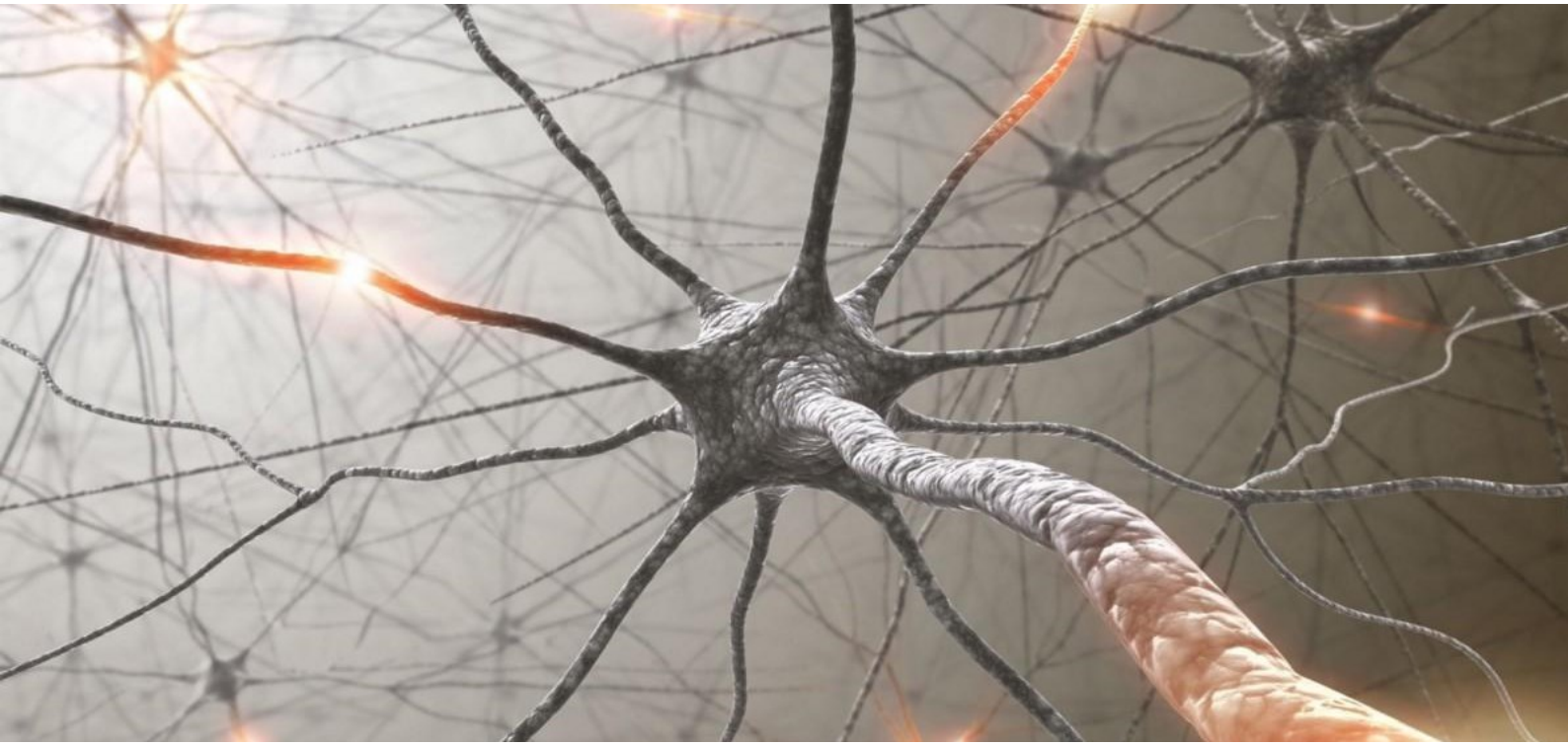
Práctica 2

Perceptrón



Objetivos

El objetivo principal de esta práctica es comprender el funcionamiento del perceptrón como bloque de construcción elemental de las redes neuronales. Adicionalmente se realiza un repaso sobre el análisis y preprocesamiento de los datos.

