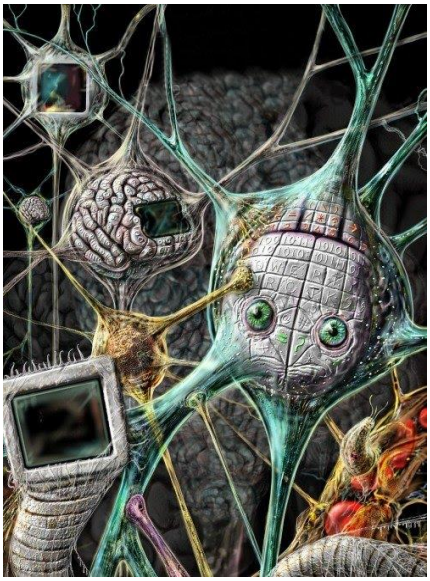
Inteligencia Computacional

Unidad 5: Redes Neuronales



Las redes neuronales han vuelto al protagonismo en la actualidad debido al avance de la capacidad de cómputo en el hardware y de los logros que se están consiguiendo con ello. A pesar de su nombre, las redes neuronales no tienen un concepto demasiado complicado detrás de ellas. El objetivo es imitar el funcionamiento de las redes neuronales de los organismos vivos: un conjunto de neuronas conectadas entre sí que trabajan en conjunto, sin que haya una tarea concreta para cada una. Con la experiencia, las neuronas van creando y reforzando sus conexiones para "aprender" algo que se queda fijo en el tejido.

Las redes neuronales artificiales son un modelo matemático para encontrar las mejores conexiones (parámetros) para lograr un objetivo. En el lenguaje propio, encontrar los parámetros que mejor ajustan es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" lo que aprendió. Desde el simple perceptrón hasta las redes convolucionales que dieron lugar al avance del aprendizaje profundo y gracias al avance tecnológico, las redes neuronales han colaborado con áreas desde la medicina y el comercio hasta las redes sociales.

1. Identifique aplicaciones concretas de las Redes Neuronales. Según lo visto hasta el momento ¿Qué tipos de modelos son capaces de implementar (regresión, clasificación, clustering)? ¿Cuál es la parametrización requerida en cada caso?

2. **RNA Perceptrón Simple:**
   a) Describa las funciones de activación escalón, lineal y sigmoidea.
   b) Caracterice sus parámetros.
   c) Describa la regla de aprendizaje del Perceptrón Simple. Dé un ejemplo.

d) Implementar las funciones lógicas AND y OR. Pensarlo en papel y luego implementarlo en software.
e) ¿Qué problema sucede con la OR EXCLUSIVA? Explicar las razones por las que ocurre.

3. Describa conceptualmente, de manera que quede bien claro, los siguientes términos: **Sobreajuste, generalización, validación**.

4. Estudie técnicas para evitar sobreajuste.

https://deep-learning.ikor.org/entrenamiento/regularizaci%C3%B3n

https://elvex.ugr.es/decsai/deep-learning/slides/NN5%20Regularization.pdf

5. Utilice la herramienta de Google "Playground" para verificar lo aprendido hasta el momento: https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises?hl=es-419

6. **Aproximación de funciones (regresión)**:
a) Diseñe e implemente una RNA para aproximar el valor de las funciones.
b) Evalúe la performance de la red.

$f(x) = cos(a) + sen(b)$ en el intervalo $[-\pi, +\pi]$.
$f(x) = \frac{sin(2x)}{\exp\left(\frac{x}{5}\right)}$

7. **Clasificación** de páginas web según su idioma
a) Los conjuntos de datos *textck.mat* y *textpt.mat* contienen dos vectores: P (entradas = cantidad de letras c/k o p/t) y T (salidas deseadas = idioma). En ellos figuran los porcentajes de letras, 'c' y 'k' en un archivo y 'p' y 't' en otro. Analice el uso de la herramienta *regularización* para mejorar el entrenamiento.
b) Presente una tabla donde figure el error obtenido en diferentes ensayos y entrenamientos. Se sugiere variar las funciones de activación, la arquitectura, el método y velocidad de aprendizaje y los grupos de datos presentados a la red.
c) Evalúe la red utilizando los métodos de medición del error y análisis de *performance*.
d) Elija una arquitectura de red y justifique su elección teniendo en cuenta el sobreentrenamiento y el error de consulta.
e) Una vez elegida la arquitectura utilice la técnica *K-fold y Bootstrap (error .632)* para brindar una estimación del error que cometerá la red en la etapa de simulación o consulta.

8. **Clasificación de pacientes con riesgo hepático:**
El conjunto de datos acerca del estado de hígado que se encuentra en el repositorio *UCI Machine Learning Repository* contiene 416 registros de pacientes hepáticos y 167 registros de pacientes no hepáticos de pacientes masculinos y femeninos. Se recopiló en el noreste de Andhra Pradesh, India.
https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset
Cada clase posee una etiqueta que se usa para dividirlo en dos clases (pacientes con riesgo hepático o no). Contiene 10 variables que son edad, género, bilirrubina total, bilirrubina directa, proteínas totales, albúmina, relación A / G, SGPT, SGOT y Alkphos. Replique los pasos del ejercicio anterior para realizar un clasificador para estos datos.
¿Cómo se utilizaría luego? ¿Cómo analiza cuán confiable es una predicción de este clasificador para detectar casos positivos (riesgo presente) y negativos (sin riesgo)?

