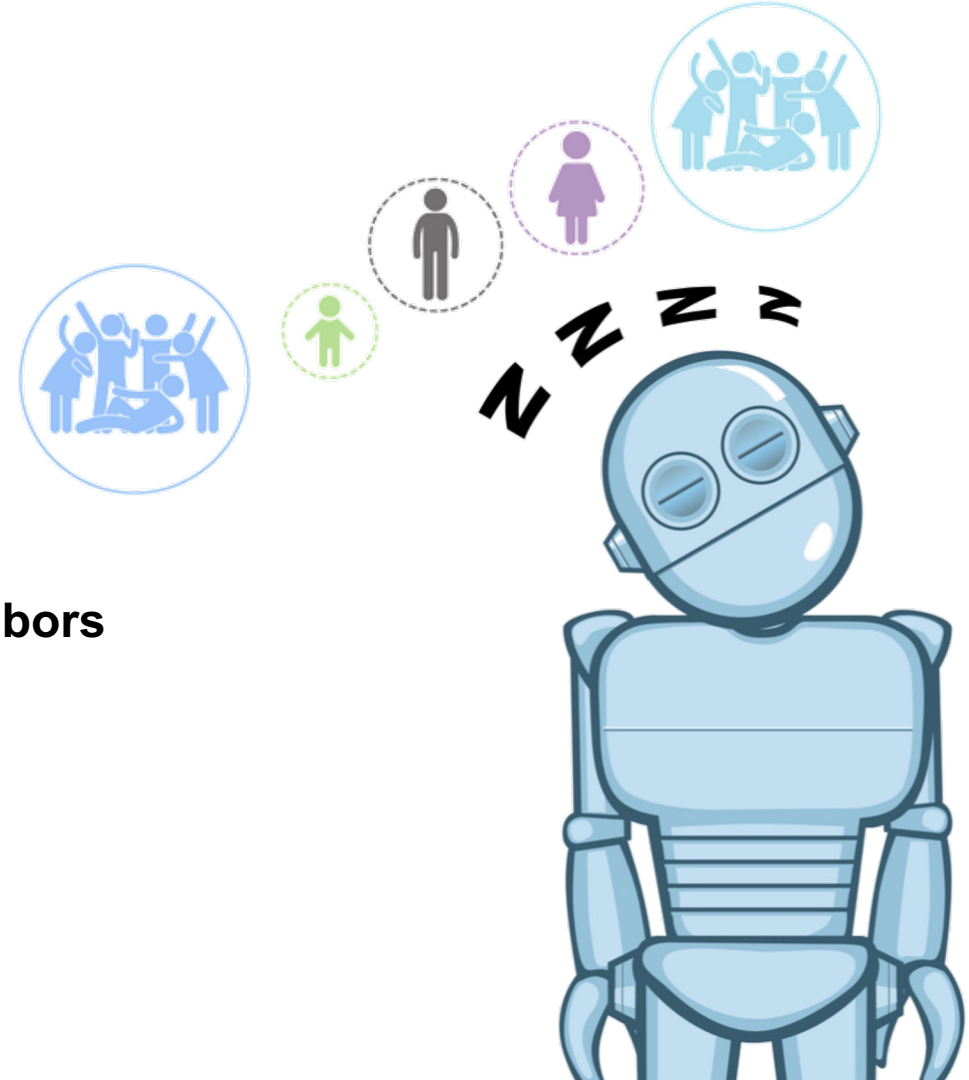
Capítulo 05

Lazy Learning

Classification Using Nearest Neighbors

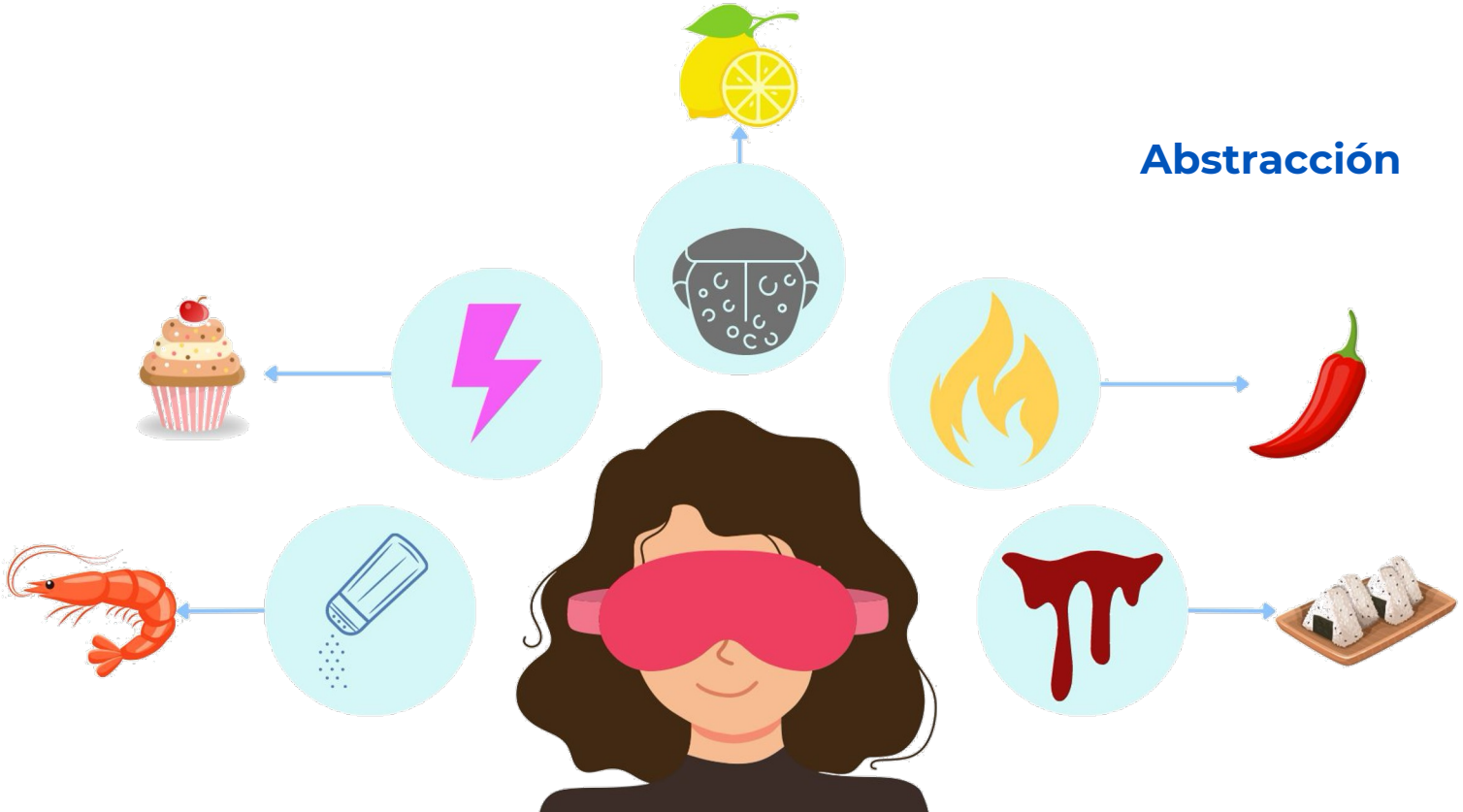
Creado por: Ing. Marcela Parra, Mg



5.1. Introducción.

Lazy Learning

5.1. Introducción.



Lazy Learning

5.1. Introducción.

- Las cosas que son parecidas tienen **propiedades similares.**

- Este principio permite **clasificar datos** colocándolos en la categoría con los **vecinos más similares o "más cercanos".**

Lazy Learning

5.1. Introducción.

“Pájaros del mismo plumaje se juntan”

Lazy Learning

5.2. Objetivos

- Definir conceptos clave que definen a los **clasificadores de vecinos más cercanos** y por qué se les considera **estudiantes "vagos".**
- Medir la similitud de dos ejemplos **usando la distancia**.
- Implementa el **algoritmo k-Vecinos** más cercanos (kNN) para diagnosticar el cáncer de mama
- Analizar el enfoque **kNN** con **Big Data**.

