1) **Explique qué es un algoritmo de clustering, qué se requiere para utilizarlo y cuál es su salida o resultado. Enuncie los pasos del algoritmo k-means (HCM) conceptualmente, sin matemática.**
Clustering → agrupamiento de datos
El algoritmo de clustering es una técnica de aprendizaje no supervisado se utiliza para agrupar datos con distintas features (D). El objetivo es descubrir regiones del espacio de datos que presenten una alta densidad de puntos. Es un método que organiza los datos en grupos de manera que los elementos dentro de un mismo grupo sean más parecidos entre sí que con los de otros grupos.

Pasos:

1. Se elige una cantidad de clusters K y se los posiciona en un lugar aleatorio.
2. Genero la matriz de distancia de cada punto a cada cluster.
3. Asigno cada dato a un centro de clusters
4. Se recalculan los centros haciendo el promedio entre los datos que pertenecen a cada cluster.
5. Este proceso se repite (2-4) hasta una determinada cantidad de iteraciones o hasta que los centro en un instante t sean iguales al instante anterior.

La salida principal de un algoritmo de clustering es la asignación de cada dato a un cluster, lo cual se visualiza en una matriz de pertenencias.

Los resultados dependen del tipo de algoritmo utilizado si se aplica el Hard C Means la matriz de pertenencias es binaria, si el valor es un 1 el dato pertenece al cluster si no es 0. Si se utiliza Fuzzy C Means la matriz es difusa, cada valor representa el grado de pertenencia de un dato al cluster.

Para los algoritmos de clustering se necesitan datos no clasificados.

Clustering sustractivo: además de agrupar, **estima automáticamente cuántos clusters existen** en el conjunto de datos.

2) **¿Cómo asigna nuevos datos a alguno de los clusters obtenidos por un algoritmo? ¿La pertenencia a un cluster es absoluta?**
Una vez que el algoritmo generó los centros de los clusters, cuando llega un nuevo dato se compara con esos centros.
Para el hard C-Means se calcula la distancia entre el nuevo punto y cada centro, el punto se le asigna al cluster más cercano. La pertenencia es absoluta, el nuevo punto pertenece a un único cluster.
En el fuzzy C-Means el nuevo punto no se asigna a un único cluster, se calcula su grado de pertenencia a cada cluster, según las distancias a los centroides. La pertenencia no es absoluta, es difusa.
Grado de pertenencia de cada dato a cada cluster:

$$m_{ik} = \frac{\left(\frac{1}{\|\mathbf{u}_i - \mathbf{c}_k\|^2}\right)^{\frac{1}{q-1}}}{\sum_{j=1}^{K}\left(\frac{1}{\|\mathbf{u}_i - \mathbf{c}_j\|^2}\right)^{\frac{1}{q-1}}}$$

Exponente de fuzzificación q ∈ [1.5, 3.5]
[fuzziness exponent / fuzzifier]

Actualización de los centros de cluster (promedio pesado):

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} m_{ik}^{q}\,\mathbf{u}_i}{\sum_{i=1}^{N} m_{ik}^{q}}$$

3) **¿Qué diferencia un algoritmo de clustering de uno de clasificación?**
En un algoritmo de clustering los datos no tienen etiquetas o clases conocidas. El algoritmo aprende por sí solo grupos de datos, basándose en su similitud. Es un proceso no supervisado, porque no se le indica cuál debería ser el resultado esperado, en base a features del dato me lo agrupara a un cluster en el cual las features tengan alguna similitud. Mientras que el de clasificación los datos si tienen etiquetas o clases conocidas. El modelo aprende con los datos y con sus resultados y a partir de ellos predice la clase de los nuevos datos. Es un proceso de aprendizaje supervisado porque se entrena con datos donde ya se conoce el resultado correcto.

4) **Indique claramente la diferencia entre un modelo supervisado y no supervisado.**
Arriba

5) **¿Cómo puede evaluarse la calidad de un método de clustering si no se cuenta con datos previamente etiquetados? Defina medidas internas de calidad del clustering. ¿Qué es lo que intentaría maximizar o minimizar para decir que un algoritmo es mejor que otro?**
La evaluación de la calidad de un método de clustering cuando no se dispone de datos previamente etiquetados se realiza utilizando medidas internas de calidad.
- Distancia intra-Cluster: esta evalúa que tan cerca están los puntos de su centro de cluster. Se mide con la suma de las distancias al cuadrado de cada punto a su centro.

$$\sum_{k=1}^{K} \sum_{i, ui \in ck} \| ui - ck \|^2$$

K= número de clusters
Ui= cada dato del cluster
Ck= centro de cluster
Está distancia se tiende a minimizar.

- Distancia inter-Cluster: esta evalúa qué tan lejos están los clusters entre sí (distancia entre los centros)

$$\sum_{i=1}^{K}\sum_{k=i}^{K} \quad \| ci - ck \|^2$$

Esta distancia se tiende a maximizar. Los grupos están más diferenciados.

- Índice de Silhouette (S): para cada dato se calcula:

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

a(i) es la distancia promedio del dato i con todos los datos de su mismo cluster
bk(i) son las distancias promedio del dato i con todos los datos de otro cluster k
b(i) es la mínima bk(i) para el dato i
Varía entre –1 y 1 y se buscan índices altos para todos los datos (para determinar que los clusters este bien definidos) → a bajo, b alto.