9. **Precio de casas:**

La propuesta en este problema es construir un algoritmo que prediga el valor de las casas en ciertos barrios de Boston, en función de una serie de variables que contienen información sobre características de las casas y características de la zona.

Los datos y toda la información sobre el dominio están disponibles en el ***UCI Machine Learning Repository.*** http://archive.ics.uci.edu/ml/datasets/Housing.

En el archivo **housing.data** se encuentran todos los datos disponibles. Se componen de 13 variables de entrada, 12 de ellas toman valores continuos y una, valor binario. La última columna es la variable de salida que representa el precio medio de las casas con las características indicadas en las variables de entrada.

Se requiere utilizar una Red Neuronal de forma que, a partir de los ejemplos disponibles, se obtenga un sistema que sea capaz de predecir a partir de un dato de entrada el precio de una determinada casa.

Diseñe, implemente y evalúe la red para lograr este objetivo. Se debe plantear una arquitectura y parámetros justificando a través de índices la elección. Tenga en cuenta la inicialización, la normalización, la aleatorización y la formación de conjunto de datos.

10. **Predicción de series temporales:**

El archivo *Datosseriesautos* contiene datos de la producción de automóviles de cierto país, desde enero de 1999 hasta abril de 2006, tomados mensualmente. Dada la serie temporal que este archivo contiene:

a) Armar la línea temporal correspondiente y programar un algoritmo para crear el conjunto de datos de entrenamiento a fines de predecir valores futuros mediante un Perceptrón Multicapa, utilizando solo los datos hasta 2005. Se debería poder elegir cuántas entradas se considerarán.
b) Dividir adecuadamente el conjunto de datos en datos de validación y de test para entrenar la red. Entrenar y predecir los valores del 2006.
c) Comparar los obtenidos con los verdaderos de la serie hasta abril. Calcular el error cometido. ¿De los valores obtenidos, cuáles son los más confiables?

11. **Redes Convolucionales:**

Ajuste los parámetros y evalúe una CNN para aplicarla al reconocimiento de imágenes del siguiente conjunto. Agregar ruido a las imágenes y evaluar *performance*.

http://yann.lecun.com/exdb/mnist/

12. **Redes no supervisadas (SOM)**

a) Ver el material propuesto por Kohonen: http://www.cis.hut.fi/somtoolbox/theory/somalgorithm.shtml
b) Indique los pasos necesarios para implementar y evaluar una RNA no supervisada.
c) Describa el entrenamiento de un SOM.
d) ¿Qué hiperparámetros requieren y cuáles podrían ser sus valores por defecto?
e) ¿Qué características deben tener los datos de manera de contribuir con un adecuado entrenamiento?
f) Describa los métodos de inicialización propuestos. ¿Cuáles son las ventajas de cada uno de ellos? ¿En qué casos se utilizan?
g) ¿Cómo se mide el error de una red no supervisada? ¿Qué métodos hay y cuándo se utilizan? Justifique.

13. **Redes No supervisadas (SOM)**

Descargar una librería de SOM y trabajar con los datos contenidos en el archivo IRIS.DAT.
http://www.cis.hut.fi/somtoolbox/

a) Investigar los parámetros de la función de MATLAB® 'newsom'. Crear un mapa autoorganizado y luego graficar sus celdas con 'plotsom' para ver la topología. Si se prefiere utilizar la Toolbox de Helsinki,

explorar las funciones 'som_make' y 'som_show'. Si no utiliza MATLAB, buscar funciones similares en su librería.

b) Definir un mapa de Kohonen con el fin de visualizar la agrupación que permite clasificar los mismos datos del problema anterior.

c) Graficar las BMU de los datos de entrenamiento con distintos colores para las diferentes clases. Observar la ubicación en el mapa luego del entrenamiento. Las funciones de la Toolbox para MATLAB son 'som_bmus' y 'som_show_add'.

d) Comparar conceptualmente este método de clasificación con el Perceptrón Multicapa o con Kmeans. ¿Qué ventajas ofrece el mapa autoorganizado?

14. En la base de datos del repositorio UCI se almacenan diversos descriptores de geometría de tres clases diferentes de trigo (Kama, Rosa y Canadiense):
https://archive.ics.uci.edu/ml/datasets/seeds

a) Visualice los datos e indique la dimensión y características de sus atributos.

b) Proponga una RNA ***supervisada*** y otra ***no supervisada*** junto con ensayos y métodos de medición del error para clasificar las semillas. Implemente y evalúe estas RNA. Recuerde que el SOM en principio NO ES un clasificador, por requerir entrenamiento no supervisado.

c) Presente una tabla donde figuren los errores de cada prueba realizada.

d) Justifique la elección de la arquitectura y sus parámetros en cada caso de prueba. ¿Cómo son los resultados obtenidos? Compare los resultados de ambas redes.

e) Compare los algoritmos en base a la elección de cada parámetro, el error del resultado y el costo computacional resultado en cada caso.

