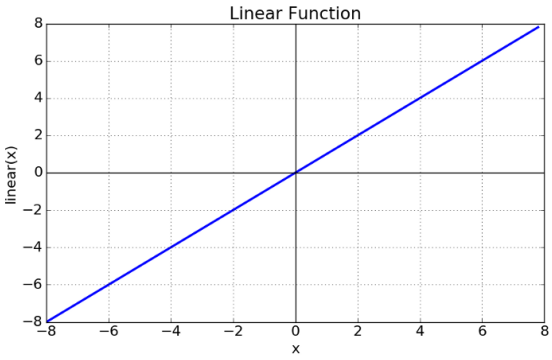
1.1 Funciones de activación:

La función de activación devuelve una salida que será generada por la neurona dada una entrada o conjunto de entradas. Cada una de las capas que conforman la red neuronal tienen una función de activación que permitirá reconstruir o predecir.

Las funciones de activación se dividen en dos tipos como: lineal y no lineal

1.1.1 Función lineal

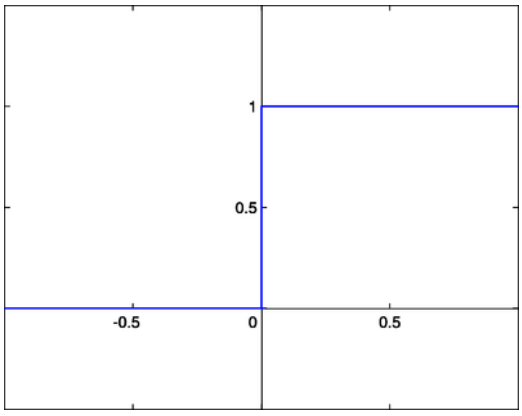
Esta función también conocida como identidad, permite que lo de la entrada sea igual a la salida por lo que si tengo una red neuronal de varias capas y aplicó función lineal se dice que es una regresión lineal. Por lo tanto, esta función de activación lineal se usa si a la salida se requiere una regresión lineal y de esta manera a la red neuronal que se le aplica la función va a generar un valor único. Por ejemplo, se usa cuando se solicita predecir el valor de un número de ventas.



1.1.2 Funciones no lineales

1.1.2.1 Función Escalón

Indica que si la x es menor que cero la neurona va a ser cero pero cuando es mayor igual a cero dará como salida igual 1. Esta función se usa cuando se quiere clasificar o cuando se tiene salidas categóricas. Por ejemplo, se puede usar para predecir si compro algo o no. En el caso de clasificación múltiple no sirve, ya que puede dar 1 en varios casos.

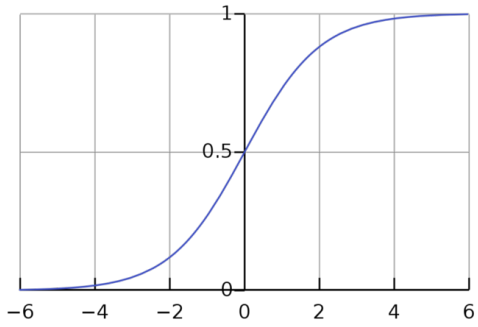


03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

1.1.2.2 Función Sigmoide

Nos da resultado entre 0 y 1.

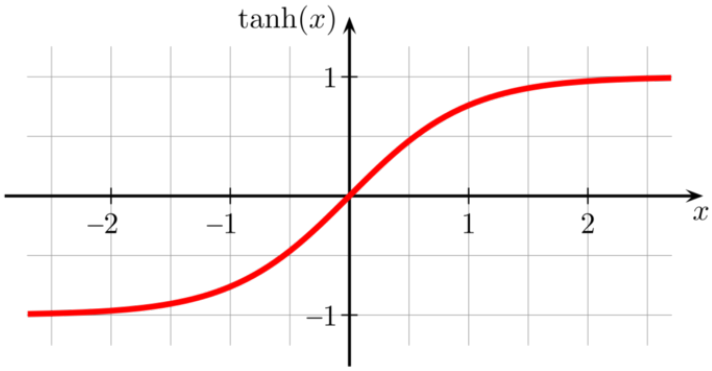
$$f(x) = \frac{1}{1 - e^{-x}}$$



Esta función está en un rango de valores de salida está entre cero y uno por lo que la salida es interpretada como una probabilidad. Si se evalúa la función con valores de entrada muy negativos, es decir x<0 la función será igual a cero, si se evalúa en cero la función dará 0.5 y en valores altos su valor es aproximadamente a 1. Por lo que esta función se usa en la última capa y se usa para clasificar datos en categorías.

1.1.2.3 Función tangente hiperbólica

$$f(x) = \frac{2}{1 + e^{-2x}} - 1$$

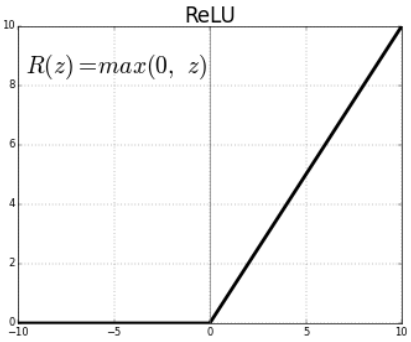


Esta función tiene un rango de valores de salida entre -1 y 1. Se dice que está función es un escalamiento de la función logística, por lo que a pesar que esta función está centrada tiene un problema similar a la sigmoide debido al problema de desaparición del gradiente, que se da cuando en el entrenamiento se genera un error con el algoritmo de propagación hacia atrás y debido a esto el error se va propagando entre las capas, por lo que en cada iteración toma un valor pequeño y la red no puede obtener un buen aprendizaje.

