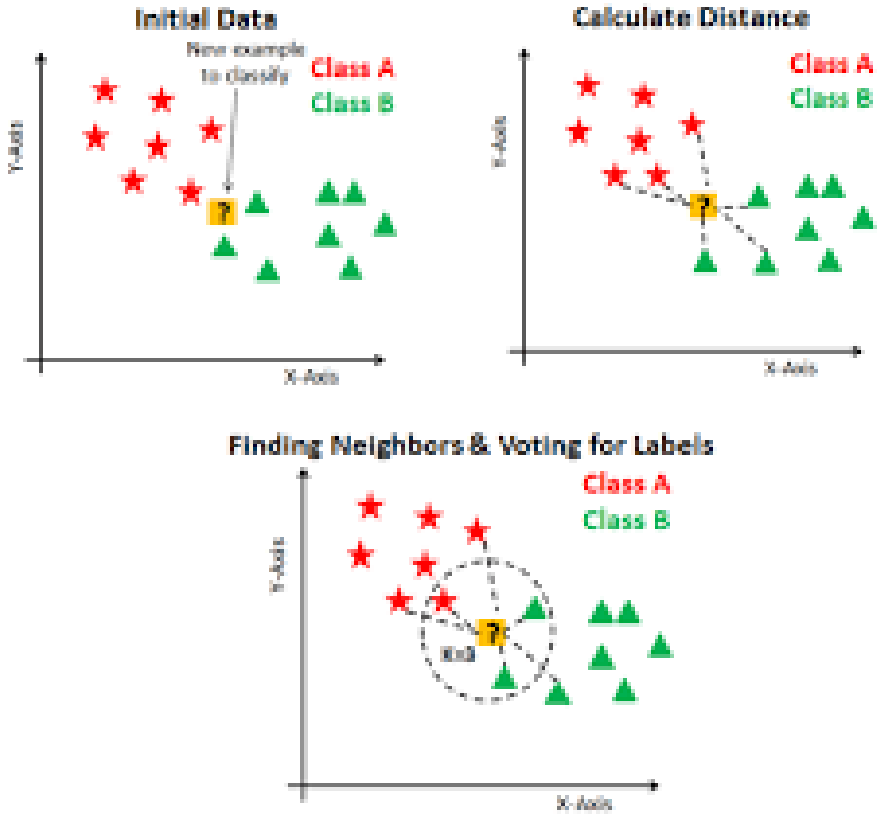


K-Nearest Neighbors

Jonnatan Arias Garcia – Jonnatan.arias@utp.edu.co – jariasg@uniquindio.edu.co

David Cardenas peña - dcardenasp@utp.edu.co

Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Contenido

- Introducción
- La idea del Knn
- Como funciona Knn
- Métrica de clasificación
- Ejemplo de Knn
- El mejor K
- Consideraciones
- Limitaciones
- Conclusiones
- Ventajas – Desventajas

KNN

K vecinos mas cercano, en una técnica de aprendizaje automático supervisado enfocado en tareas de regresión y clasificación.

La idea detrás de la técnica:

- Si tienes un amigo cercano y pasas la mayor parte de tu tiempo con él/ella, terminarás teniendo intereses similares y amando las mismas cosas. Eso es kNN con k=1.
- Si constantemente sales con un grupo de 5, cada uno del grupo tiene un impacto en tu comportamiento y terminarás convirtiéndose en el promedio de 5. Eso es kNN con k=5.

KNN

kNN identifica la clase de un punto de datos utilizando el principio de votación por mayoría.

Si k = 5, se examinan las clases de los 5 puntos más cercanos.

La predicción se realiza según la clase predominante.

De manera similar, la regresión kNN toma el valor medio de las 5 ubicaciones más cercanas.

KNN

¿Como validamos la cercanía?

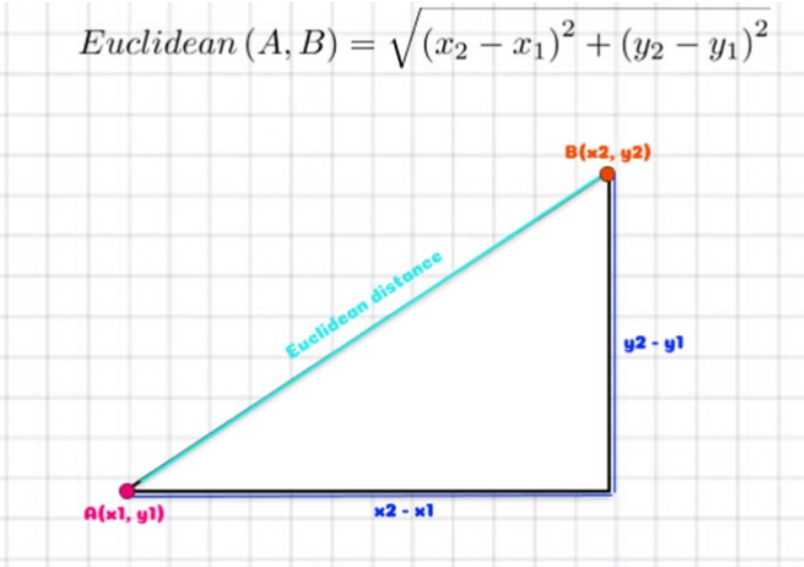
Distancia entre puntos de datos.

