COLOMBIA POTENCIA DE LA VIDA

TIC

Lección 2
Algoritmos de clasificación

TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

Universidad Tecnológica de Pereira

Tiempo de ejecución: 4 horas

PLANTEAMIENTO DE LA SESIÓN	MATERIALES
La clasificación es una tarea fundamental en el aprendizaje automático supervisado, donde el objetivo es asignar una etiqueta o clase a un objeto basándose en sus características observadas. Por ejemplo, podemos clasificar correos electrónicos como spam o no spam, pacientes como enfermos o sanos, o imágenes como gatos o perros.	

En la clasificación se tienen unas clases predeterminadas y atributos a partir de los cuales queremos inferir la clase. Similar al caso de regresión, lo que se intenta es particionar los datos en dos o tres subconjuntos que permiten entrenar un modelo y evaluar su comportamiento. Algunos de los modelos más empleados son:

- Máquinas de soporte vectorial
- K nearest neighbour
- Regresión logística.

Máquinas de soporte vectorial:

En el aprendizaje automático, las máquinas de vectores de soporte (SVM) son modelos de aprendizaje supervisados con algoritmos de aprendizaje asociados que analizan los datos utilizados. Pueden emplearse para la clasificación y la regresión. Dado un conjunto de ejemplos de entrenamiento, cada uno marcado como perteneciente a una u otra categoría, un algoritmo de entrenamiento SVM crea un modelo que asigna nuevos ejemplos a una categoría u otra, convirtiéndolo en un clasificador lineal binario no probabilístico.

Además de realizar una clasificación lineal, las SVM pueden realizar una clasificación no lineal de manera eficiente utilizando kernels. Estos son representaciones de un sistema numérico que pueden agregar no linealidad y solucionar problemas de clasificación en espacios de alta dimensionalidad.

En ocasiones los datos no están etiquetados, allí se necesita un enfoque de aprendizaje no supervisado, que depende de ejecutar algoritmos de clústering. Estos algoritmos se encargan de crear grupos, los que pueden servir como etiquetas para un algoritmo de clasificación. En el caso de las SVMs se tienen algoritmos de agrupamiento de vectores que permiten realizar los grupos de forma automática.

La "Máquina de vectores de soporte" es un algoritmo de aprendizaje automático supervisado que se puede utilizar para problemas de clasificación o regresión. Sin embargo, se utiliza principalmente en problemas de clasificación. En este algoritmo, cada elemento de datos se traza como un punto en el espacio n-dimensional (donde n es el número de características que tiene) siendo el valor de cada característica el valor de una coordenada particular. A continuación, la clasificación se realiza encontrando el hiperplano que distingue bastante bien las dos clases.

