PROMiDAT Iberoamericano
Machine Learning con Python IV
Calibración de Modelos

- Las tareas tienen fecha de entrega una semana después a la clase y deben ser entregadas antes del inicio de la clase siguiente.
- Cada día de atraso en implicará una pérdida de 10 puntos.
- Las tareas son estrictamente de carácter individual, tareas iguales se les asignará cero puntos.
- En nombre del archivo debe tener el siguiente formato: `Tarea1_nombre_apellido.pdf`. Por ejemplo, si el nombre del estudiante es Luis Pérez: `Tarea1_luis_perez.pdf`. Para la tarea número 2 sería: `Tarea2_luis_perez.pdf`, y así sucesivamente.
- Esta tarea tiene un valor de un 25 % respecto a la nota total del curso.

Tarea Número 1

- **Pregunta 1:** [25 puntos] En este ejercicio usaremos los datos (`voces.csv`). Se trata de un problema de reconocimiento de género mediante el análisis de la voz y el habla. Esta base de datos fue creada para identificar una voz como masculina o femenina, basándose en las propiedades acústicas de la voz y el habla. El conjunto de datos consta de 3.168 muestras de voz grabadas, recogidas de hablantes masculinos y femeninos.

  El conjunto de datos tiene las siguientes propiedades acústicas (variables) de cada voz:

  - `meanfreq`: frecuencia media (en kHz).
  - `sd`: desviación estándar de frecuencia.
  - `median`: frecuencia mediana (en kHz).
  - `Q25`: primer cuantil (en kHz).
  - `Q75`: tercer cuantil (en kHz).
  - `IQR`: rango intercuantile (en kHz).
  - `skew`: sesgo (ver nota en la descripción de specprop).
  - `kurt`: kurtosis (ver nota en la descripción de specprop).
  - `sp.ent`: entropía espectral.
  - `sfm`: planitud espectral.
  - `mode`: modo frecuencia.
  - `centroide`: centroide de frecuencia (ver specprop).
  - `peakf`: frecuencia de pico (frecuencia con mayor energía).
  - `meanfun`: promedio de la frecuencia fundamental medida a través de la señal acústica.
  - `minfun`: frecuencia mínima fundamental medida a través de la señal acústica.
  - `maxfun`: máxima frecuencia fundamental medida a través de la señal acústica.
  - `meandom`: promedio de la frecuencia dominante medida a través de la señal acústica.

- `mindom`: mínimo de la frecuencia dominante medida a través de la señal acústica.
- `maxdom`: máximo de la frecuencia dominante medida a través de la señal acústica.
- `dfrange`: rango de frecuencia dominante medido a través de la señal acústica.
- `modindx`: índice de modulación. Calculado como la diferencia absoluta acumulada entre las mediciones adyacentes de las frecuencias fundamentales dividida por la gama de frecuencias.
- `género`: Masculino o Femenino (variable a predecir).

Realice lo siguiente:

1. Cargue la tabla de datos `voces.csv` en `Python`.
2. Genere al azar una tabla de testing con una 20 % de los datos y con el resto de los datos genere una tabla de aprendizaje.
3. Usando `MLPClassifier` genere un modelo predictivo para la tabla de aprendizaje. Utilice una cantidad suficiente de capas ocultas y nodos para que la predicción sea buena.
4. Con la tabla de testing calcule la matriz de confusión, la precisión, la precisión positiva, la precisión negativa, los falsos positivos, los falsos negativos, la acertividad positiva y la acertividad negativa. Luego construya un cuadro comparativo.
5. Construya un cuadro comparativo con respecto a las tareas del curso anterior. ¿Cuál método es mejor?
6. Repita los ejercicios anteriores, pero esta vez utilice el paquete `Keras`, utilice la misma cantidad de capas ocultas y nodos que la usada arriba. ¿Mejora la predicción?
7. Compare los resultados con los obtenidos en las tareas del curso anterior.

