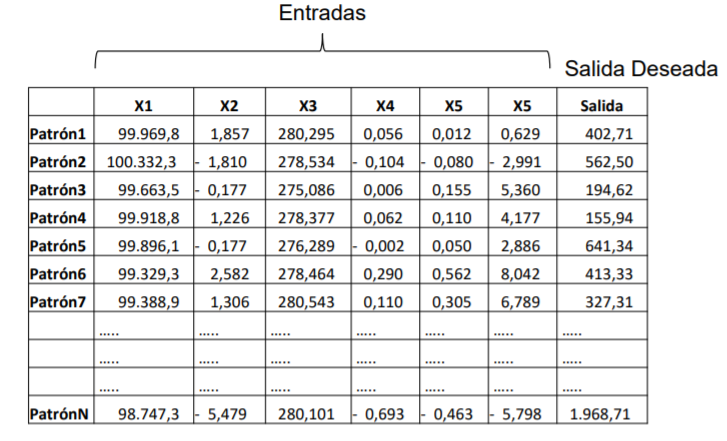
Datos

• Serán variables, datos o patrones de **entrada y de salida** deseada para aquellas redes que utilicen **aprendizaje supervisado;** o variables, datos o patrones de **entrada** para aquellas redes que utilicen aprendizaje **no supervisado (no hay salida).**

• Las redes de neuronas **sólo** trabajan **con atributos numéricos.**

• Los **atributos no numéricos** (nominales) deben **discretizarse** a valores numéricos. Esta discretización puede influir en los resultados de la red.



Transformación de los datos

# Normalización

Consiste en transformar los datos en un intervalo [0,1]. Si bien **no es obligatorio**, siempre es **recomendable** porque evita problemas durante el aprendizaje, como la **saturación de neuronas** (suele pasar en Perceptrón Multicapa).

Se aconseja normalizar cada variable de manera independiente.



# Aleatorización

Para **evitar sesgos**, se aconseja aleatorizar los datos ya que pueden aparecer ordenados.

Evaluación

# Separación del conjunto de entrenamiento, validación y test (Práctica 1)

Necesitamos que la red generalice para datos que no se han utilizado durante el entrenamiento.

*Nota: Parámetros se refiere a los pesos. A los parámetros de más alto nivel como la tasa de aprendizaje o el número de ciclos se conocen como Hiperparámetros*

El conjunto de datos disponible se separa en tres subconjuntos:

* Conjunto de **entrenamiento** 🡪 Para el proceso de aprendizaje.
* Conjunto de **validación** 🡪 Para decidir la parada del aprendizaje y determinación de los hiperparámetros de la red como la tasa de aprendizaje, el número de ciclos, el número de neuronas por capa, etc. si los hiperparámetros se eligieran con el conjunto de entrenamiento podría haber un problema de **sobreaprendizaje o overfitting**
* Conjunto de **test** 🡪 Para generalización

Esta división se debe hacer **aleatoriamente**

# Validación cruzada (Práctica 2)

Se usa para evitar sesgos en los datos de entrenamiento y test. Muy útil cuando hay pocos datos disponibles.

Se divide el conjunto de datos en k partes o subconjuntos. Supongamos k=3 (generalmente k=10) generando los subconjuntos A, B, y C. Se realizan k (3) iteraciones:

* Aprender con A, B y test con C (T1 = medida de evaluación con C)
* Aprender con A, C y test con B (T2 = medida de evaluación con B)
* Aprender con B, C y test con A (T3 = medida de evaluación con A)
* Medida de evaluación final T = (T1+T2+T3)/3

La elección de los hiperparámetros se puede hacer para cada subconjunto, extrayendo de cada uno un subconjunto de validación (validación anidada)

Cuando se utiliza Validación Cruzada, el **modelo final se construye con todos los datos**.

# Validación cruzada estratificada

En problemas de clasificación con desbalanceo entre clases, si se utiliza VC es recomendable utilizar la VC estratificada.

– El procedimiento es igual, pero al dividir los datos en k partes, se deberá mantener la proporción de instancias de las clases del conjunto total en cada una de las partes A, B, y C (en el caso k=3)

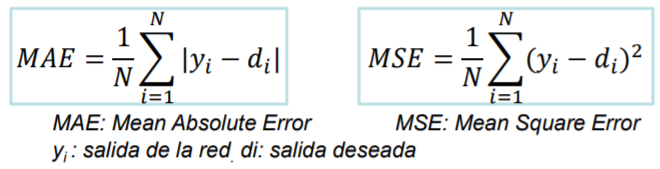
Por ejemplo, si se dispone de 100 patrones con 2 clases (C1 y C2) y si el número de patrones para la C1 es 75 y para la clase C2 es 25, entonces cuando se dividan los datos en las partes A, B y C, cada parte debe contener 75% de C1 y 25% de C2.

– Un posible mecanismo: dividir los patrones de cada clase en k partes y posteriormente juntarlos

– Muchas herramientas y/o librerías de Aprendizaje Automático contienen la VC y VC Estratificada.

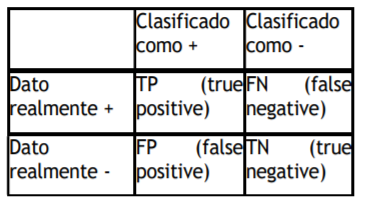
# Medida de evaluación

* En **problemas de regresión,** la más habitual es aplicar medidas de error sobre los datos de entrenamiento y test:



* En **problemas de clasificación,** lo más habitual es calcular el **porcentaje de aciertos**
* En **problemas de agrupación o clustering** se suele medir la cohesión de cada grupo o clúster y la separación entre los grupos (distancia Euclídea)

# Matriz de confusión



* El porcentaje de aciertos total es (TP+TN)/(TP+TN+FN+FP)
* El porcentaje de aciertos de + es: TP/(TP+FN)
* El porcentaje de aciertos – es: TN/(FP+TN)