Existen varias técnicas para estimar la distancia.

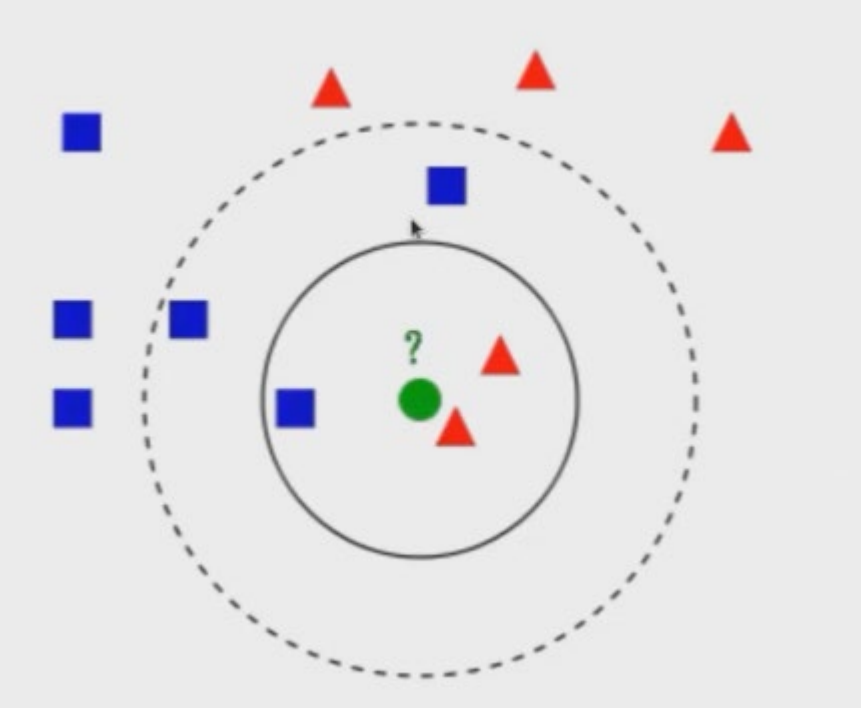
La distancia euclidiana (distancia de Minkowski con p=2) es una de las medidas de distancia más utilizadas.

*se puede usar cualquier otra distancia/función.

$$Euclidean\,(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



KNN



Ejemplo

Azules: Perros

Rojos: Lobo

Verde?

En función de las características mas similares con sus vecinos, clasificaríamos a verdi.

Si k=3

Verdi es? (línea continua) -> Az=1 RJ=2 ∴ Verdi es Lobo

Si k=5

Verdi es? (línea punteada) -> Az=3 RJ=2 ∴ Verdi es Perro

El mejor k para kNN

- **K bajo (por ejemplo, k=1)**: Sobreajuste: se adapta demasiado a los datos de entrenamiento y no generaliza bien para nuevos datos.
- **K alto (por ejemplo, k=100)**: Desajuste: el modelo es poco fiable tanto en los datos de entrenamiento como en los nuevos.

Y Entonces?

El mejor k para kNN

Y Entonces?

K-Fold Cross-Validation: Probar varios y elegir el mejor.

Una muy buena opción para encontrar el mejor k, es usar la herramienta grid search de sklearn.

El mejor k para kNN

Consideraciones Adicionales:

1. **Prueba con una Gama de Valores**: Comienza probando una amplia gama de valores de "k" (por ejemplo, desde 1 hasta raíz de n, donde n es el número de muestras en el conjunto de datos).
2. **Ajuste según el Tamaño del Conjunto de Datos**: Elige "k" más pequeños para conjuntos de datos más ruidosos o más pequeños, y "k" más grandes para conjuntos de datos más grandes o menos ruidosos.
3. **Balance entre Sesgo y Varianza**: Encuentra un equilibrio entre el sesgo (error debido a suposiciones incorrectas en el modelo) y la varianza (sensibilidad a pequeñas fluctuaciones en los datos) al seleccionar "k".

Limitaciones de kNN

KNN es simple y no depende de un modelo interno de aprendizaje automático. Trabaja con cualquier cantidad de categorías, lo que lo hace rápido para evaluar la adición de nuevas categorías. Sin embargo, su simplicidad impide anticipar eventos raros.

Aunque KNN puede lograr alta precisión en pruebas, es lento y costoso en tiempo y memoria. Requiere una gran cantidad de memoria para almacenar el conjunto de datos de entrenamiento, y la distancia euclidiana puede ser afectada por la magnitud de las características, dando mayor peso a características con magnitudes grandes.

- En resumen, KNN no es ideal para conjuntos de datos de grandes dimensiones.

Conclusión de kNN

Se suele usar mas para tareas especificas como agrupar puntos cercanos que de forma general.

El costo computacional depende de los datos, muchas datos, suele tener alto costo computacional

La interpretación suele ser intuitiva y ligada a la característica de la “distancia”.

Ventajas y Desventajas

Ventajas	Desventajas
Extremadamente fácil de implementar	No funciona bien con alta dimensionalidad
Puede realizar tareas complejas	Alto costo en datos grandes
No requiere entrenamiento	No funciona bien con variables categoricas
Nuevos datos, se agregan fácilmente	
Solo se necesita k y la distancia.	