Temas

- Normalización y tratamiento de datos
- Correlación
- Representaciones Gráficas
- Perceptrón. Entrenamiento

Lectura

Material de Lectura: Capítulo 13 del Libro Introducción a la Minería de Datos de Hernández Orallo

Ejercicio 1

El archivo **Hawks.csv** contiene mediciones de casi 900 aves de tres especies diferentes. Los datos registrados son los siguientes:

- **Epecies.** Es la etiqueta de clase e indica la especie de cada gavilán: gavilanes de Cooper (CH), gavilanes colirrojos (RT) y gavilanes rastreros (SS).
- **Año.** Indica el año en que se avistó el ave.
- **Hallux.** Indica la medida en milímetros del hallux (el dedo posterior del pie, que poseen todas las aves cazadoras para sujetar mejor a sus presas).
- **Ala.** Indica la longitud en milímetros de la pluma más larga del ala.
- **Peso.** Indica el peso del ave en gramos.
- **Cola.** Indica la longitud de la cola del ave en milímetros.

a) Calcule la correlación lineal entre los atributos **Ala** y **Cola**. Indique la intensidad de la correlación (no hay correlación/débil/fuerte) y el tipo (positiva/negativa)

Valor	
Intensidad	
Tipo	

b) Complete el cuadro con los valores del atributo ***Hallux*** antes y después de normalizarlos utilizando media y desvío.

Medida	Sin normalizar	Normalizado
Media		
Desvío		
Mínimo		
Q1		
Q2		
Q3		
RIC		
Máximo		
Bigote Inferior *(valor extremo mínimo dentro del rango de normalidad)*		
Bigote Superior *(valor extremo máximo dentro del rango de normalidad)*		

Compare los valores obtenidos e indique el valor de verdad de las siguientes afirmaciones:

I. El valor de Q2 normalizado permite afirmar que la media y la mediana tienen valores cercanos.
II. El máximo valor de Hallux se encuentra a más de 9 desvíos por encima de la media.
III. Un hallux que mida más de 600 mm es considerado atípico extremo.
IV. Un valor normalizado del atributo “hallux” de 2 se considerará un valor atípico extremo.
V. En un gavilán es atípico que su hallux mida menos de 100 mm.

Ejercicio 2

El archivo Globos.csv contiene el registro de 16 intentos para inflar globos. Los atributos registrados fueron el color del globo, el tamaño, si se estira o no y si la acción de inflarlo fue realizada por un adulto o por un niño. En cada registro se anotó si el globo pudo ser inflado o no.

Numerice los atributos de la siguiente forma:

- Color = 1 si es rojo y 0 si es amarillo
- SeEstira? = 1 para Si y 0 para No
- Edad = 1 si es adulto y 0 si es Nene
- Tamanio = 0 si es Chico, 0.5 si es mediano y 1 si es grande

a) Indique cuáles de los siguientes perceptrones pueden clasificar correctamente todos los ejemplos normalizados del archivo Globos.csv

Configuración	w(Color)	w(Tamanio)	w(Se_estira?)	w(Edad)	b
Perceptrón 1	0,050	0,006	-0,068	-0,047	0,005
Perceptrón 2	-0,494	-0.003	0,722	0,489	-0.4943
Perceptrón 3	0,06	0,004	0,1278	0,021	-0,1132
Perceptrón 4	0,296	0,026	-0,576	-0,287	0,371
Perceptrón 5	0,994	0,006	-1,466	-0,989	0,988

b) Indique cuál de los 5 perceptrones del inciso anterior presenta el peor desempeño, es decir, cuál es el que clasifica la menor cantidad de ejemplos correctamente.

Ejercicio 3

La Tabla 1 muestra información correspondiente de pacientes para determinar si deben realizarse un examen médico en función de su edad, altura y riesgo médico.