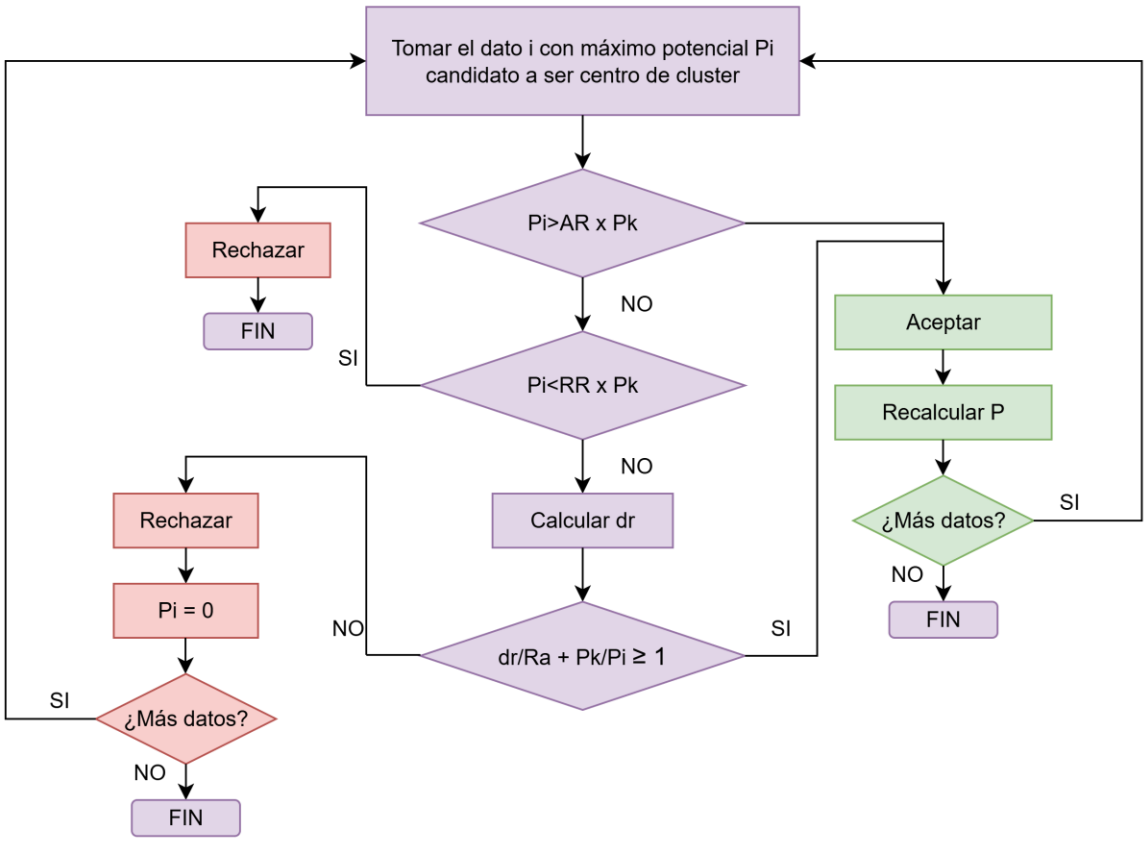
6) Enuncie los pasos del algoritmo k-means (HCM) conceptualmente, sin matemática.
En el 1.

7) Enuncie conceptualmente los pasos del algoritmo Clustering Substractivo. Destaque los parámetros de configuración que provocarían diferentes resultados. Compare el algoritmo con el k-means, descubra semejanzas y diferencias.

Pasos:

- ✓ Calcular un valor numérico que evalúe cuál es el punto con más "vecinos" (con algún criterio para definir quiénes son "vecinos"). Este valor numérico se lo conoce como potencial de cluster. Solo aportarán al valor de P los "vecinos fuzzy" que se encuentran aproximadamente dentro del radio que determina el hiperparámetro Ra. El dato con el máximo valor de P será el primer centro de cluster.
- ✓ Descartar a los vecinos "demasiado cercanos" (con algún criterio para definir quiénes son "demasiado cercanos"). Normalmente Rb es mayor que Ra para prevenir centros de cluster muy cercanos. Los puntos "cercanos" (según el valor del parámetro Rb) al centro de cluster C1 verán muy reducidas sus medidas de densidad y se descartarán los puntos con potencial menor que cierto umbral. El dato con el máximo valor de P (recalculado) será el segundo centro de cluster.
- ✓ Repetir este "descarte" con los datos que quedaron. Los puntos elegidos sucesivamente serán los centros de clusters.

El objetivo de este algoritmo es descubrir regiones del espacio de datos con alta densidad de puntos y estimar la cantidad de clusters para los mismos.



Parámetros para provocar diferentes resultados:

- Ra: radio de “vecinos” (radio de vecindad). Determina el tamaño del vecindario usado para calcular el potencial. Radios pequeños producen muchos clusters, radios grandes producen pocos.
- Rb: radio de demasiado cercanos (radio de sustracción). Suele tomarse Rb=1.5Ra. Controla cuánto se reduce el potencial alrededor de cada centro.
- AR: coeficiente de aceptación Adjusted Rand Index (Índice de Rand ajustado)
- RR: coeficiente de rechazo Recall Rate (Tasa de Recuperación/Sensibilidad).
- Factores de aceptación/rechazo → controlan qué tan distintos deben ser los centros nuevos respecto a los existentes.

