Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ciencias y Sistemas

Inteligencia Artificial 1

Vacaciones de diciembre de 2020

Catedrático: Ing. Jorge Gutiérrez

Auxiliar: Nery Galvez

Proyecto

Tabla de Contenido

Objetivos

Objetivo General

Implementar una aplicación web donde se apliquen los conceptos adquiridos en el laboratorio de Inteligencia Artificial 1.

Objetivos Específicos

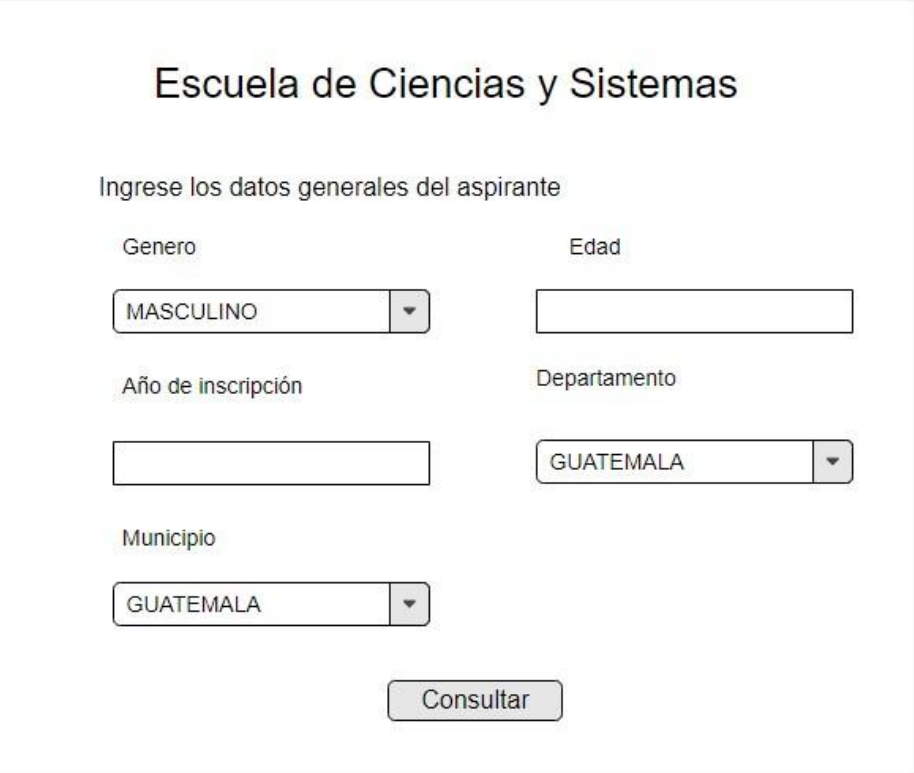
- Implementar una red neuronal que permita realizar predicciones sobre un conjunto de datos.
- Implementar un algoritmo genético que permita seleccionar la mejor configuración de híper parámetros para una red neuronal.

Descripción de la Solución

El prototipo consistirá en una pequeña página web que solicitará la información del aspirante y mostrará si la persona se cambiará en un futuro de carrera o no. La información que se solicitará es la siguiente:

- Genero
- Edad
- Año de inscripción
- Departamento
- Municipio

El prototipo no realizará ningún entrenamiento, sino que ya únicamente se encargará de predecir los resultados. Por lo que se podrán realizar N cantidad de pruebas seguidas sin ningún problema. A continuación, se muestra una posible interfaz de la solución.



Red Neuronal

Conjunto de datos para la red neuronal

Para entrenar el algoritmo se utilizará el conjunto de datos proporcionado junto con el enunciado, el cual es el siguiente:

	E	F	H	J	K	L	M	P
1	Estado	Genero	edad	cod_depto	nombre	cod_muni	municipio	Año
2	Activo	MASCULINO	67	1	Guatemala	1	Ciudad de Gu	2020
3	Traslado	MASCULINO	39	1	Guatemala	1	Ciudad de Gu	2015
4	Activo	MASCULINO	50	1	Guatemala	64	Villa Nueva	2010
5	Activo	MASCULINO	51	1	Guatemala	66	San Miguel P	2012
6	Traslado	MASCULINO	57	1	Guatemala	1	Ciudad de Gu	2011
7	Activo	MASCULINO	46	1	Guatemala	57	Mixco	2010
8	Traslado	FEMENINO	54	3	Sacatepeque	8	San Lucas Sa	2014
9	Activo	MASCULINO	48	1	Guatemala	1	Ciudad de Gu	2013
10	Traslado	MASCULINO	44	1	Guatemala	1	Ciudad de Gu	2010
11	Traslado	MASCULINO	48	1	Guatemala	65	Villa Canales	2013
12	Traslado	MASCULINO	48	1	Guatemala	1	Ciudad de Gu	2010
13	Traslado	MASCULINO	54	1	Guatemala	1	Ciudad de Gu	2020
14	Traslado	MASCULINO	46	1	Guatemala	1	Ciudad de Gu	2011
15	Traslado	MASCULINO	54	1	Guatemala	57	Mixco	2019
16	Traslado	MASCULINO	44	1	Guatemala	57	Mixco	2010
17	Traslado	MASCULINO	44	1	Guatemala	64	Villa Nueva	2010
18	Activo	MASCULINO	51	1	Guatemala	1	Ciudad de Gu	2015
19	Traslado	MASCULINO	44	1	Guatemala	1	Ciudad de Gu	2011
20	Traslado	FEMENINO	47	1	Guatemala	1	Ciudad de Gu	2011
21	Traslado	FEMENINO	50	1	Guatemala	64	Villa Nueva	2012
22	Activo	MASCULINO	51	1	Guatemala	1	Ciudad de Gu	2014

y **x**

De todos los datos proporcionados el estudiante deberá realizar la división entre los conjuntos de entrenamiento, validación y prueba de la forma que más crea conveniente.

Tratamiento de datos

A veces los datos de entrada no poseen suficiente información y es necesario generar nuevos datos que aporten mayor información a partir de los datos originales. Por ejemplo, si queremos predecir el precio de una casa en base a las siguientes características:

- Número de habitaciones
- Ancho
- Fondo
- Número de niveles

- Número de baños

Es posible que obtengamos mejores resultados si agregamos nuevas características como el área de la casa, la cual podemos calcular usando el ancho y fondo.

Área = Ancho * Fondo

Observando la información que se posee del aspirante se puede detectar que el departamento y el municipio por si solos no aportan mayor información, por lo que ambos campos se utilizaran para calcular un nuevo campo que es la distancia aproximada del municipio del estudiante con la universidad. El campo se calculará empleando las coordenadas geográficas del municipio con las coordenadas geográficas de la Universidad. Cada coordenada geográfica se compone de dos valores que corresponden la latitud y longitud.

Para calcular la distancia entre dos coordenadas geográficas se utiliza la fórmula de Haversine.

$$D = 2 * R * Asin(\sqrt{Sin^2\left(\frac{\Delta lat}{2}\right) + Cos(lat_1) * Cos(lat_2) * Sin^2\left(\frac{\Delta lon}{2}\right)}$$

Donde:

- $(lat1, lon1)\ coordenadas\ del\ punto\ 1$
- $(lat2, lon2)\ coordenadas\ del\ punto\ 2$
- $R\ radio\ de\ la\ tierra$
- $\Delta lat = lat1 - lat2$
- $\Delta lon = lon1 - lon2$

Pueden basarse en el siguiente ejemplo: http://pythoninicios.blogspot.com/2016/03/como-calcular-la-distancia-entre-dos.html

Para la coordenada de la Universidad utilizaran la siguiente:

(14.589246, -90.551449)

Para las coordenadas de los municipios se les proporcionara un archivo con las coordenadas de los municipios utilizados en el conjunto de datos.

Depto	Muni	Nombre	Lat	Lon
4	11	Acatenango	**14.555038**	**-90.944041**
22	4	Agua Blanca	**14.497781**	**-89.649232**
1	63	Amatitlan	**14.477957**	**-90.630915**
3	1	Antigua Guatemala	**14.556672**	**-90.733692**
22	5	Asuncion Mita	**14.333415**	**-89.711587**
22	7	Atescatempa	**14.175275**	**-89.741092**
6	2	Barberena	**14.305514**	**-90.360918**
19	7	CabaÃ±as	**14.934045**	**-89.796378**
20	5	Camotan	14.8167	-89.3667
9	14	Cantel	14.8167	-91.45
12	16	Catarina	**14.850455**	**-92.077725**
4	1	Chimaltenango	**14.667456**	**-90.820375**
1	55	Chinautla	**14.705282**	**-90.49713**
20	1	Chiquimula	**14.705282**	**-90.49713**
6	8	Chiquimulilla	**14.088689**	**-90.37968**
1	1	Ciudad de Guatemala	**14.628867**	**-90.51179**
3	12	Ciudad Vieja	**14.523103**	**-90.764837**
9	20	Coatepeque	**14.705032**	**-91.865766**
16	1	Coban	**15.467662**	**-90.384487**