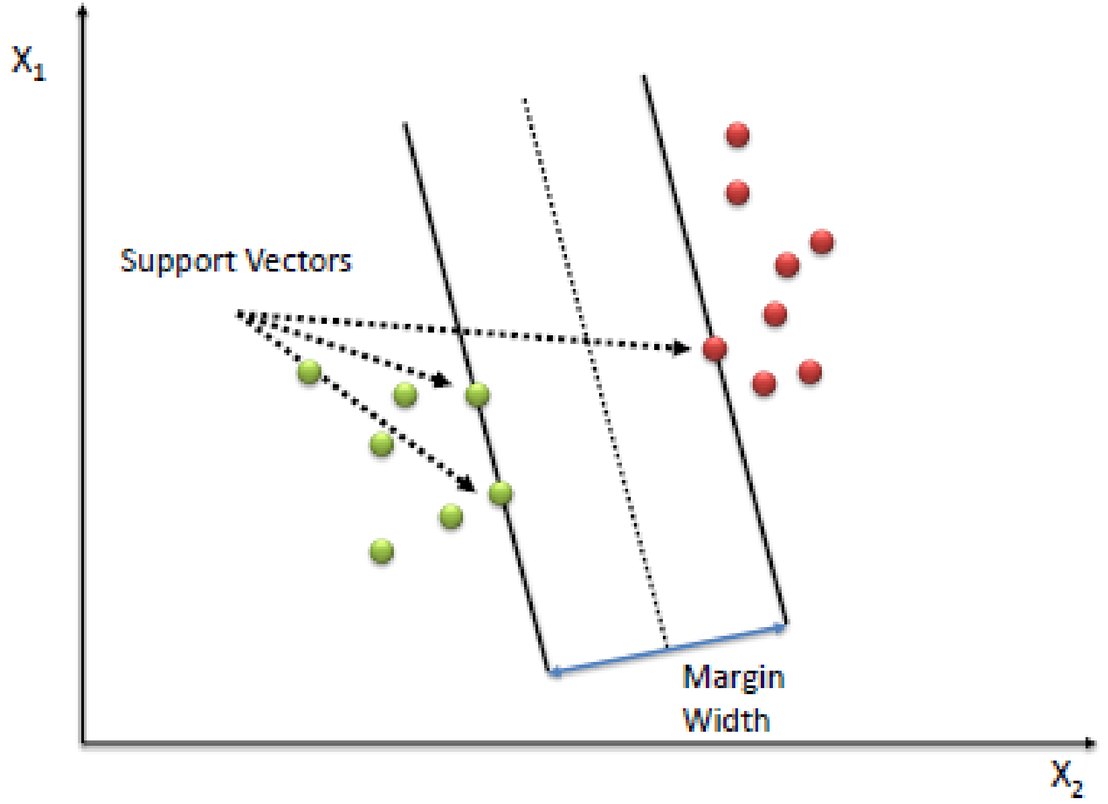

Figura 1: vectores de soporte y superficie de decisión.

En esencia, los vectores de soporte son las coordenadas de una observación. El concepto de máquina de vectores de soporte es un límite que permite separar efectivamente entre dos clases. Este límite puede ser una curva o un hiperplano, dependiendo de las dimensiones de los datos. Su utilidad en clasificación se puede ver en incluso en sistemas de reconocimiento de voz y reconocimiento facial (asignar la cara a una clase que es determinada persona).

Ejercicio orientado en clase:

Para explicar mejor el funcionamiento de los algoritmos de clasificación, emplearemos el cuaderno de jupyter llamado clasificación.ipynb. En este cuaderno encontraremos los ejemplos de la clase que emplean modelos de clasificación con algunos conjuntos de datos conocidos como es el caso de Iris (link: https://archive.ics.uci.edu/dataset/53/iris).

Al iniciar, realizaremos un breve análisis exploratorio de los datos. Esto con el fin de conocer cómo se distribuyen los conjuntos de datos y tener una visión del contenido. Al hacerlo podremos abordar mejor la creación de los modelos de clasificación.

En la sección 1.1 analizaremos la distribución de las variables, separadas por clases en diagramas de violín como el de la figura 1. Estos diagramas permiten ver la distribución de los datos y la densidad de probabilidad, es un gráfico que combina las características del diagrama de cajas y bigotes con un diagrama de densidad rotado 90 grados y colocado a cada lado para mostrar la forma de distribución de los datos. La barra negra en el centro representa el intervalo intercuartil, la barra negra fina representa el 95% de los intevalos de confianza y el punto blanco se sitúa en la mediana. En este caso es posible saber si una distribución es unimodal o multimodal, cosa que en los diagramas de cajas y bigotes no se puede saber.

