Modelo no paramétrico: k Vecinos cercanos

Héctor Selley

Universidad Anáhuac México

13 de junio de 2023

Contenido

1 Introducción

2 Ejemplo

Regla del vecino cercano

Descripción

El método del k-ésimo vecino cercano utiliza el conjunto de entrenamiento y uno de prueba. Para cada renglón del conjunto de prueba, se encuentran los k vectores más cercanos (distancia Euclidiana) del conjunto de entrenamiento y la clasificación se decide por mayoría de votos, en el caso de los empates se decide aleatoriamente. Si hubiera empates para el k-ésimo vector más cercano, todos los candidatos se consideran en la votación.

Regla del vecino cercano

- Nearest neighbour

[1] Pieza de piedra coloreada para confeccionar un mosaico.

Regla del vecino cercano

- Nearest neighbour
- Identifica la categoría, con base en la de su vecino más cercano de acuerdo con alguna medida de distancia.

[1] Pieza de piedra coloreada para confeccionar un mosaico.

Regla del vecino cercano

- Nearest neighbour
- Identifica la categoría, con base en la de su vecino más cercano de acuerdo con alguna medida de distancia.
- Divide el espacio de características de forma no lineal.

[1] Pieza de piedra coloreada para confeccionar un mosaico.

Regla del vecino cercano

- Nearest neighbour
- Identifica la categoría, con base en la de su vecino más cercano de acuerdo con alguna medida de distancia.
- Divide el espacio de características de forma no lineal.
- Tesela[1] o diagrama de Voronoi.

[1] Pieza de piedra coloreada para confeccionar un mosaico.

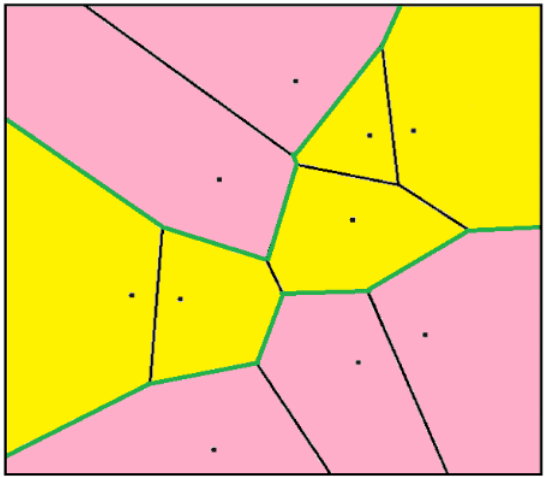
Regla del vecino cercano

- Nearest neighbour
- Identifica la categoría, con base en la de su vecino más cercano de acuerdo con alguna medida de distancia.
- Divide el espacio de características de forma no lineal.
- Tesela[1] o diagrama de Voronoi.
- Celdas formadas por todos los puntos, que se encuentran más cerca de un punto dado del conjunto de datos.

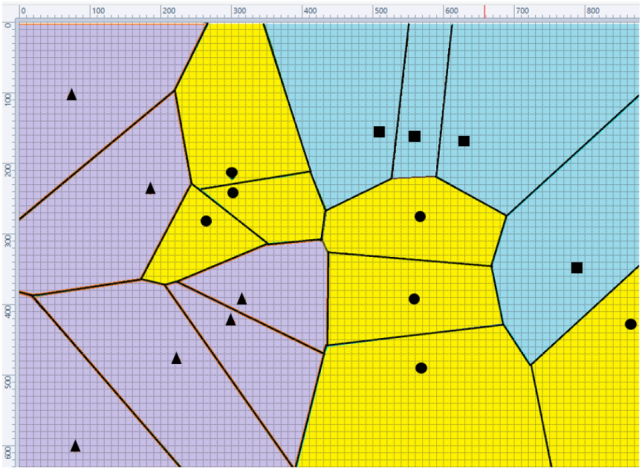
[1] Pieza de piedra coloreada para confeccionar un mosaico.