Semejanzas:

- Ambos buscan encontrar centros de cluster que representen bien la distribución de los datos.
- En ambos casos, los clusters se forman alrededor de estos centros mediante una medida de distancia.

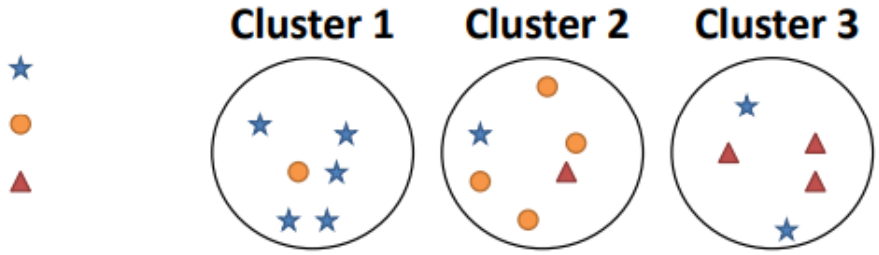
Diferencias:

	Clustering substractivo	K-Means

Número de clusters	No requiere fijar K de antemano, lo determina según densidad	Se debe especificar K antes de ejecutar
Inicialización de centros	Evalúa todos los puntos como candidatos, seleccionando en base a densidad. Automática	Inicia con K centros elegidos (aleatorios o heurísticos)
Criterio de agrupamiento	Basado en densidad y radios de influencia/rechazo	Minimización de distancia intra-Cluster (distancias medias)
Iteración	No es iterativo como K-Means, es más bien un procedimiento de selección progresiva	Iterativo: asigna puntos y recalcula centroides hasta convergencia
Sensibilidad	Muy sensible a los parámetros de radios y umbrales	Muy sensible a la elección inicial de los centroides.
Complejidad	Puede ser más costoso en grandes datasets (calcula densidad para cada punto)	Generalmente más eficiente en bases grandes.

8) **Busque en bibliografía (online o no) algún otro método de clustering que no se haya visto en el curso para intentar comprenderlo.**

9) **¿Cómo puede evaluarse la calidad de un método de clustering si se cuenta con algunos datos previamente etiquetados? Defina medidas externas de calidad del clustering.**
Purity → se calcula buscando la clase mayoritaria presente en cada cluster generado



- Se busca la **clase mayoritaria** en cada *cluster* (5, 4, 3 respectivamente).

$$purity = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap c_j|$$

$$purity = \frac{5+4+3}{17} \approx 0.71$$

Donde:

- N es el número total de datos en el conjunto
- k representa el cluster generado por el algoritmo
- j representa la clase verdadera pre-etiquetada
- wk es el número de datos que pertenecen tanto al cluster k como a la clase j

.....

Fuzzy logic

La lógica difusa se usa para manejar información imprecisa o incierta, como la que encontrarías en lenguaje humano ("la temperatura es alta"). Los sistemas de lógica difusa permiten:

- ✓ Transformar valores crisp (números concretos) en valores difusos mediante funciones de membresía.
- ✓ Aplicar reglas "IF-THEN" difusas para tomar decisiones.
- ✓ Convertir el resultado difuso a un valor concreto (defuzzification) si se necesita.

1) **Explique la generalización de la lógica booleana para pasar a conceptos de lógica difusa. ¿Qué es lo que se generaliza?**
   La lógica booleana clásica se opera con 0 y 1, en la lógica difusa se abre el panorama y no es todo 0 y 1. Se generaliza el concepto de verdad, no es llueve o no es llueve mucho, poco, nada y a cada una se les asigna un valor entre 0 y 1, donde cada número representa el grado de verdad o nivel de pertenencia para cada antecedente (es la parte de la regla que dice "si llueve mucho".
   Lo que se generaliza son los **valores de verdad** y las **operaciones lógicas** (AND, OR, NOT), que pasan de ser **discretas** a **numéricas y continuas**.

2) **Defina un conjunto difuso y todo lo necesario para su determinación a través de una función de pertenencia.**
   Un conjunto difuso A es A= {[x, μA (x))] x ∈ X} donde: μA (x)= a la función de pertenencia del valor x en A el cual va de 0 a 1.
   Primero en base a un valor de x y una función de pertenencia se obtiene un valor difuso de x que me dice que tan verdad es lo que afirmé, se le asigna un valor de pertenencia.

3) **Indique de forma genérica una regla difusa. Reconozca que en antecedente y consecuente incluye proposiciones difusas.**
   Una regla difusa es "Si x es A (antecedente) entonces y es b (consecuente)". También puede pasar que tenga operadores lógicos "Si x es A Y/O z es c entonces y es b"

   Antecedente (parte SI): Es una **proposición difusa compuesta**, que evalúa el grado de verdad de las condiciones.
   Consecuente (parte ENTONCES): También es una **proposición difusa**, que establece la acción o resultado.