Problemas Opcionales:

Clasificación de la actividad en personas

Considere los siguientes datos en el repositorio de datos UCI Dataset:

https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Corresponde a experimentos llevados a cabo con un grupo de 30 voluntarios dentro de un rango de edad de 19-48 años. Cada persona realizó seis actividades (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) con un teléfono inteligente (Samsung Galaxy S II) en la cintura. Usando su acelerómetro y giroscopio integrados, se captura la aceleración lineal 3-axial y la velocidad angular 3-axial a una velocidad constante de 50Hz. Los experimentos han sido grabados en video para etiquetar los datos manualmente. Se obtuvo un vector de características calculando variables del dominio del tiempo y la frecuencia. El conjunto de datos obtenido se ha dividido aleatoriamente en dos conjuntos, donde se seleccionó el 70% de los voluntarios para generar los datos de entrenamiento y el 30% de los datos de prueba. Diseñe, implemente y evalúe una RNA para clasificar los datos que se encuentran en el repositorio. Se debe plantear una arquitectura y parámetros justificando a través de índices la elección. Tenga en cuenta la inicialización, la normalización, la aleatorización y la formación de conjunto de datos.

Base de Datos de Animales

Considere los siguientes datos en el repositorio de datos UCI Dataset :

https://archive.ics.uci.edu/ml/datasets/Zoo

Contiene 17 atributos de valor booleano. El atributo "tipo" es un atributo de clase. Clase # - Conjunto de animales:

1 - (41) aardvark, antílope, oso, jabalí, búfalo, ternera, cavy, guepardo, ciervo, delfín, elefante, fruitbat, jirafa, niña, cabra, gorila, hámster, liebre, leopardo, leopardo, lynx, visón mole, mangosta, zarigüeya, oryx, ornitorrinco, polecat, pony, marsopa, puma, gatito, mapache, reno, foca, leña, ardilla, vampiro, vole, wallaby, lobo

2 - (20) pollo, cuervo, paloma, pato, flamenco, gaviota, halcón, kiwi, alondra, avestruz, periquito, pingüino, faisán, rhea, skimmer, skua, gorrión, cisne, buitre, buitre

3 - (5) pitviper, seanake, slowworm, tortuga, tuatara

4 - (13) lubina, carpa, bagre, cacho, pez gato, eglefino, arenque, lucio, piraña, caballito de mar, lenguado, raya, atún

5 - (4) rana, rana, tritón, sapo

6 - (8) pulga, mosquito, abeja, mosca doméstica, mariquita, polilla, termita, avispa

7 - (10) almeja, cangrejo, cangrejo de río, langosta, pulpo, escorpión, avispa marina, babosa, estrella de mar, gusano.

a) Describa las dimensiones, características y tipos de los atributos recolectados:
b) Proponga una RNA ***supervisada*** y otra ***no supervisada*** junto con ensayos y métodos de medición del error para clasificar los animales. Implemente y evalúe estas RNA.
c) Presente una tabla donde figuren los errores de cada prueba realizada.
d) Justifique la elección de la arquitectura y sus parámetros en cada caso de prueba. ¿Cómo son los resultados obtenidos? Compare los resultados de ambas redes.
e) ¿Los datos de con que se implementa y evalúa la red deben contener **clases balanceadas**? Justifique.

Neurona Real vs Neurona Artificial

Describa las similitudes entre la neurona real y la artificial.

https://www.youtube.com/watch?v=e_BOJS1BLj8

Bioética y Confidencialidad de la información

Ver el siguiente video:

https://www.infobae.com/america/tecno/2018/01/30/el-algoritmo-de-google-que-predice-si-un-paciente-morira-o-no-en-base-a-46-mil-millones-de-datos/

¿Cuáles son las ventajas y desventajas del uso de datos? ¿Qué cuestiones se deben cuidar y preservar? ¿Cuál es la potencialidad de la aplicación?

Aprendizaje profundo

https://www.youtube.com/watch?v=FTr3n7uBIuE&t=1878s

¿Qué es el **aprendizaje profundo**? ¿Cuáles son sus requerimientos y aplicaciones actuales?

Google Deep Mind

¿Qué es **Google Deep Mind** y cuáles son sus posibles usos?

Investigue en: https://deepmind.com/

Google Deep Mind

http://sinc.unl.edu.ar/sinc-publications/2017/FRG17/sinc_FRG17.pdf

Describa brevemente, indique las ventajas de **DropOut** y **Mini Batch** y cuáles sus usos y ventajas.

Libro gratuito de Deep Learning

http://www.deeplearningbook.org/contents/ml.html

Playground de Google

http://playground.tensorflow.org

Tensorflow de Google

¿Cuál es la filosofía de Tensorflow?
¿Qué son los tensores? ¿Que representan? ¿Qué algoritmos tienen implementados?