Preparación de Datos. Evaluación

- Los datos y sus transformaciones
- Evaluación de una red de neuronas

Los datos y sus transformaciones

- Variables, datos o patrones de entrada y variables, datos o patrones de salida deseada para aquellas redes que utilicen aprendizaje supervisado.
- Variables, datos o patrones de entrada para aquellas redes que utilicen aprendizaje no supervisado.
- 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.

Los datos y sus transformaciones

- Trasformaciones de los datos
  - **Normalización de los datos.** Es aconsejable trabajar con datos normalizados en un cierto intervalo, generalmente el intervalo [0,1]. Si bien no es obligatorio, siempre es recomendable pues evita problemas durante el aprendizaje, como la saturación de las neuronas.

$$\mathrm{VarNor}_i = (\mathrm{Var}_i - \mathrm{VMin}_i)/(\mathrm{VMax}_i - \mathrm{VMin}_i)$$

  - **Aleatorización de los datos**. Para evitar sesgos en el aprendizaje, siempre es conveniente aleatorizar los patrones o datos disponibles.
  - Otras posibles transformaciones:
    - **Eliminación de atributos irrelevantes**, pues no aportan información al problema
    - **Reducción dimensionalidad**: Para algunos problemas, los datos de entada poseen alta dimensión. Es conveniente aplicar técnicas de reducción de dimensionalidad, que bien seleccionan un subconjunto de atributos, bien transforman los datos de entrada en otro conjunto de menor dimensión. No se entra en detalle en los diferentes métodos de reducción de dimensionalidad, pues no es objetivo del curso.

Evaluación de una red de Neuronas

- **Separación del conjunto de entrenamiento y test**.
  - El conjunto de datos disponible se separa en dos subconjuntos: Conjunto de entrenamiento (aprendizaje) y conjunto de test (generalización).
  - Esta división debe realizarse aleatoriamente, o al menos utilizando un mecanismo que garantice que los datos de test están representados en el conjunto de entrenamiento
- **Separación del conjunto de entrenamiento, validación y test**.
  - El conjunto de datos disponible se separa en tres subconjuntos: Conjunto de entrenamiento (aprendizaje), conjunto de validación (parada del aprendizaje y determinación de los parámetros de la red) y conjunto de test (generalización).
  - Esta división debe realizarse aleatoriamente
  - El conjunto de validación puede ser útil para elegir los mejores parámetros (número de ciclos, neuronas, etc). Si la elección se hace utilizando el conjunto de entrenamiento, puede darse sobreaprendizaje.

Evaluación de una red de Neuronas

- **Validación cruzada**.
  - Para evitar sesgos en los datos de entrenamiento y test. Pocos datos disponibles
  - Se divide el conjunto de datos en k partes o subconjuntos. Supongamos k=3 (generalmente k=10) ;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

Evaluación de una red de Neuronas

- **Medida de Evaluación**: dependen de la tarea o problema a resolver.
  - En **problemas de regresión** (aproximación) o predicción la manera más habitual de evaluar la red es mediante el error medio o error cuadrático medio cometido por la red sobre las datos de entrenamiento y test (son equivalentes) :

$$\text{EMedio} = \frac{1}{N}\sum_{i=1}^{N}\left|sm_i - sd_i\right| \qquad \text{ECuadráticoMedio} = \frac{1}{N}\sum_{i=1}^{N}(sm_i - sd_i)^2$$

  - **En problemas de clasificación** lo más habitual es evaluar la calidad de la red en base a su precisión predictiva (% de aciertos): el número de patrones del conjunto de entrenamiento o test clasificadas correctamente dividido por el número total de instancias en dicho conjunto
  - En problemas de agrupación o clustering se suele medir la cohesión de cada grupo o cluster y la separación entre los grupos (distancia Euclídea)

Evaluación de una red de Neuronas

- A lo hora de decidir si una red resuelve con éxito o no un problema
  - **En regresión**:
    - Un error medio (o error cuadrático medio) aceptable
    - Visualización gráfica de las salidas deseadas y las salidas proporcionadas por la red
  - **En clasificación**
    - Un alto % de aciertos, aunque hay que tener presente:
      - En problemas con M clases, el porcentaje de aciertos debe superar el 100*1/M. De otra manera, sería mejor tirar una moneda (azar) que utilizar el clasificador
      - Si se dispone de una clase con muchos más datos que otra, el porcentaje de aciertos a superar es el porcentaje de datos de la clase mayoritaria. Por ejemplo, si tenemos dos clases (+ y -) y hay 90 datos + y 10 -; un clasificador que prediga siempre + ya acertará en un 90%. Hay que hacerlo mejor que eso.
    - En ocasiones el coste de fallar en una clase no es el mismo que fallar en otra. Por ejemplo, para un clasificador de cáncer si/no, es preferible predecir que una persona tiene cáncer (sin tenerlo) que predecir que no lo tiene (teniéndolo). Para estas situaciones se suele utilizar la matriz de confusión, que contiene falsos positivos y falsos negativos.

Evaluación de una red de Neuronas

Matriz de confusión (problemas de clasificación)

	Clasificado como +	Clasificado como -
Dato realmente +	TP (true positive)	FN (false negative)
Dato realmente -	FP (false positive)	TN (true negative)

- Los datos correctamente clasificados están en la diagonal, los incorrectos fuera de ella.
- 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)