Sugeno

1) **Escriba una regla típica de un FIS de tipo Sugeno. Compare claramente las diferencias con respecto a las de Mamdani.**

Si x es pequeño entonces y=a1x+b1
La diferencia con mamdani es que este tiene como salida un valor. No hace falta la defuzzificación.

2) **Indique los pasos de procesamiento de un dato de entrada hasta llegar a la salida en un FIS de tipo Sugeno.**
   1. Fuzzificación de las entradas: cada dato de entrada se evalúa respecto a las funciones de pertenencia, obteniendo los grados de pertenencia de cada entrada.
   2. Evaluación de las reglas: para cada regla se calcula su grado de activación, para ello se combinan los antecedentes con operadores lógicos difusos (ejemplo: min para AND, max para OR)
   3. Cálculo de las salidas: cada regla genera una salida numérica (en sugeno la consecuencia no es un conjunto difuso es una función matemática). Puede ser constante (y=ck) o lineal (y=a0+a1X.…)
   4. Agregación: las salidas numéricas de todas las reglas se ponderan por el grado de activación de las reglas.
   5. Obtención de la salida: la salida del sistema se obtiene mediante el promedio ponderado de las salidas de todas las reglas activadas
      $y = \frac{\mu 1 * y1 + \mu 2 * y2}{\mu 1 + \mu 2}$ sí tengo 2 reglas por lo tanto 2 funciones de pertenencia

3) **Explique diferentes maneras de generar los antecedentes en un FIS de tipo Sugeno: partición por grillas, clustering substractivo y FCM. Determine la relación que guarda la cantidad y posición de los clusters hallados en los datos o la cantidad de particiones de las variables de entrada con la cantidad de reglas.**
   - Partición por grillas (Grid Partitioning): Se divide cada variable de entrada en un numero de particiones difusas. Luego se forma una regla por cada combinación posible.
     Ejemplo:
     Variable 1 3 particiones
     Variable 2 2 particiones
     Total de reglas → 3*2 = 6 reglas
   - Clustering substractivo: Se aplica un método de clustering a los datos de entrada-salida. Cada punto de datos se considera candidato a centro de cluster, y se eligen los más representativos según una medida de densidad. Los clusters encontrados se usan como antecedentes, y a partir de ellos se construyen las reglas. El número de reglas se ajusta automáticamente a la distribución real de los datos. Desventaja →hay que elegir parámetros adecuados.
   - FCM: cada punto de datos pertenece a varios clusters con un grado de pertenencia. Los centros definen las funciones de pertenencia de los antecedentes. Cada cluster da lugar a una regla.

     Tanto en cluster substractivo como en FCM la cantidad de clusters determinan la cantidad de reglas. En cambio, en partición por grillas la productoria de la cantidad de partición por variable nos da la cantidad de reglas

4) **Plantee la ecuación típica para determinar los parámetros de los consecuentes de las reglas de un Fis de tipo Sugeno. ¿Cuántas ecuaciones y cuántos**

parámetros dispone, en función del número de datos (N) y la cantidad de variables (D)? ¿Cómo se resuelve el sistema de ecuaciones generado?

$$t = y(x1) = \frac{\mu_{p1}(x1, c1, \sigma1) \cdot (a_{01} + a_{11}x1) + \mu_{p2}(x1, c2, \sigma2) \cdot (a_{02} + a_{12}x1)}{\mu_{p1} + \mu_{p2}}$$

$T = X \cdot A$ A=T:X

Voy a tener una ecuación por cada dato de entrenamiento (N).

Y voy a tener D+1 parámetros por regla.

Se soluciona mediante mínimos cuadrados.

5) **En base a todo lo anterior, compare similitudes y diferencias de los FIS de tipo Mamdani y Sugeno. Indique cuándo debería utilizar cada uno.**
   - ✓ Similitudes: ambos sistemas son de inferencia difusa las cuales utilizan operadores difusos y trabajan en base reglas del tipo si entonces. Tienen las mismas etapas: fuzzificación, inferencia, agregación, salida. Ambos utilizan un operador difuso (mínimo o máximo) para evaluar los antecedentes y obtener los grados de pertenencia.
   - × Diferencias:
     - La creación del modelo con mamdani es en base a experiencia de expertos mientras que con Sugeno es basada en datos
     - Los consecuentes de mamdani son conjuntos difusos y los de sugeno son numéricos.
     - Mamdani requiere una defuzzificación para convertir el conjunto difuso a un valor real. Sugeno requiere un promedio pesado de las salidas y su grado de pertenencia.

   Conviene usar Mamdani cuando:

   - ❑ Es crucial capturar el conocimiento experto de una manera intuitiva y humanística. Ideal cuando los expertos expresan su conocimiento en términos vagos y ambiguos que se mapean a conjuntos difusos de salida, como “la distancia de frenado es larga”
   - ❑ La interpretabilidad del modelo es una prioridad. Los consecuentes difusos de mamdani son más fáciles de entender y verificar por los expertos humanos.
   - ❑ El problema se presta a una descripción cualitativa que refleja la forma en que los expertos piensan sobre un problema complejo, incluso si el proceso es más exigente.

