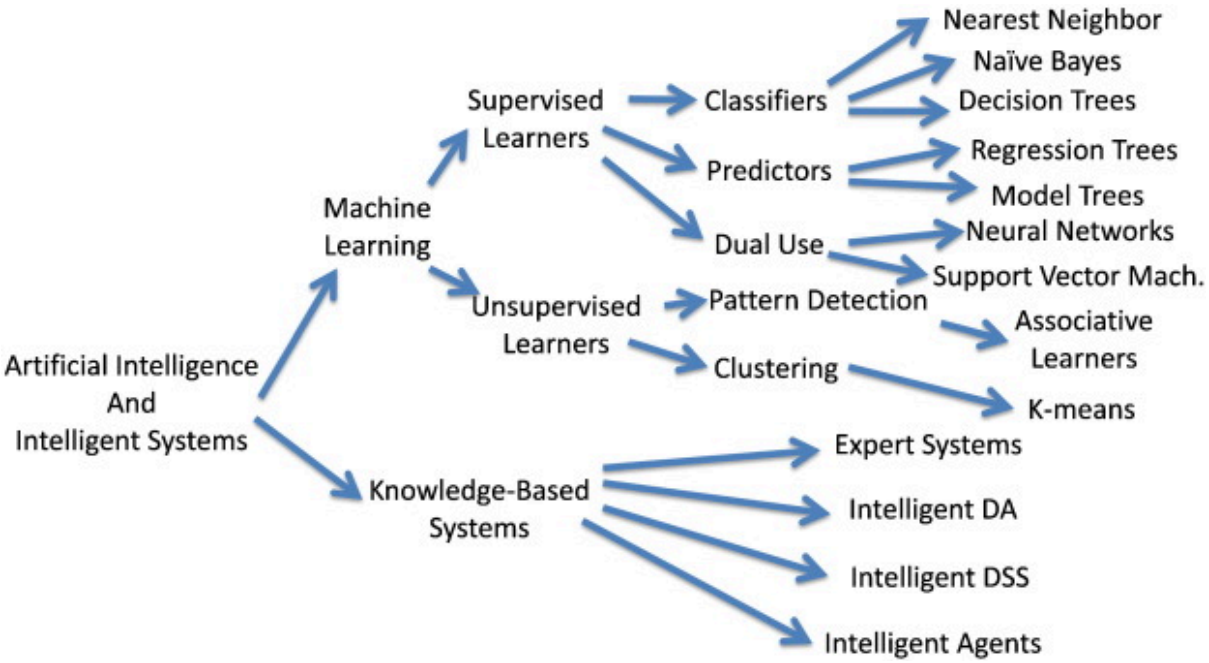
LIC. EN SISTEMAS DE INFORMACIÓN

INTELIGENCIA ARTIFICIAL



PARTE 1 - Introducción a la clasificación

La clasificación en inteligencia artificial (IA) es un tipo de aprendizaje supervisado donde un algoritmo aprende a asignar etiquetas o categorías a instancias de datos basándose en características de entrada. Se utiliza en una amplia gama de aplicaciones, desde el reconocimiento de imágenes hasta la filtración de correos electrónicos y la predicción de decisiones de crédito

Ejemplo: Clasificador de Correos Electrónicos

Supongamos que quieres crear un sistema que automáticamente clasifique los correos electrónicos entrantes en dos categorías: "**spam**" y "**no spam**". Esto ayudará al usuario a enfocarse solo en los mensajes importantes y evitar perder tiempo con correos no deseados.

Para clasificar correos electrónicos automáticamente en categorías de "spam" y "no spam", se pueden considerar diferentes tipos de características, ejemplo:

1. **Frecuencia de palabras**: La frecuencia y distribución de ciertas palabras pueden indicar spam. Los correos spam a menudo repiten palabras que suenan atractivas para incentivar clics.
2. **Palabras clave**: Algunas palabras o frases son comúnmente asociadas con spam, como "oferta", "gratis", "promoción exclusiva", "haz clic aquí", etc.

¿Cómo funciona la clasificación en IA?