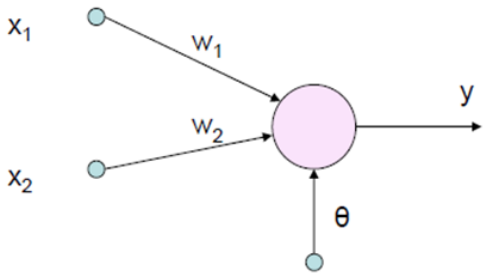
- **Ejercicio 2:** [25 puntos] Esta pregunta utiliza los datos (`tumores.csv`). Se trata de un conjunto de datos de características del tumor cerebral que incluye cinco variables de primer orden y ocho de textura y cuatro parámetros de evaluación de la calidad con el nivel objetivo. La variables son: Media, Varianza, Desviación estándar, Asimetría, Kurtosis, Contraste, Energía, ASM (segundo momento angular), Entropía, Homogeneidad, Disimilitud, Correlación, Grosor, PSNR (Pico de la relación señal-ruido), SSIM (Índice de Similitud Estructurada), MSE (Mean Square Error), DC (Coeficiente de Dados) y la variable a predecir `tipo` (1 = Tumor, 0 = No-Tumor).

  1. Usando el paquete `MLPClassifier` y el paquete `Keras` en `Python` genere modelos predictivos para la tabla `SpamData.csv` usando 70 % de los datos para tabla aprendizaje y un 30 % para la tabla testing. Utilice una cantidad suficiente de capas ocultas y nodos para que la predicción sea buena.
  2. Calcule para los datos de testing la precisión global y la matriz de confusión. Interprete la calidad de los resultados. Además compare respecto a los resultados obtenidos en las tareas del curso anterior.
  3. Compare los resultados con los obtenidos en las tareas del curso anterior.

- **Pregunta 3:** [25 puntos] [**no usar** `MLPClassifier` **ni** `Keras`] Diseñe una Red Neuronal de una capa (Perceptron) para la tabla de verdad del `nand`:

x_1	x_2	y
0	0	1
1	0	1
0	1	1
1	1	0

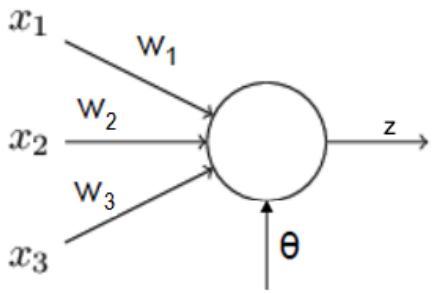
Es decir, encuentre los pesos w_1, w_2 y el umbral θ para la Red Neuronal que se muestra en el siguiente gráfico, usando una función de activación tipo Sigmoidea:



- **Pregunta 4:** [25 puntos] [**no usar** `MLPClassifier` **ni** `Keras`] Para la Tabla de Datos que se muestra seguidamente donde x^j para $j = 1, 2, 3$ son las variables predictoras y la variable a predecir es z diseñe y programe a pie una Red Neuronal de una capa (Perceptron):

x^1	x^2	x^3	z
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	0

Es decir, encuentre todos los posibles pesos w_1, w_2, w_3 y umbrales θ para la Red Neuronal que se muestra en el siguiente gráfico:



Use una función de activación tipo Tangente hiperbólica, es decir:

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

Para esto escriba una Clase en `Python` que incluya los métodos necesarios pra implementar esta Red Neuronal.

Se deben hacer variar los pesos w_j con $j = 1, 2, 3$ en los siguientes valores `v=(-1,-0.9,-0.8,...,0,...,0.8,0.9,1)` y haga variar θ en `u=(0,0.1,...,0.8,0.9,1)`. Escoja los pesos w_j con $j = 1, 2, 3$ y el umbral θ de manera que se minimiza el error cuadrático medio:

$$E(w_1, w_2, w_3) = \frac{1}{4} \sum_{i=1}^{4} \left[I \left[f \left(\sum_{j=1}^{3} w_j \cdot x_i^j - \theta \right) \right] - z_i \right]^2,$$

donde x_i^j es la entrada en la fila i de la variable x^j e $I(z)$ se define como sigue:

$$I(t) = \begin{cases} 1 & \text{si } t \geq 0 \\ 0 & \text{si } t < 0. \end{cases}$$