   Conviene usar Sugeno cuando:

   - ➢ La eficiencia computacional es crucial. Su defuzzificación más sencilla lo hace más rápido
   - ➢ Se necesitan técnicas de optimización y adaptación. El método de Sugeno funciona muy bien con ellas, lo que permite que el sistema aprenda y ajuste sus parámetros a partir de los datos. Particularmente para sistemas dinámicos no lineales.
   - ➢ Las funciones de salida singleton (constantes) son suficientes para las necesidades del problema.

- Se busca un enfoque sistemático para generar reglas difusas a partir de un conjunto de datos de entrada y salida, como el ANFIS.
- La generación del modelo está basada en datos o un problema de control, donde la velocidad y la capacidad de ajuste automático de los parámetros son ventajosas.
  Mamdani sobresale la representación del conocimiento intuitivo y la interpretabilidad, mientras que Sugeno es preferible por su eficiencia computacional y su aptitud para la optimización y el control adaptativo.

6) ¿Cómo puede evaluar objetivamente si un Sistema de Inferencia Difusa de tipo Sugeno está funcionando correctamente?

Dado que estos modelos se entrenan con datos, su rendimiento se mide comparando los valores de salida predichos con los valores reales esperados.

Algunos indicadores de error:

◊ MAE (Error absoluto medio):

Valor real de salida del dato x (Target) — **Valor predicho por el modelo (salida para el dato x)**

$$mae = \frac{1}{N}\sum_{x \in D} |t(x) - y(x)|$$

N: Cantidad de datos de D.

◊ RMAE (root mean absolute error): para conocer la magnitud real de los errores √mae

◊ MSE (error medio cuadrático): Minimizarlo

Error de **aproximación** considerando un conjunto de datos D:

Valor real de salida del dato x (Target) — **Valor predicho por el modelo (salida para el dato x)**

$$mse = \frac{1}{N}\sum_{x \in D} [t(x) - y(x)]^2$$

N: Cantidad de datos de D.

◊ RMSE (root medium square error): Para conocer la magnitud real √mse

◊ RSE (errores relativos): mide el error cuadrático del modelo con respecto a un modelo base que siempre predice el valor medio de los targets en el

conjunto de datos.

$$rse = \frac{\sum_{x \in D}[t(x) - y(x)]^2}{\sum_{x \in D}[y(x) - \langle f \rangle]^2}$$

$$\langle f \rangle = \frac{1}{N} \sum_{x \in D} t(x)$$

$$rse = \frac{1}{N} \frac{\sum_{x \in D}[t(x) - y(x)]^2}{\sum_{x \in D}[y(x) - \langle f \rangle]^2}$$

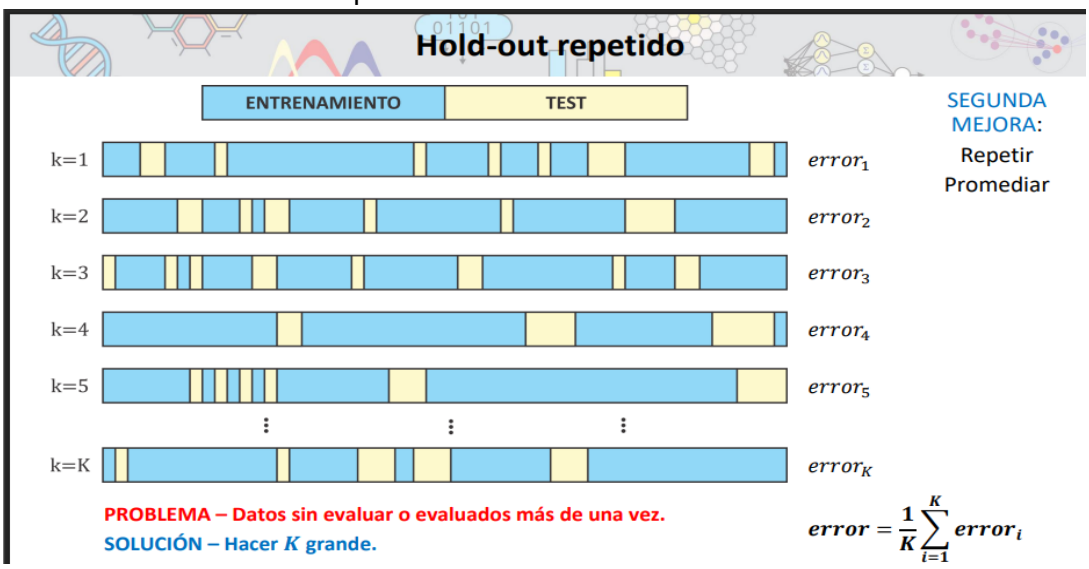
A estas medidas también se les puede aplicar la raíz cuadrada ($rrse$).

$$rse = \frac{\frac{1}{N}\sum_{x \in D}[t(x) - y(x)]^2}{\langle f \rangle}$$

Si el sistema está funcionando correctamente, los valores de estas métricas de error deben ser bajos, lo que indica que la salida del modelo se aproxima a los datos reales de entrenamiento

Metodologías de Validación (Generalización): para comprobar se debe probar con datos que el modelo no utilizó en su entrenamiento. Métodos:

- Error de entrenamiento o resustitución: es la primera aproximación, donde se utiliza la evidencia completa (todos los datos) para el aprendizaje y se usa la misma evidencia para calcular el error.
- Hold-out: consiste en separar el conjunto de datos en dos subconjuntos: uno para entrenar el modelo y otro para calcular el error. El valor del error obtenido depende de la partición.
- Hold-out repetido: para mejorar ele hold-out, se repite la partición y se promedia el erro obtenido en cada repetición.