5.4. Algoritmo kNN

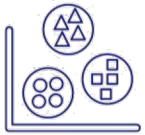
Fortalezas	Debilidades
• Sencillo y eficaz. • **No hace suposiciones** sobre la distribución de datos subyacente. • Fase de **entrenamiento rápido**.	• **No produce un modelo**, lo que limita la capacidad de encontrar ideas novedosas en las relaciones entre características • Fase de **clasificación lenta.** • Requiere una **gran cantidad de memoria** • Las **características nominales** y los datos faltantes requieren **procesamiento adicional.**

5.5. Pasos del Algoritmo kNN

PREPARAR DATOS

Datos de entrenamiento **etiquetados** y un conjunto de datos de prueba **sin etiquetar**

01



ELEGIR K

Determine el **número k de vecinos más cercanos** a considerar.

02

CALCULAR SIMILITUD

identifique los **k registros más similares** en los datos de entrenamiento.

03

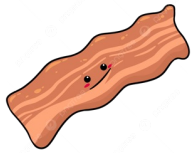
ASIGNAR CLASE

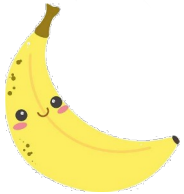
Cada instancia de prueba sin etiquetar la **clase mayoritaria entre sus k vecinos más cercanos.**

04

INTRODUCCIÓN A ML

5.6. Ejemplo Algoritmo kNN





Ingrediente	Dulzura	Crujiente	Tipo de Comida
Manzana	10	9	Fruta
Tocino	1	4	Proteína
Banana	10	1	Fruta
Zanahoria	7	10	Vegetales
Apio	3	10	Vegetales
Queso	1	1	Proteína

INTRODUCCIÓN A ML

5.6. Ejemplo Algoritmo kNN

- El algoritmo **kNN** trata las características como **coordenadas en un espacio** de **características multidimensional**.

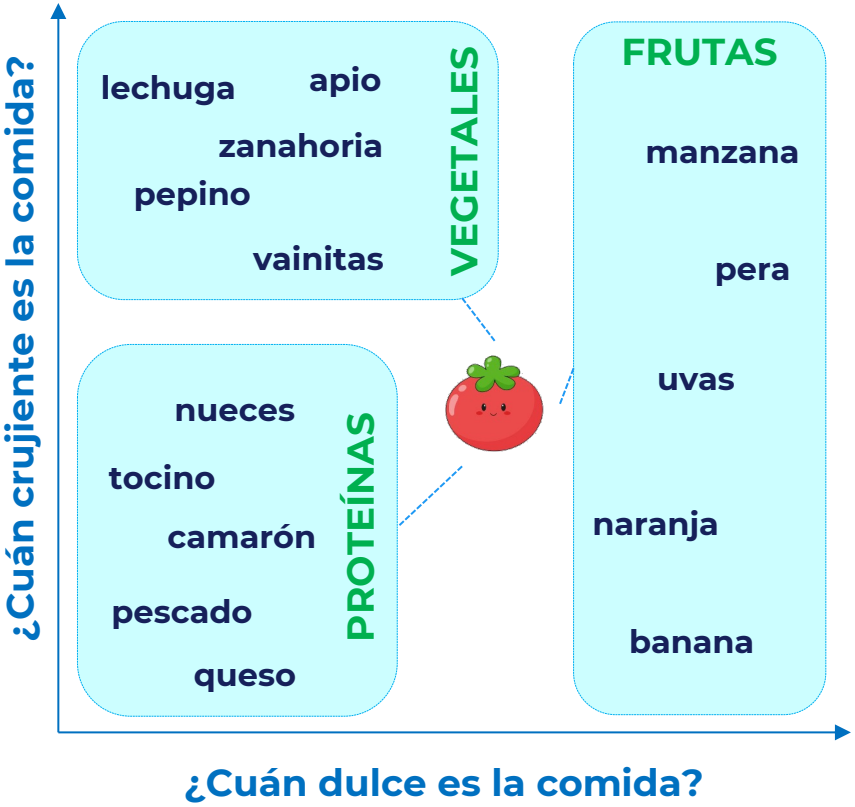
- El conjunto de datos incluye solo **dos características**, el espacio de **características es bidimensional**.