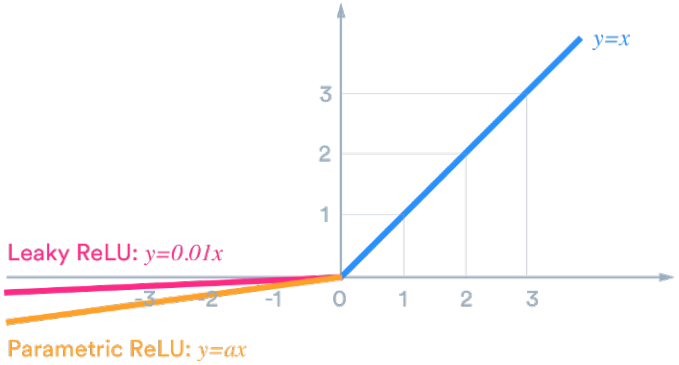
En las capas intermedias es mejor usar la función de activación **tanh** que el sigmoide. Pero en la capa final, usar sigmoide.

03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

1.1.2.4 Función RelU



Está función es la más utilizada debido a que permite el aprendizaje muy rápido en las redes neuronales. Si a esta función se le da valores de entrada muy negativos el resultado es cero, pero si se le da valores positivos queda igual y además el gradiente de esta función será cero en el segundo cuadrante y uno en el primer cuadrante. Cuando se tiene que la función es igual a cero y su derivada también lo es se genera lo que es la muerte de neuronas, a pesar de que puede ser un inconveniente en algunos casos permite la regularización Dropout. Por esta razón la función ReLu tiene una variante denominada Leaky ReLu que va a prevenir que existan neuronas muertas debido a la pequeña pendiente que existe cuando x<0.



Para capas intermedias, si no se está seguro de qué activación usar, usa RelU

1.1.2.5 Función Softmax

Esta función se usa para clasificar datos, por ejemplo si le damos de entrada la imagen de una fruta y se solicita saber el tipo de fruta a que pertenece, aplicando softmax la red nos dará la probabilidad de que puede ser 0.3 o 30% melón, 0.2 o 20% sandía y 0.5 o 50% papaya, por lo que nos el resultado será el que tenga mayor probabilidad y cabe recalcar que la suma de estas probabilidades será igual a 1. En otras palabras, Softmax se usa para clases múltiples y cuando se va a asignar probabilidades a cada clase que pertenezca a clases múltiples.

La función softmax es una generalización de la regresión logística que puede ser aplicada a datos continuos.

1.2 Resumen:

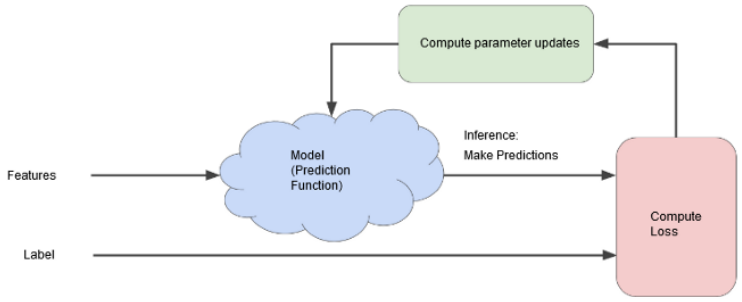
Para clasificaciones binarias, sigmoide y para capas intermedias RelU, pero para cada problema puede ser diferente y el resultado puede variar dependiendo de qué función de activación se escoge.

03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa Tel 966870140 Fax 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

2 Entrenamiento y pérdida

El objetivo de una red neuronal es aprender comportamientos que permitan destacar características de los objetos que se están tomando como input, todo este proceso es llamado Entrenamiento, sin embargo, es casi imposible obtener un cien por ciento de predicciones correctas, por lo que a la penalidad por una predicción incorrecta se la llama Pérdida. El que exista pérdida en nuestro modelo no es malo, ya que en caso de no tenerla es muy posible que se esté realizando **overfitting**, *es decir que nuestro modelo es incapaz de generalizar.*

El objetivo del entrenamiento es medir la pérdida para luego ir iterando hasta que el algoritmo descubra los parámetros del modelo con la pérdida más baja posible una vez que esto ocurra el modelo ha convergido.

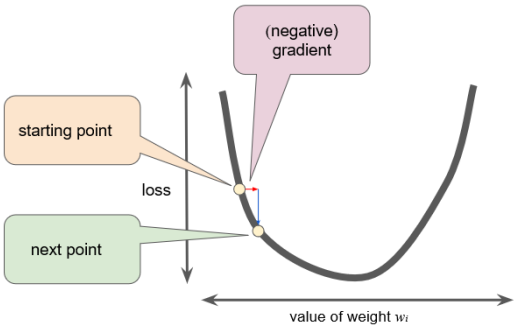


2.1 Reducción de la pérdida

Para poder reducir la pérdida es necesario modificar los parámetros (pesos) y volver a recalcular el modelo tal como lo indica la figura anterior.

Uno de los métodos más efectivos para determinar los parámetros con los que obtendremos la menor pérdida sería **calcular la pérdida con cada uno de los pesos posibles y luego determinar el punto en el que la pérdida es mínima,** pero esto conlleva un muy largo tiempo y demasiado poder computacional, por lo que **se usa el algoritmo de descenso de gradientes.**

El primer paso en el algoritmo de descenso de gradientes es escoger un punto inicial, luego se calculará el gradiente de la curva de pérdida en el punto escogido, para poder escoger el siguiente punto se multiplica la gradiente encontrada por un escalar llamado tasa de aprendizaje que se encuentra entre los valores 0.0 y 1.0.



Este escalar es un hiper parámetro muy importante, usualmente el valor de este es 0.05, controla qué tan rápido el modelo se adapta al problema.