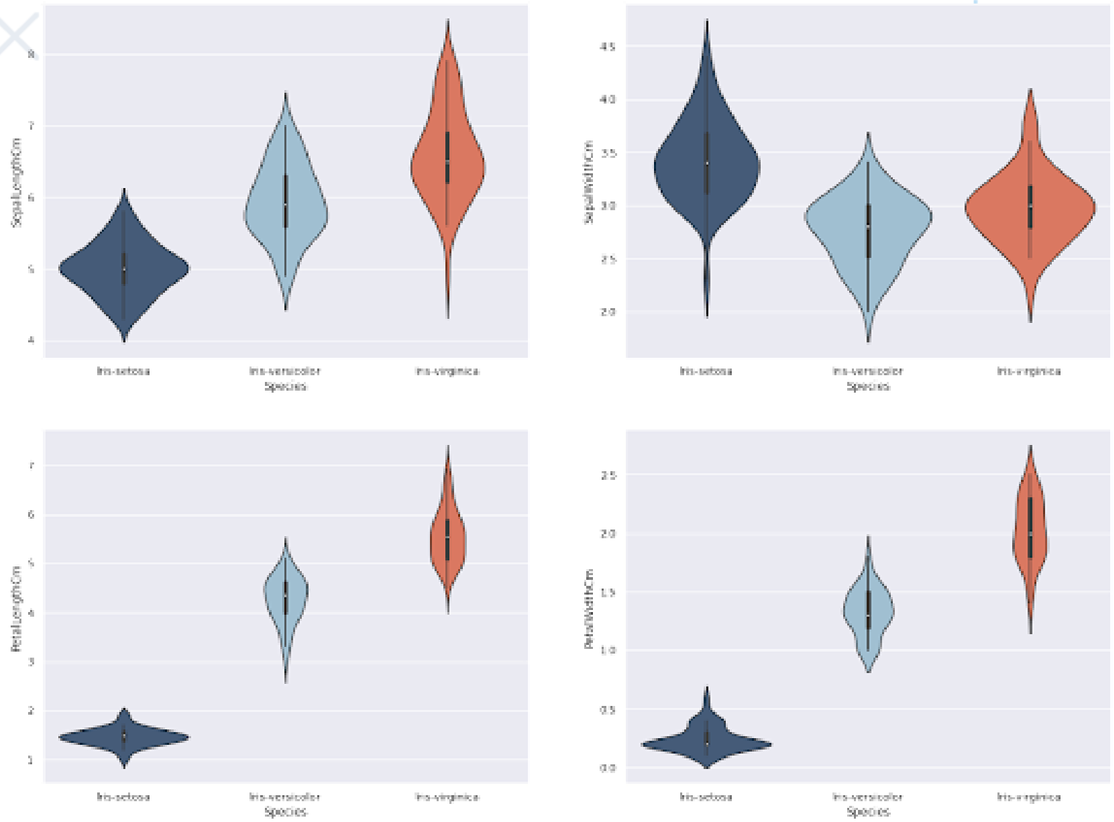

Figura 1: diagramas de violín a generar para el dataset Iris.

Analizando el diagrama vemos que la especie iris setosa tiene los pétalos más pequeños (alto y ancho) con lo que podría ser fácilmente clasificable por medio de estas características.

También podemos emplear la función pairplot de seaborn que permite combinar gráficos de varias variables de forma que podamos comparar los atributos y tener un entendimiento amplio de su comportamiento, separado en colores por clases, como en la figura 2.