- Podemos trazar datos bidimensionales en un **diagrama de dispersión,** donde la dimensión x indica el dulzor del ingrediente y la dimensión e indica el carácter crujiente.

5.6. Ejemplo Algoritmo kNN

5.6. Ejemplo Algoritmo kNN

5.6. Ejemplo Algoritmo kNN



5.7. Cálculo de la Distancia Euclidiana

- El algoritmo kNN utiliza la **distancia euclidiana** para medir la **similitud entre puntos**, representando la **ruta directa más corta entre ellos**.
- La distancia se calcula a partir de **p y q, que son ejemplos para comparar**; cada uno tiene **n características**.

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 - (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

- Donde:
- p_1 primer valor de la primera característica del ejemplo p,
- q_1 primer valor de la primera característica del ejemplo q.

5.7. Cálculo de la Distancia Euclidiana

- La fórmula de la distancia **compara los valores de cada característica.** La distancia entre el **tomate** (dulzura=6, crujiente =4) y **vainita** (dulzura=3, crujiente =7) y se aplicaría:

$$dist(tomate, vainita) = \sqrt{(6-3)^2 - (4-7)^2} = 4.2$$

Ingrediente	Dulzura	Crujiente	Tipo de Comida	Distancia al tomate
Uva	8	5	Fruta	$\sqrt{(6-8)^2 - (4-5)^2} = 2.2$
Vainita	3	7	Vegetal	$\sqrt{(6-3)^2 - (4-7)^2} = 4.2$
Nueces	3	6	Proteína	$\sqrt{(6-3)^2 - (4-6)^2} = 3.6$
Naranja	7	3	Fruta	$\sqrt{(6-7)^2 - (4-3)^2} = 1.4$

5.7. Cálculo de la Distancia Euclidiana

- La fórmula de la distancia **compara los valores de cada característica.** La distancia entre el **tomate** (dulzura=6, crujiente =4) y **vainita** (dulzura=3, crujiente =7) y se aplicaría:

$$dist(tomate, vainita) = \sqrt{(6-3)^2 - (4-7)^2} = 4.2$$

Ingrediente	Dulzura	Crujiente	Tipo de Comida	Distancia al tomate
Uva	8	5	Fruta	$\sqrt{(6-8)^2 - (4-5)^2} = 2.2$
Vainita	3	7	Vegetal	$\sqrt{(6-3)^2 - (4-7)^2} = 4.2$
Nueces	3	6	Proteína	$\sqrt{(6-3)^2 - (4-6)^2} = 3.6$
Naranja	7	3	Fruta	$\sqrt{(6-7)^2 - (4-3)^2} = 1.4$

$$k = 1 \rightarrow 1NN$$

5.8. Selección del valor de k

- El equilibrio entre el sobreajuste y el desajuste de los datos de entrenamiento es un problema conocido como **equilibrio entre sesgo y varianza.**

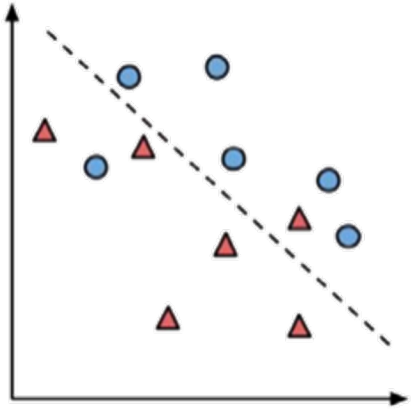
- Elegir una **k grande reduce el impacto o la variación** causada por **datos ruidosos**, pero puede sesgar al de tal manera que corre el riesgo de ignorar patrones pequeños pero importantes.

- El uso de un único **vecino más cercano** permite que **datos ruidosos o valores atípicos** influyan indebidamente en la **clasificación de los ejemplos**.