7) ¿Qué es la capacidad de generalización de un modelo? Defina error de resustitución y error de generalización o de test.

La capacidad de generalización es la habilidad de un modelo de aprendizaje (como un FIS Sugeno, una red neuronal) para producir predicciones correctas sobre datos nuevos, distintos de los usados en su entrenamiento.
Si un modelo solo funciona bien con los datos de entrenamiento, pero mal con datos nuevos →sobreajuste (overfitting)
Si un modelo generaliza bien, significa que capturó las regularidades reales del problema y no simplemente memorizó los ejemplos.
El error de sustitución es el que se obtiene al evaluar los valores dador con los datos de entrenamiento. Se realiza con datos previamente analizados.
En el error de generalización el error se evalúa con datos distintos a los del entrenamiento.

.....

Mamdani

1) **Escriba una regla típica de un Sistema de Inferencia Difusa (FIS) de tipo Mamdani. ¿Cómo se genera un modelo de este tipo?**
   Si X es baja e Y es Alta entonces Z es media.
   SI x1 es A1 Y x2 es A2 ENTONCES y es B
   donde:
   X1, x2, ... de entrada.
   A1, A2, ...: **conjuntos difusos** asociados a etiquetas lingüísticas de los antecedentes.
   y: variable de salida.
   B: **conjunto difuso** asociado al consecuente.

   1. Definir las variables de entrada y salida
   2. Asignar conjuntos difusos a cada variable con sus funciones de pertenencia.
   3. Definir las reglas difusas
   4. Elegir el método de agregación e inferencia (min para AND y max para OR).
   5. Elegir el método de defuzzificación.

2) **Indique los pasos de procesamiento de un dato de entrada hasta llegar a la salida en un FIS de tipo Mamdani.**
   Fuzzificación: convertir valores numéricos de entrada en grados de pertenencia a cada conjunto difuso. Ejemplo: temperatura = 28 °C → 0.7 Media, 0.3 Alta.
   Evaluación de reglas: calcular el grado de activación de cada regla usando los valores de pertenencia. Ejemplo: si regla = “Si temperatura es Alta y humedad es Baja → ventilador Rápido”, aplicamos min (0.3, 0.8) = 0.3.
   Agregación de las salidas difusas: combinar los efectos de todas las reglas en un conjunto difuso de salida único. Usualmente se usa el operador max.
   Defuzzificación: transformar el conjunto difuso de salida en un valor numérico concreto. Método más común centroide (promedio ponderado de los valores posibles).

3) **Reconozca todos los parámetros que definen y podría modificar en un FIS de tipo Mamdani.**
   Los parámetros que podría modificar son:
   - La cantidad de variables.
   - La función de pertenencia.
   - La cantidad de reglas.

- Las condiciones lógicas.
- Operadores de inferencia: And → min o producto; Or → max o suma algebraica.
- Método de defuzzificación

4) **¿Cómo puede evaluar (aunque sea subjetivamente) si un sistema de inferencia difusa de tipo mamdani está funcionando correctamente?**
   Su evaluación puede hacerse de forma cualitativa, observando si el comportamiento del sistema coincide con el razonamiento humano o el conocimiento experto.
   Puede evaluarse con algún profesional del tema el cual diga que tanto se ajustó lo que se hizo a lo que realmente es y que se debería esperar de ciertas entradas. También podríamos probar con casos conocidos.

Algoritmos genéticos

1) **¿Para qué sirve un algoritmo genético?**
   Un algoritmo genético es una técnica de optimización y búsqueda inspirada en la evolución biológica. Su propósito es hallar la mejor solución entre todas las posibles, evaluando y combinando soluciones mediante selección, cruzamiento y mutación. Es un método estocástico.

2) **¿Qué es la función de evaluación?**
   La función de evaluación o de aptitud (fitness) se aplica a todos los individuos y debe ser capaz de identificar las malas soluciones y "premiar" a las buenas, de forma que se propaguen. Me determina que tan buena es la solución encontrada en esa generación, además compara con otras y se queda con la mejor. Es lo que se quiere optimizar.

3) **¿A qué se denomina cromosoma y genes en la jerga de los algoritmos de este tipo?**
   Cromosoma se le denomina al conjunto de datos que hacen al individuo, el cual es una representación de una posible solución al problema que se desea optimizar.)
   Los genes son las unidades básicas que componen un cromosoma (se usan en crossover).

    cada cuadradito es un gen y el conjunto el cromosoma.

4) **Describa el flujo de procesamiento de un Algoritmo Genético convencional. Mencione algunos métodos de selección de padres de cruzamiento y de mutación.**
   1. Se genera una población inicial, puede ser de forma aleatoria o valores previamente obtenidos.
   2. Se elige un método de selección de padres y se aplica a la población inicial.
   3. Con los padres elegidos se elige un método de cruzamiento para cruzar los distintos genes y obtener distintos hijos.
   4. Para que el programa no caiga siempre en los mínimos locales y para explorar nuevas zonas del espacio de búsqueda (da aleatoriedad no se estanque en una

solución simple) se le aplica, con baja probabilidad, a algún individuo una mutación, un cambio en algún gen. Esto se hace ya que el sistema puede estar convergiendo a una buena solución, pero en realidad esta no es la solución óptima.