Regla del vecino cercano

Diagrama de Voroni



Regla del vecino cercano



K-Nearest Neighbour

- Consiste en encontrar en un conjunto de datos etiquetados, los k vecinos más cercanos (k-NN) y asignar el nuevo patrón a la clase mayoritaria.

K-Nearest Neighbour

- Consiste en encontrar en un conjunto de datos etiquetados, los k vecinos más cercanos (k-NN) y asignar el nuevo patrón a la clase mayoritaria.
  - A la de mayor probabilidad aposteriori.

K-Nearest Neighbour

- Consiste en encontrar en un conjunto de datos etiquetados, los k vecinos más cercanos (k-NN) y asignar el nuevo patrón a la clase mayoritaria.
  - A la de mayor probabilidad aposteriori.
- k es un número positivo entero

K-Nearest Neighbour

- Consiste en encontrar en un conjunto de datos etiquetados, los k vecinos más cercanos (k-NN) y asignar el nuevo patrón a la clase mayoritaria.
  - A la de mayor probabilidad aposteriori.
- k es un número positivo entero
  - En problemas binarios (de dos clases), es provechoso elegir un número impar.

K-Nearest Neighbour

Algoritmo k-NN

K-Nearest Neighbour

Algoritmo k-NN

1. Elegir el valor de k y la distancia a ocupar

K-Nearest Neighbour

Algoritmo k-NN

1. Elegir el valor de k y la distancia a ocupar
2. Obtener la distancia del objeto a clasificar a cada elemento del conjunto de datos

K-Nearest Neighbour

Algoritmo k-NN

1. Elegir el valor de k y la distancia a ocupar
2. Obtener la distancia del objeto a clasificar a cada elemento del conjunto de datos
3. Tomar k vecinos más cercanos y contar el número de elementos que pertenecen a cada categoría

K-Nearest Neighbour

Algoritmo k-NN

1. Elegir el valor de k y la distancia a ocupar
2. Obtener la distancia del objeto a clasificar a cada elemento del conjunto de datos
3. Tomar k vecinos más cercanos y contar el número de elementos que pertenecen a cada categoría
4. Asignar la categoría a la que pertenecen más vecinos

Contenido

1 Introducción

2 Ejemplo

Ejemplo: Clasificación de color

Problema:

- Se tienen las coordenadas (a, b) del modelo CIELAB[2,3] de color para pixeles rojos y naranjas.

[2] Sistema de interpretación de color CIE 1976 L*a*b*

[3] `https://es.wikipedia.org/wiki/Espacio_de_color_Lab`

Ejemplo: Clasificación de color

Problema:

- Se tienen las coordenadas (a, b) del modelo CIELAB[2,3] de color para pixeles rojos y naranjas.
- Llega una nueva observación ω con las coordenadas $x = (172, 160)$.

[2] Sistema de interpretación de color CIE 1976 L*a*b*

[3] `https://es.wikipedia.org/wiki/Espacio_de_color_Lab`

Ejemplo: Clasificación de color

Problema:

- Se tienen las coordenadas (a, b) del modelo CIELAB[2,3] de color para pixeles rojos y naranjas.
- Llega una nueva observación ω con las coordenadas $x = (172, 160)$.
- ¿Cómo se clasifica ω eligiendo de 1 hasta 8 vecinos cercanos usando distancia euclidiana?

[2] Sistema de interpretación de color CIE 1976 L*a*b*

[3] `https://es.wikipedia.org/wiki/Espacio_de_color_Lab`

Ejemplo: Clasificación de color

Formulación:

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles
- Clases

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles
- Clases
  - Ω_1: naranja

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles
- Clases
  - Ω_1: naranja
  - Ω_2: rojo

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles
- Clases
  - Ω_1: naranja
  - Ω_2: rojo
- Vector de características: $X = [a, b]$ donde $X : \Omega \to \mathbb{R}^2$

Ejemplo: Clasificación de color

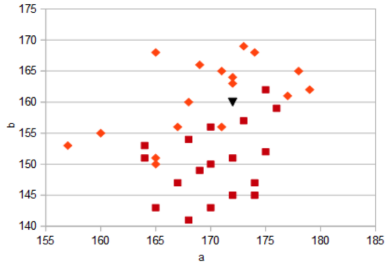
Formulación:

- Población Ω: pixeles
- Clases
  - Ω_1: naranja
  - Ω_2: rojo
- Vector de características: $X = [a, b]$ donde $X : \Omega \to \mathbb{R}^2$
- Función de distribución de probabilidad: no se asume

Ejemplo: Clasificación de color

Formulación:

- Población Ω: pixeles
- Clases
  - Ω_1: naranja
  - Ω_2: rojo
- Vector de características: $X = [a, b]$ donde $X : \Omega \to \mathbb{R}^2$
- Función de distribución de probabilidad: no se asume



Ejemplo: Clasificación de color

- Se utilizará un archivo de datos: *datosAB.txt*
- Conjunto de datos de clasificación de pixeles en color de acuerdo a su valor (a, b)

¿Clasificación?

clase	distancia	k–vecinos	votos rojo	votos naranja
naranja	3.00	1		1
rojo	3.16	2	1	1
rojo	3.61	3	2	1
naranja	4.00	4	2	2
naranja	4.00	5	2	3
rojo	4.12	6	3	3
naranja	4.12	7	3	4
rojo	4.47	8	4	4

Matriz de confusión

- Comparar modelos de clasificación

Matriz de confusión

- Comparar modelos de clasificación
- Utilizar CP para evaluar el rendimiento del clasificador

Matriz de confusión

- Comparar modelos de clasificación
- Utilizar CP para evaluar el rendimiento del clasificador

		clase estimada			
		$\bar{\Omega} = \Omega_0$	$\bar{\Omega} = \Omega_1$	...	$\bar{\Omega} = \Omega_j$
clase real	Ω_0	n_{00}	n_{01}	...	n_{0j}
	Ω_1	n_{10}	n_{11}	...	n_{1j}
	⋮	⋮	⋮	⋮	⋮
	Ω_i	n_{i0}	n_{i1}	...	n_{ij}

Matriz de confusión

- Proporción de clasificación correcta

$$cc = \frac{1}{N} \sum_{i=0}^{g-1} n_{ii}$$

Matriz de confusión

- Proporción de clasificación correcta

$$cc = \frac{1}{N} \sum_{i=0}^{g-1} n_{ii}$$

  - N: total de datos en CP

Matriz de confusión

- Proporción de clasificación correcta

$$cc = \frac{1}{N} \sum_{i=0}^{g-1} n_{ii}$$

  - N: total de datos en CP
  - Probabilidad de éxito

Matriz de confusión

- Proporción de clasificación correcta

$$cc = \frac{1}{N} \sum_{i=0}^{g-1} n_{ii}$$

  - N: total de datos en CP
  - Probabilidad de éxito
  - En un problema de 2 clases:

$$cc = \frac{TP + TN}{TP + FN + TN + FP}$$

Matriz de confusión

- Proporción de clasificación correcta

$$cc = \frac{1}{N} \sum_{i=0}^{g-1} n_{ii}$$

  - N: total de datos en CP
  - Probabilidad de éxito
  - En un problema de 2 clases:

$$cc = \frac{TP + TN}{TP + FN + TN + FP}$$

- Proporción estimada del error $= 1 - cc$

Matriz de confusión

		clase estimada		
		$\bar{\Omega} = \Omega_0$ **Clase no referencia**	$\bar{\Omega} = \Omega_1$ **Clase referencia**	**Total por renglón**
clase real	Ω_0 **Clase no referencia**	Verdadero Negativo **TN**	Falso Positivo **FP**	Total sin la condición **N- = TN + FP**
	Ω_1 **Clase referencia**	Falso Negativo **FN**	Verdadero Positivo **TP**	Total con la condición **N+ = TP + FN**
	Total por columna	**$\bar{N}$- = TN+FN** Total de pruebas negativas	**$\bar{N}$+ = FP+TP** Total de pruebas positivas	**N=TP+TN+FP+FN** Total de individuos en el CP