- **Recopilación de datos**: Se recopila un conjunto de datos que incluye ejemplos de entrada y sus correspondientes etiquetas.
- **Preprocesamiento de datos:** Los datos se limpian y se transforman para facilitar el aprendizaje del algoritmo. Esto puede incluir la normalización, la gestión de datos faltantes y la codificación de variables categóricas.
- **División de datos**: El conjunto de datos se divide en un conjunto de entrenamiento y un conjunto de pruebas. El conjunto de entrenamiento se utiliza para enseñar al modelo, y el conjunto de prueba para evaluar su rendimiento.
- **Selección de un modelo**: Se selecciona un algoritmo de clasificación adecuado para el problema y los datos disponibles. Algunos ejemplos populares incluyen árboles de decisión, redes neuronales y máquinas de vectores de soporte.
- **Entrenamiento del modelo**: El modelo aprende a clasificar las entradas basándose en los datos de entrenamiento. Aprende patrones y asociaciones entre las características de los datos y las etiquetas.
- **Evaluación del modelo:** Se evalúa el modelo utilizando el conjunto de prueba para verificar su precisión y capacidad de generalización. Esto se hace comparando las predicciones del modelo con las etiquetas reales.
- **Ajuste y optimización**: Basado en la evaluación, el modelo puede ser ajustado para mejorar su rendimiento. Esto puede implicar la modificación de parámetros, la elección de diferentes características o el uso de técnicas de regularización.

Métodos comunes de clasificación

- **Árboles de decisión**: Utilizan una estructura de árbol donde cada nodo representa una característica del dato, cada rama representa una regla de decisión, y cada hoja representa un resultado de la clasificación.
- **Regresión logística**: A pesar de su nombre, es un modelo de clasificación que estima la probabilidad de que una instancia pertenezca a una categoría particular.
- **Máquinas de vectores de soporte** (SVM): Encuentran un hiperplano en un espacio de muchas dimensiones que clasifica los datos de manera óptima separando las categorías con la mayor distancia posible.
- **Redes neuronales**: Sistemas que imitan el funcionamiento del cerebro humano para reconocer patrones y características en los datos.
- **K-vecinos más cercanos (KNN)**: Clasifica una instancia basándose en la mayoría de votos de sus K vecinos más cercanos, con K siendo un número predefinido.

EJERCICIOS:

1. Presentar 3 problemas de clasificación, con su dataset, explique cuál es la variable a clasificar y cuales son las variables que se utilizan para la clasificación.
2. Investigar que es un clasificador en IA y cómo funciona.
3. Investigar métricas que se utilizan para elegir entre clasificador.
4. Investigar por qué se dividen los datos para entrenar los clasificadores, que es el sesgo y la varianza en contexto de IA?
5. Para cada uno de los puntos, desarrollarlo con el fin de defenderlo posteriormente.

PARTE 2 - ¿Cómo determinar las variables más importantes en un problema de clasificación?

Un ejemplo clásico de clasificación en inteligencia artificial es el diagnóstico de tumores como benignos o malignos a partir de características médicas extraídas de imágenes de biopsias.

Características de interés

Algunas características de las células en las imágenes pueden ser particularmente reveladoras sobre la naturaleza del tumor:

1. Textura: La variabilidad en la tonalidad de gris puede indicar desorden en la estructura celular, típico en tumores malignos.
2. Forma y bordes de la célula: Los tumores malignos a menudo tienen bordes irregulares o espiculados, mientras que los benignos tienden a ser más redondos y uniformes.
3. Tamaño de la célula: Un aumento en el tamaño y variabilidad del tamaño celular puede indicar malignidad.
4. Núcleos celulares: Núcleos grandes y atípicos son comunes en las células cancerosas.

Algunas características pueden no contribuir significativamente a la clasificación o podrían ser redundantes debido a la correlación con otras características más informativas:

1. Color general de la imagen: Puede ser influenciado por la técnica de tinción y no necesariamente por la naturaleza del tejido.
2. Área total de la imagen: No distingue entre tejido normal y tumoral.

EJERCICIOS

Investigar

1. Correlación entre características y etiquetas: Puedes calcular el coeficiente de correlación entre cada característica (variable independiente) y la etiqueta (variable dependiente) del problema de clasificación. Esto puede ayudarte a entender qué características tienen una relación más fuerte con la etiqueta

2. Random Forest
Los modelos basados en árboles de decisión, como los árboles de decisión, bosques aleatorios (Random Forest) y Gradient Boosting Machines, ofrecen una evaluación directa de la importancia de las características. La importancia se mide generalmente por el grado en que cada atributo mejora el rendimiento del modelo, como la pureza de los nodos, y se calcula durante la construcción del árbol.

3. Métodos de eliminación recursiva de características (RFE)
RFE trabaja eliminando iterativamente las características menos importantes. Se entrena un modelo con el conjunto completo de características, se evalúa la importancia de cada una y se elimina la menos significativa. Este proceso se repite hasta que se alcanza el número deseado de características.