5. Se ejecuta la función fitness para cada individuo y se realiza un método de selección para quedar con la misma cantidad de individuos que la población inicial. Se elige el criterio de detención, si no se detiene vuelve al 2.

Métodos de selección de padres:

- Roulette
- Elite
- Al azar

Métodos de cruzamiento (crossover):

- Crossover de un punto: se corta en un punto y se intercambian las partes.
- Crossover de dos puntos: se cortan dos posiciones y se intercambia el segmento central.
- Crossover uniforme: cada gen del hijo se toma aleatoriamente de uno de los padres.
- Crossover aritmético: hijo=α·Padre1 + (1–α)·Padre2.

Ejemplos

	P1	P2	H1	H2
Un punto	11010110	01101001	11011001	01100110
	18 -3	9 23	18 23	9 -3
Dos puntos	11010110	01101001	11001010	01110101
Uniforme	11010110	01101001	11111111	11011101

Métodos de mutación:

- Bit flip: en binario cambio un bit (0 por 1 o 1 por 0).
- Mutación uniforme: se reemplaza un gen por un valor aleatorio dentro del rango.
- Mutación gaussiana: se añade un pequeño ruido normal al valor del gen.
- Swap: se intercambian las posiciones de dos genes.
- Mutación de inserción/desplazamiento: se selecciona un gen y se inserta en otra posición.

5) Dé un ejemplo de aplicación de un algoritmo genético. Piense en algo real, quizá relacionado con algo de su trabajo, de su carrera o de sus hobbies.

Problema: optimizar un plan de estudio
Cromosoma: planificación semanal de estudio
Genes: cada hora de estudio por cada materia.
Fitness: mide que tan bueno es el plan, cuanto se dedica a materias más importantes, se fija exceso de horas
Aplicación real: encontrar el mejor plan de estudio, cumpliendo con las horas semanales disponibles y priorizando materias críticas.

6) **¿Podría con algún algoritmo genético optimizar un modelo de sugeno? ¿Cuál podría ser la función de evaluación? ¿Cuál sería el cromosoma?**
SI.
Se pueden optimizar los parámetros de las funciones de pertenencia de las variables de entrada (antecedentes) y los coeficientes (a0, a1, ...) de las funciones lineales del consecuente.
La función de evaluación mide que tan bueno es un conjunto de parámetros respecto a los datos reales. podría ser el error cuadrático medio (mse).
El cromosoma codifica todos los parámetros del modelo Sugeno que se van a optimizar. Cada gen corresponde a un parámetro.
Ej.
2 variables de entrada, 2 reglas, 3 a
Centro y ancho por cada antecedente 2*2= 4 + 3a =7 * R =14 →R*(D+1)*parámetros gauss R*(2D+D+1)
Cromosoma [c11, c12, c21, c22, σ11, σ12, σ21, σ22, a01, a02, a11, a10, a21, a22]

.....

Perceptrón

1) **Indique el modelo matemático de una neurona con salida 0 o 1 (escalón).**
Se calcula:$v = \sum_{i=1}^{N} w_i\, x_i + b$
X → entradas
W → peso de cada entrada
B → umbral
La función de activación produce una salida binaria (0 o 1):
1 si v es >=0 y 0 en el caso contrario. Si a la hora de calcular el error da +1 está mal.

$$y = \begin{cases} 1 \text{ si } v \geq 0 \\ 0 \text{ si } v < 0 \end{cases}$$

V es la recta o hiperplano que los separa y yo busco los pesos para que la recta este bien ubicada.

2) **Escriba la ecuación de discriminación de clases de un perceptrón simple de 2 y de 3 entradas. Dibújela para el caso de 2 entradas, adoptando valores para los pesos y el umbral.**
La función de discriminación de clases es la que define la frontera de decisión, donde la neurona cambia de clase → v=0
Para 2 entradas:

$$v = w_1 x_1 + w_2 x_2 + b = 0$$

Para 3 entradas

$$v = w_1 x_1 + w_2 x_2 + w_3 x_3 + b = 0$$

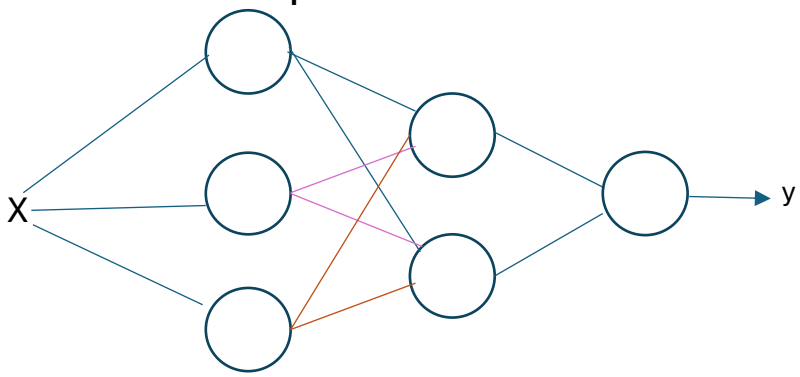
3) **¿Qué ocurre si se quita el umbral al perceptrón simple?**
Al quitarse el umbral la recta siempre pasa por el origen, el perceptrón pierde la capacidad de resolver muchos problemas de clasificación linealmente separables. Si

los datos no pueden ser partidos por una recta que pase por el origen, el perceptrón será incapaz de resolver el problema.