EDAD	RIESGO	EXAMEN
55	ALTO	SI
56	MEDIO	NO
58	MEDIO	SI
56	BAJO	NO
59	BAJO	NO
57	MEDIO	SI
60	BAJO	SI
53	ALTO	NO
59	MEDIO	SI

Tabla 1

Donde:

- **EDAD** es un atributo numérico que indica la edad del paciente.
- **RIESGO** es el nivel de riesgo del paciente.
- **EXAMEN** indica si debe realizarse un examen extra.

Para obtener transformar el atributo nominal RIESGO en uno numérico se lo numerizó de la siguiente forma: BAJO = 1, MEDIO = 2 y ALTO = 3.

a) Luego de la numerización se calculó el coeficiente de correlación lineal entre los atributos EDAD y RIESGO y se obtuvo como resultado -0.71. ¿Cómo debe interpretarse este valor?

b) Luego de numerizar el atributo RIESGO y de normalizar los atributos de manera lineal entre 0 y 1, los ejemplos fueron utilizados para entrenar un perceptrón capaz de predecir correctamente el atributo EXAMEN. Los pesos obtenidos fueron los siguientes:

W(EDAD) = 0.0807 W(RIESGO) = 0.074 Sesgo o bias = -0.0742

¿Cuál será la respuesta del perceptrón para los siguientes valores?

(EDAD, RIESGO) = (50, BAJO), (51, MEDIO), (54, ALTO)

Ejercicio 4

Se busca predecir si el tipo de fármaco que se debe administrar a un paciente afectado de rinitis alérgica es el habitual o no. Se dispone de información correspondiente a las historias clínicas de pacientes atendidos previamente. Las variables relevadas son las siguientes:

- **Age**: Edad
- **Sex**: Sexo
- **BP (Blood Pressure):** Presión sanguínea.
- **Cholesterol**: nivel de colesterol.
- **Na**: Nivel de sodio en la sangre.
- **K**: Nivel de potasio en la sangre.
- **Class**: Fármaco suministrado. Cada paciente ha sido medicado con un único fármaco de 5 posibles: DrugA, DrugB, DrugC, DrugX, DrugY

a) Utilice el archivo **Drug_train.csv** para entrenar un perceptrón que sea capaz de predecir si el tipo de fármaco que se debe administrar a un paciente afectado de rinitis alérgica es el habitual (suministro de DrugY) o no.

b) Luego utilice el archivo **Drug_test.csv** para medir la calidad del modelo.

c) Resuelva el problema:

- numerizando los atributos ordinales utilizando dos representaciones diferentes: como entero único y de manera binaria (dummy).
- Normalizando de diferentes formas: Sin normalizar, normalización lineal, normalización con media y desvío

Ejercicio 5

El archivo **SEMILLAS.csv** contiene información de granos que pertenecen a tres variedades diferentes de trigo: Kama, Rosa y Canadiense. El total es de 210 ejemplos a razón de 70 ejemplos para cada tipo de grano, seleccionados al azar para el experimento. La información registrada corresponde al resultado de la visualización de alta calidad de la estructura interna del núcleo efectuada utilizando una técnica de rayos X blandos. Este tipo de estudio no es destructivo y es considerablemente más económico que otras técnicas de imagen más sofisticadas como la microscopía de barrido o la tecnología láser. Las imágenes se grabaron en placas KODAK de rayos x de 13x18 cm. Los estudios se realizaron utilizando granos de trigo cosechados combinados procedentes de campos experimentales, explorados en el Instituto de Agrofísica de la Academia Polaca de Ciencias en Lublin. Para construir los datos, se midieron siete parámetros geométricos de cada grano de trigo:

- área A
- perímetro P
- compacidad C = 4 * pi * A / P ^ 2
- longitud del núcleo
- ancho del núcleo
- coeficiente de asimetría
- longitud del surco del núcleo

A partir de los 210 ejemplos, luego de normalizarlos utilizando los valores de media y desvío, se logró entrenar un perceptrón capaz de identificar, con una precisión del 100%, uno de los tres tipos de semillas. Para realizar el entrenamiento se utilizó una velocidad de aprendizaje de 0.05 y un máximo de 200 iteraciones. Indique cuál es el tipo de semilla que puede ser reconocido correctamente por un perceptrón.