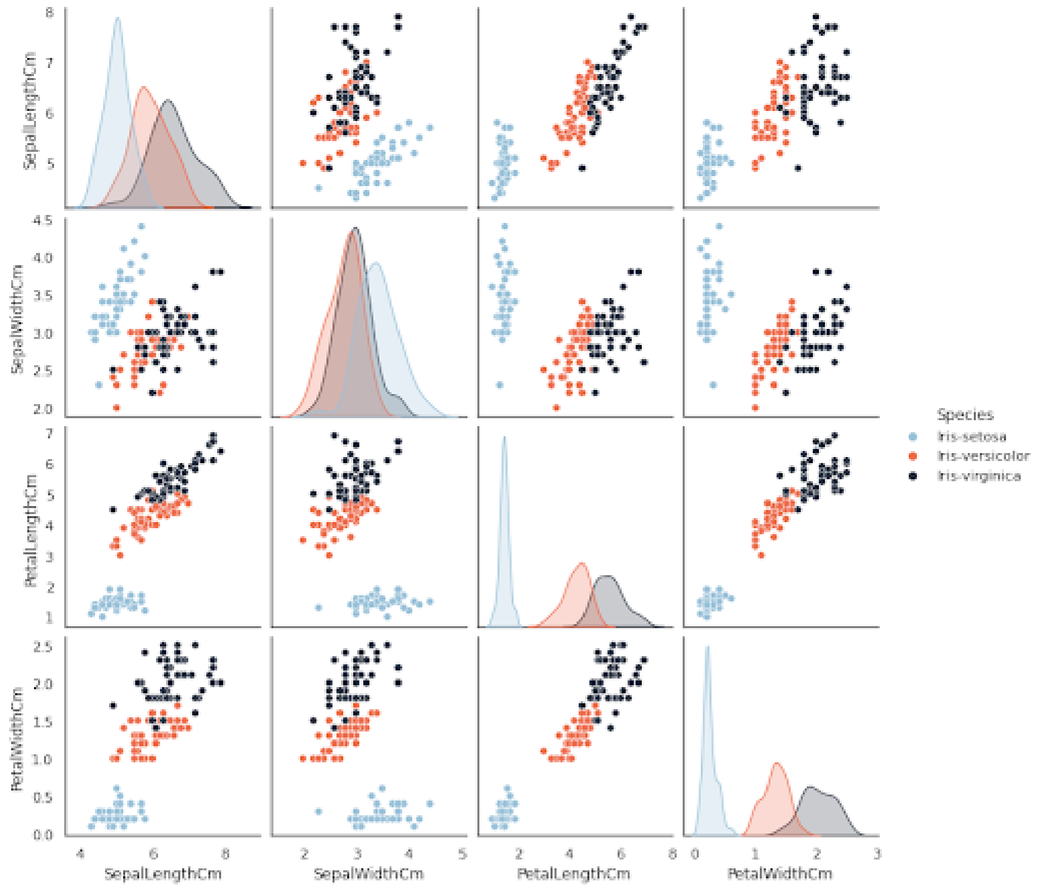

Figura 2: pairplot del dataset iris.

Luego preprocesaremos los datos y emplearemos una máquina de vectores de soporte para la clasificación.

Un clasificador SVM lineal simple conecta dos clases trazando una línea recta entre ellas. Es decir, todos los puntos de datos de un lado de la línea se asignarán a una categoría, mientras que los puntos de datos del otro lado de la línea se asignarán a una categoría diferente. Esto implica que puede haber un número ilimitado de líneas entre las que elegir.

Lo que distingue al método SVM lineal de otros algoritmos, como los k vecinos más cercanos, es que selecciona la línea óptima para categorizar los puntos de datos.

Tipos de SVM

Las SVM se clasifican en dos tipos, cada una de las cuales se utiliza para un propósito diferente:

- SVM simple: este tipo de SVM se usa comúnmente para tareas de clasificación y regresión lineal.
- Kernel SVM: tiene flexibilidad adicional para datos no lineales, ya que puede ajustarse a un hiperplano en lugar de a un espacio bidimensional.

En la sección 2 del cuaderno de clasificación.ipynb se encontrarán las funciones para entrenar una SVM, calcular la matriz de clasificación

K nearest neighbours

K-NN es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión. Es uno de los algoritmos más simples y versátiles en el campo del aprendizaje automático. La idea principal detrás de K-NN es bastante intuitiva: clasificar o predecir el valor de un nuevo punto de datos basándose en los puntos de datos más cercanos que ya han sido etiquetados o tienen valores conocidos.

Para clasificar un nuevo punto de datos, primero calculamos su distancia a todos los puntos de datos en el conjunto de entrenamiento. La distancia puede calcularse usando diferentes métricas, como la distancia euclidiana o la distancia Manhattan.

Después de calcular las distancias, seleccionamos los "K" puntos de datos más cercanos al nuevo punto. "K" es un parámetro que se elige antes de aplicar el algoritmo y representa el número de vecinos más cercanos que se utilizarán para clasificar o predecir el nuevo punto.

En el caso de clasificación, los K vecinos votan por la clase a la que pertenecen. La clase más común entre los vecinos se asigna al nuevo punto de datos como su clase predicha. Finalmente, evaluamos el rendimiento del algoritmo utilizando métricas como precisión, F1-score (en clasificación) o error cuadrático medio (en regresión).

Knn es un algoritmo fácil de entender e implementar, no requiere entrenamiento explícito, funciona bien con datos no lineales y es robusto a datos ruidosos. Sin embargo, computacionalmente costoso en conjuntos de datos grandes, sensible a la elección de la métrica de distancia y sensible a la escala de las características.