4) **Defina conjunto de datos linealmente separable.**
Un conjunto de datos es linealmente separable si existe una recta que puede separar perfectamente los patrones de las dos clases sin errores de clasificación.

Redes Feedforward

1) **Dibuje una red neuronal feed-forward con 2 capas ocultas de 3 y 2 neuronas, una entrada y una salida. No considere umbrales en las neuronas y considérelas lineales. Plantee la expresión matemática de salida.**



Y=W*x1
W1= [w11
w21
w31]
W2= [w11 w12 w13
w21 w22 w23]
W3= [w11 w12]

Y (1) = w1*x1
Y (2) = y (1) *w2
Y (3) = y (2) *w3 → final

2) **¿Cómo se aconseja que sea la arquitectura de una red neuronal feed-forward para aproximar una función de 3 variables independientes y una variable dependiente? Dibuje una posible arquitectura.**
La arquitectura de una red para aproximar una función de 3 variables independientes y una dependiente es de regresión ya que la salida es una variable continua.
En problemas de regresión, se necesita que la salida sea un valor real continuo (R), no una probabilidad limitad.
X1 O
X2 O O → y
X3 O

3) **¿Cómo entrenaría con un conjunto de datos etiquetado en 3 clases a una red neuronal? Dibuje una posible arquitectura.**
Red de clasificación supervisada.

La salida no es un número real continuo, sino una etiqueta de clase, la red debe aprender a distinguir entre esas 3 clases.
Codificación One hot encoding las etiquetas se convierten en formato numérico
Las características se normalizan para tener una media de 0 y una desviación de 1
El número de neuronas de la entrada debe coincidir con el número de características en los datos de entrada.
Debe haber 3 salidas
Se usa softmax que convierte las salidas en una distribución de probabilidades. La neurona con la probabilidad más alta indica la clase.

X1	O	SO	
X2	O	FT	O → CLASE 1
X3	O	M	O → CLASE 2
X4	O	A	O → CLASE 3
X5	O	X	

4) ¿Qué es la "capacidad de generalización" de una red neuronal (o en general de un sistema con entrenamiento supervisado)?

Es la habilidad que tiene la red de responder correctamente a datos nuevos, no vistos durante el entrenamiento. Uno de los problemas relacionados con esto es que se produzca un overfitting, red sobreentrenada con los valores de entrenamiento, pero a la hora de estimar con los datos de validación lo más probable es que funcione mal. Con una buena generalización aprende los patrones esenciales para el buen funcionamiento.

5) ¿Qué son los datos de entrenamiento, validación y test? ¿En qué etapa actúa cada uno de ellos y en qué influyen?

Los datos de entrenamiento se utilizan en la etapa inicial, con ellos se entrena a la red haciendo que esta aprenda, ajustando los pesos internos, para poder estimar valores nuevos. (Se intenta ajustar el modelo a los datos lo justo y necesario para que funcione correctamente con nuevos valores).

Los datos de validación se utilizan en cada época del modelo y lo que hacen es comprobar que el modelo haya aprendido bien, se comparan las salidas esperadas con las obtenidas, se obtiene un error, que permite evaluar el desempeño del modelo sobre datos que no intervienen en el ajuste de los pesos. Permiten detectar overfitting (si el error aumenta mientras el entrenamiento baja). Se pueden utilizar para detener el entrenamiento (early stopping).

Los datos de test son los que se utilizan al final del entrenamiento para comprobar si la red generaliza correctamente y es útil para estimar/calcular valores nuevos.

6) ¿Qué similitudes y diferencias hay entre un modelo de Sugeno y una red neuronal multicapa para aproximar una función?

Similitudes:

- ✓ Los dos son modelos de regresión. Buscan aproximar una función a partir de datos de entrenamiento.
- ✓ En sugeno cada regla actúa como una neurona. DUDOSO
- ✓ Ambos se entrenan con datos → aprendizaje supervisado.
- ✓ Ambos buscan minimizar el error de estimación

Diferencias:

- Sugeno basado en reglas IF-THEN difusas. Redes basado en pesos y umbrales distribuidos a través de múltiples capas de neuronas.
- En sugeno es consecuente es una función lineal de las variables de entrada. En redes la salida es el resultado de la función de activación de la última neurona.
- Aprendizaje: sugeno →ajuste de reglas y parámetros de funciones de pertenencia. Redes → backpropagation (descenso de gradiente) para ajustar los pesos y umbrales.
- Sugeno es fácil de interpretar (reglas lingüísticas), las redes son más difíciles (caja negra) es difícil extraer reglas entendibles.
- Sugeno requiere una estructuración inicial del conocimiento (definición de términos lingüísticos y reglas) que puede venir de expertos. Mientras que redes necesita una gran cantidad de datos para aprender sin conocimiento previo.