Ejercicio 6

El archivo **Zoo.csv** contiene información de 101 animales caracterizados por los siguientes atributos

A1. Nombre del animal	A6. Vuela	A11. Branquias	A16. Domestico
A2. Tiene Pelo	A7. Acuático	A12. Venenoso	A17. Tamaño gato
A3. Plumas	A8. Depredador	A13. Aletas	A18. Clase
A4. Huevos	A9. Dentado	A14. Patas	
A5. Leche	A10. Vertebrado	A15. Cola	

Salvo los atributos A1 y A18 que contienen texto y el A14 que contiene el número de patas del animal, el resto toma el valor 1 si el animal posee la característica y 0 si no. Hay 7 valores de clase posible (atributo A18): mamífero, ave, pez, invertebrado, insecto, reptil y anfibio.

a) Realice un gráfico que visualice de la cantidad de ejemplos por cada valor del atributo **clase** y analice que tipos de problema podrían surgir al entrenar un modelo para clasificación.

b) Utilice todos los ejemplos para entrenar un perceptrón que sea capaz de reconocer si un animal es un mamífero. Entrene varias veces si es necesario y verifique que funcione correctamente.

c) Observe los pesos del perceptrón entrenado en a) ¿Puede determinar cuáles son las características más relevantes para decidir si se trata de un mamífero o no? Realice varias ejecuciones independientes y observe si las características más relevantes siguen siendo las mismas.

d) Repita b) y c) para las aves.

e) Repita b) y c) para los reptiles.

Ejercicio 7

El archivo ***automobile-simple.csv*** contiene 11 atributos de automóviles de un total de 205 registros. Es una versión modificada y simplificada del dataset disponible en el repositorio UCI https://archive.ics.uci.edu/ml/datasets/Automobile. La siguiente tabla contiene una breve descripción de los atributos que contiene el archivo y caracterizan a cada vehículo.

Atributo	Descripción
make	Marca: alfa-romero, audi, bmw, chevrolet…
fuel-type	Tipo de combustible: diesel, gasolina.
num-of-doors	Número de puertas: cuatro, dos.
body-style	Tipo de carrocería: techo duro, wagon, sedán, hatchback, descapotable.
curb-weight	Peso en vacío: numérico de 1488 a 4066.
engine-size	Tamaño del motor: numérico de 61 a 326.
horsepower	Potencia: numérico de 48 a 288.
city-mpg	Rendimiento en ciudad (en millas por galón): numérico de 13 a 49.
highway-mpg	Rendimiento en ruta (en millas por galón): numérico de 16 a 54.
price	Precio en USD: numérico de 5118 a 45400.
volume	Volumen del vehículo (alto x ancho x alto).
eco-rating	Evaluación de la sostenibilidad ambiental del vehículo basado en su volumen, peso, consumo de combustible y tipo de combustible utilizado.

a) Para cada atributo indique si es Discreto, Continuo, Nominal u Ordinal.

b) Elimine los registros que presenten valores faltantes.

c) Calcule la matriz de correlación usando los atributos numéricos.

d) Realice el entrenamiento de un perceptrón para que aprenda a clasificar si un auto es ecológico. Tenga en cuenta los siguientes pasos:
   I. Utilice el atributo eco-rating para generar un nuevo atributo binario que determine si un auto es ecológico o no. Un auto es considerado ecológico si el valor de eco-rating supera la media de dicho atributo.
   II. Genere y compare 3 modelos utilizando diferentes normalizaciones (Sin normalizar, normalización lineal, normalización estándar.
   III. Teniendo en cuenta la matriz de correlación del punto c) repita el punto II) eliminando dos atributos fuertemente correlacionados (uno negativo y otro positivo). Compare y reflexione sobre los resultados obtenidos.