Por lo que las entradas del modelo realmente serán:

- Genero
- Edad
- Año de inscripción
- Distancia de la Universidad

Escalamiento de variables

Como las variables de entrada poseen distintos rangos de valores, será necesario realizar un escalamiento de las entradas para evitar que una de las entradas domine a las demás en el entrenamiento. El escalamiento que se implementará será el siguiente:

$$X_i = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$$

El escalamiento solamente se les aplicara a las siguientes entradas:

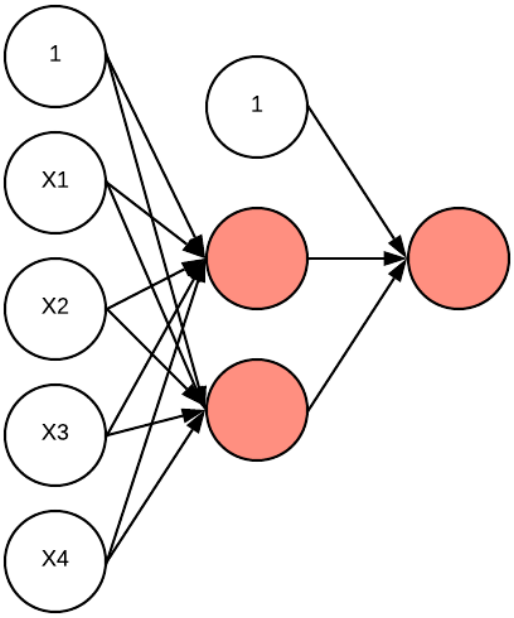
- Edad
- Año de inscripción
- Distancia de la Universidad

Arquitectura del algoritmo

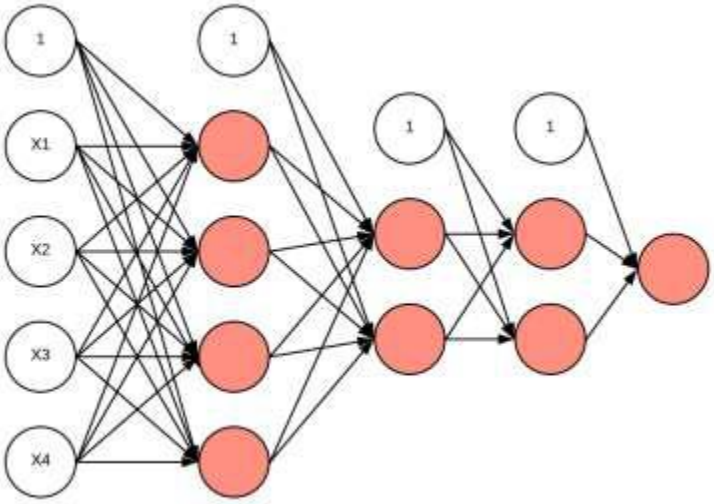
El modelo se obtendrá a partir del entrenamiento de una red neuronal, la arquitectura de esta quedará a discreción del estudiante, solamente tendrá que cumplir los siguientes requisitos:

- Contar con más de 3 capas (sin contar la capa de las entradas).
- Tener en total (por toda la red) más de 18 neuronas.

Por ejemplo, las siguientes redes no son válidas:



Tiene menos de 3 capas



Tiene menos de 18 neuronas

Algoritmo Genético

Se implementará un algoritmo genético para seleccionar la mejor configuración de híper parámetros para la red neuronal que se utilizará para realizar la predicción.

Entrenamiento

Los híper parámetros que se tomaran en cuenta son los siguientes:

- Tasa de aprendizaje
- Tasa de regularización
- Probabilidad de eliminación
- Número máximo de iteraciones

Elegirán 10 combinaciones para los valores de cada híper parámetro, como se muestra en la siguiente tabla:

alpha	lambda	Max_iteration	Keep_prob
0.00001	0	500	1
0.00005	0.01	750	0.97
0.0001	0.05	1250	0.95
0.0005	0.1	1500	0.9
0.001	0.5	2000	0.85
0.005	1	4000	0.75
0.01	1.5	7500	0.65
0.05	2	15000	0.5
0.1	5	20000	0.3
0.5	7	25000	0.1

Nota: Los valores son únicamente de ejemplo.