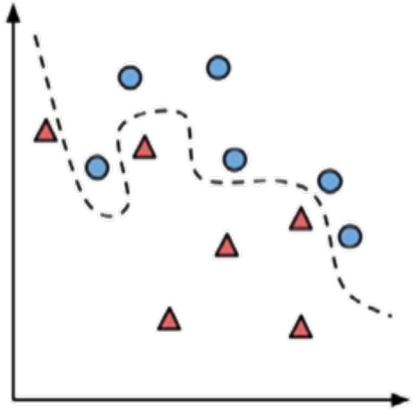
5.8. Selección del valor de k

- Una práctica común es establecer **k igual a la raíz cuadrada del número de ejemplos** de entrenamiento.

- Un enfoque alternativo es probar **varios valores de k y elegir el que ofrezca el mejor rendimiento de clasificación**.



k grande

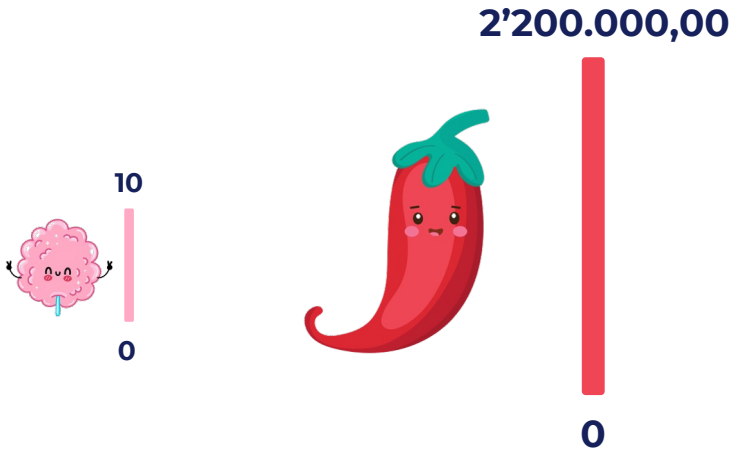


k pequeño

5.8. Preparación de los datos

- El motivo de este paso es que la fórmula de la distancia **depende de cómo se miden las características**.

- Se requiere **"reducir" o reescalar** las diversas características de manera que cada una contribuya de **manera relativamente equitativa a la fórmula de la distancia**.



5.8. Preparación de los datos

1. **Normalización mínima-máxima**, y, transforma una característica en un rango entre 0 y 1.

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)}$$

2. **Estandarización de puntuación z**. La escala de cada uno de los valores de una característica en términos de cuántas desviaciones estándar cae por encima o por debajo del valor medio. El valor resultante se llama puntuación z.

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{StdDev(X)}$$

5.8. Preparación de los datos

Para calcular la distancia entre características **nominales**, se debe convertir a un **formato numérico**.

$$mujer = \begin{cases} 1 & if\ x = mujer \\ 0 & de\ otra\ forma \end{cases}$$

$$calor = \begin{cases} 1 & if\ x = calor \\ 0 & de\ otra\ forma \end{cases}$$

$$medium = \begin{cases} 1 & if\ x = medium \\ 0 & de\ otra\ forma \end{cases}$$

5.9. ¿Por qué el algoritmo kNN es flojo?

- Es un **algoritmo de aprendizaje diferido** porque, técnicamente hablando, **no produce ninguna abstracción**.

- Esto permite que la **fase de entrenamiento ocurra muy rápidamente**, con una posible desventaja de que el **proceso de hacer predicciones tiende a ser relativamente lento**.

- El algoritmo basado en instancias **no construye un modelo.**

Resumen

1. En este capítulo, se explicó el **algoritmo de clasificación k-vecinos más cercanos (kNN)**, que **almacena los datos** de entrenamiento sin realizar aprendizaje.

1. kNN asigna etiquetas a ejemplos de prueba **comparándolos** con los registros más similares mediante una **función de distancia**.

1. A pesar de ser sencillo, kNN puede abordar **tareas complejas**, como **identificar masas cancerosas** con una precisión del 98% usando código en R.

1. Luego tratará sobre la agrupación con **k-medias, un método relacionado con kNN**.