En la sección 2.2 se puede practicar la clasificación con el algoritmo knn en Python.

Regresión logística

Es un método para determinar la relación entre dos factores de datos. Con esta relación se predicen los valores (clase a la que pertenece un elemento).

Si bien el nombre dice regresión, es un algoritmo principalmente empleado en clasificación y es ampliamente utilizado debido a que es simple, es decir, no se requiere un conocimiento avanzado de machine learning para poder implementar el algoritmo. También es un algoritmo rápido, lo que implica menos costo computacional cuando se requiere procesar grandes cantidades de datos.

La regresión logística emplea la función logística (logit) que está definida como:

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

Al graficar el dominio y el rango de la función se ve como la figura 1, en donde se aprecia que los valores de salida de la regresión logística fluctúan entre 0 y 1. Es decir, “comprime” todo el rango de los reales en el rango de 0 y 1. Esa característica permite que las salidas del modelo respondan a preguntas como “si” o “no” lógicos (salidas cercanas a cero o a uno), con lo que puede clasificar datos si pertenecen o no a una clase, maximizando la separación entre clases.

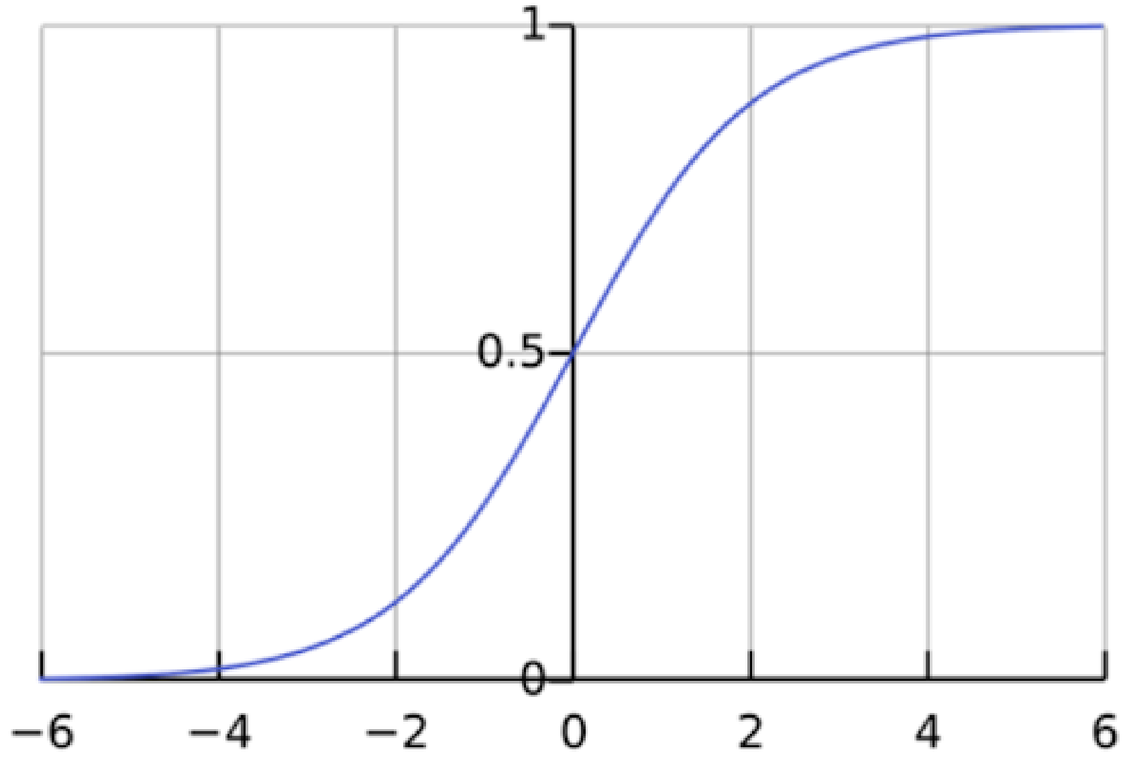

Figura 1: regresión logística.

En la sección 2.3 del cuaderno clasificación.ipynb se explicará la implementación del modelo sobre el set de datos de Iris.