Se puede representar la configuración que tendrá la red neuronal con solamente un conjunto de índices, por ejemplo:

Individuo = [2, 5, 0, 2]

alpha	lambda	Max_iteration	Keep_prob
0.00001	0	500	1
0.00005	0.01	750	0.97
0.0001	0.05	1250	0.95
0.0005	0.1	1500	0.9
0.001	0.5	2000	0.85
0.005	1	4000	0.75
0.01	1.5	7500	0.65
0.05	2	15000	0.5
0.1	5	20000	0.3
0.5	7	25000	0.1

Nota: Una de las combinaciones debe de cumplir que el valor de lambda sea 0 y donde keep_prob sea 1. Esto porque con estos valores se estaría eliminando la regularización y el Dropout.

Individuo

Corresponde a una configuración posible de los híper parámetros que se encuentran en la tabla anterior, posee la siguiente estructura Individuo = [i1, i2, i3, i4]. Cada elemento del individuo representa un índice en el rango de [0, 9]. Por ejemplo, los siguientes son individuos:

$$Individuo1 = [1, 4, 9, 2]$$

$$Individuo2 = [4, 0, 6, 3]$$

Nota: Cada individuo tendrá una longitud de 4.

Población

Es un conjunto de individuos que solucionan el problema. El tamaño de la población queda a discreción de cada estudiante, teniendo un mínimo de 4 para que el algoritmo tenga bastante variedad de soluciones. La estructura de la población quedaría de la siguiente manera:

P = [

Individuo1,

Individuo2,

Individuo3,

….

]

Cada solución será inicializada con números enteros al azar comprendidos en el rango de [0, 9].

Criterio de finalización

Para la finalización del algoritmo, se manejará únicamente el criterio de un Máximo de generaciones. Queda a discreción del estudiante cual será el máximo de generaciones que utilice para obtener el mejor modelo. Durante la calificación se validará la implementación de esta parte con un numero bajo de generaciones (2 o 3).

Cálculo del valor fitness

Para calcular el valor fitness se entrenará una red neuronal con los híper parámetros del individuo. Luego se tomará esa red neuronal y se evaluará su rendimiento en el conjunto de datos de validación. Ese valor de rendimiento representara el valor fitness del individuo. Por ejemplo:

Individuo1 = [3, 0, 5, 0]

- Alpha = 0.0005
- Lambda = 0
- Max_iteraciones = 4000
- Keep_prob = 1

Nota: Los valores fueron extraídos de la tabla anterior.

Se entrena una red neuronal con los híper parámetros anteriores y al evaluar su rendimiento se obtienen los siguientes resultados:

Exactitud Entrenamiento: 0.96

Exactitud Validación: 0.67

De los resultados anteriores el valor de 0.67 corresponde al fitness del individuo. Como ese valor representa que tan buena fue la predicción, entre más alto sea el valor mejor será.

Selección de padres

La forma en la que se seleccionaran los mejores padres dependerá de cada estudiante, así como también la cantidad de padres que se seleccionaran.

Emparejamiento

La forma en la que se emparejaran los padres quedara a discreción del estudiante, tomar en cuenta que la cantidad de parejas que se tendrán que formar dependerá de la cantidad de padres que hayan seleccionado y del tamaño de la población.

Cruzamiento

El proceso de generar a los nuevos elementos de la población se realizará de la siguiente manera. Cada posición del hijo tendrá el 50% de probabilidad de provenir de un padre, caso contrario será del otro padre.

Mutación

El proceso de mutación se realizará de la siguiente manera. Se elegirá una posición al azar y se le asignará un numero aleatorio comprendido en el rango [0, 9].

Reportes

Tabla de híper parámetros

Deberán mostrar los valores con los que llenaron la tabla de híper parámetros. No es necesario que se muestre desde la aplicación, sino que pueden mostrarla completamente aparte (por ejemplo, en Excel).

Diagrama de la red neuronal

Deberán realizar una representación gráfica de arquitectura de la red neuronal que van a utilizar. No es necesario que se muestre desde la aplicación, pueden mostrar el dibujo completamente aparte.

Consideraciones

- La aplicación deberá desarrollarse utilizando Python 3.
- El framework para la aplicación web queda a discreción del estudiante.
- El proyecto se realizará de manera individual, si se detecta algún tipo de copia el laboratorio quedará anulado.
- Se debe de crear un repositorio para llevar control de su proyecto.

Forma de Entrega

Entregables

- El entregable será el enlace al repositorio del proyecto, recordar que el auxiliar debe de tener siempre acceso a dicho repositorio, de lo contrario se anulará la nota obtenida en el proyecto.

Fecha de entrega

Martes 5 de enero de 2021, hasta las 23:59