

Clasificación de pingüinos Palmer usando clasificador perceptrón multicapa

Escalante Hernandez Alejandro
alejandro.escalante@alumno.buap.mx

Resumen El clasificador perceptron multicapa ha sido ampliamente utilizado en problemas de clasificación, debido a sus ventajas frente a un perceptron simple. En este trabajo se utiliza un clasificador perceptron multicapa para clasificar 3 especies de pingüinos tomando en cuenta como atributos el largo del pico (bill length) y largo del ala (flipper length).

I. INTRODUCCIÓN

Una red neuronal artificial es un modelo computacional inspirado en una red neuronal. Constituida por un conjunto de neuronas (Unidades que reciben datos, procesan según su configuración, y generan un resultado, estas también almacenan una configuración (pesos)) interconectadas. Cada red neuronal debe ser entrenada para que cumpla su objetivo, dicho entrenamiento modifica los pesos de cada neurona, consiguiendo así que el resultado final mejore.

El modelo de Perceptrón consiste en una capa de neuronas con pesos y umbral ajustables, este sistema puede ser llamado una red neuronal debido a las conexiones existentes en sus elementos. El desarrollo del perceptron llevó a la generación de un nuevo tipo de red cuya modificación principal respecto al perceptron se basa en el uso de varias capas de neuronas artificiales en vez de una sola capa. Además del cambio de la función de activación a una función diferenciable y no lineal como la sigmoide. Este tipo de función de activación introdujo un nuevo paradigma en el procesamiento de los sistemas neuronales permitiendo a las redes neuronales aprender las variaciones no lineales de los distintos tipos de ambientes.

II. OBJETIVO

En este trabajo se implementará y entrenará un clasificador perceptrón multicapa para clasificar los datos de una base de datos determinada, con el fin de obtener un buen porcentaje de precisión. Siendo esta entrenada con un 80% del total de los datos para posteriormente realizar la clasificación de la base de datos completa.

III. BASE DE DATOS

III-A. Descripción

Esta base de datos contiene la información de distintas características de los pingüinos Palmer los cuales pueden ser de las siguientes especies.

- Adelie
- Chinstrap
- Gentoo

Esta base de datos seleccionada contiene 7 atributos de los cuales se usarán para la clasificación sólo 2 debido a su relevancia, 3 clases, y 345 instancias.

Atributos Totales

- island
- bill_length_mm
- bill_depth_mm
- flipper_length_mm
- body_mass_g
- sex
- year

Atributos utilizados en la clasificación

- bill_length_mm
- flipper_length_mm

III-B. Lectura y limpieza de los datos

Ya que la base de datos está disponible para instalarse en R, para poder realizar la lectura en Python, se realizó un programa en R que exporta los datos a un archivo .csv.

Posteriormente la lectura se realizó a un data frame utilizando pandas, para después filtrar en el dataframe únicamente los atributos que se tomarían en cuenta en la clasificación, a su vez se eliminaron aquellas instancias con valores nulos.

Por último se realizó la codificación de las etiquetas de las clases y se escalaron los valores de los atributos.

IV. CLASIFICADOR PERCEPTRON MULTICAPA

En esta sección se describen las partes importantes de la implementación del clasificador.

IV-A. Parametros

Para instanciar el clasificador son necesarios multiples parametros para configurarlo. Para este trabajo se consideraron los parametros presentados a continuación, lo cual implica que los demás parametros conservaron los valores predeterminados.

- Solver : lbfgs
- Hidden layer Sizes : (3,5,3)
- Activation : relu
- MaxIter: 3000

Se optó por estos valores debido a las siguientes razones:

lbfgs es un optimizador de la familia de métodos cuasi-Newton. No es el solver predeterminado, sobretodo porque adam el solver predeterminado obtiene mejores resultados con grandes conjuntos de datos, sin embargo el conjunto de datos para este trabajo no es muy grande.

La configuración de los tamaños de las capas está dada de la siguiente manera el primer numero es el numero de atributos mas uno por el bias, el último es la cantidad de clases y el tercero esta dado por la suma de el primero y ultimo menos uno para que no sea el doble del primer valor.

La función de activación Relu o unidad lineal rectificada es recomendada para problemas de clasificación.

El limite maximo de iteraciones fue cambiado debido a que el valor predeterminado era muy bajo y en la mayoría de ocasiones se alcanzaba el numero maximo sin alcanzar la convergencia.

IV-B. Entrenamiento

Para realizar el entrenamiento del clasificador se utilizó un 80 % del conjunto de datos

IV-C. Medición de precisión

Con el fin de medir la precisión del clasificador se hizo uso de una matriz de confusión para posteriormente dividir la suma de la diagonal de la matriz entre la suma de todos los elementos, obteniendo así el porcentaje de la precisión.

V. EJECUCIÓN

Se realizo un conjunto de 10 ejecuciones del clasificador con las configuraciones mencionadas anteriormente obteniendo los siguientes resultados:

V-A. Predicción

Los resultados de predicción son los siguientes:

Resultados = [0.9649, 0.9707, 0.8274, 0.9502, 0.9619, 0.7953, 0.9707, 0.9736, 0.9678, 0.9590]

- Media aritmética (): 0.93415
- Mediana: 0.9634
- Moda: 0.9707
- Menor valor: 0.7953
- Mayor valor: 0.9736
- Rango: 0.1783
- Varianza (2): 0.0038626065000003
- Desviación estándar (): 0.062149871279032

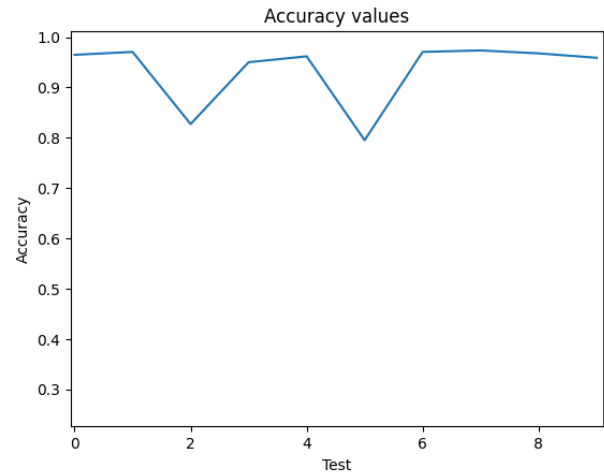


Figura 1. Valores de precisión del clasificador

VI. DISCUSIÓN

A través del análisis de los resultados obtenidos en las multiples ejecuciones del programa se puede concluir que los resultados son lo suficientemente buenos en precisión teniendo en cuenta el rango en el cual se encuentran los resultados y su distribución.

VII. REFERENCIAS

Mejías César, Yuleidys, Carrasco Velar, Ramón, Ochoa Izquierdo, Isbel, Moreno Lemus, Edel. (2013). Funciones de transferencia en el perceptrón multicapa: efecto de su combinación en entrenamiento local y distribuido. Revista Cubana de Informática Médica, 5(2), 186-199. Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18592013000200009&lng=estlng=es.