Algoritmo del vecino más cercano

DRA. CONSUELO VARINIA GARCÍA MENDOZA

Principios del algoritmo

Dos objetos similares suelen compartir características

Por ejemplo:

- Dos plantas que son parecidas suelen ser de la misma especie
- Dos pacientes que presentan los mismos síntomas suelen tener la misma enfermedad

Los algoritmos de aprendizaje automático pueden aprovechar esta similitud para determinar la clase a la que pertenece un objeto

Similitud de vectores de atributos

En el aprendizaje automático, las características de los objetos se representan como vectores

Estos vectores se pueden comparar para determinar su similitud

Para esta comparación se considera que entre menos diferencias existan entre las características mayor será la similitud

Ejemplo

Example	Shape	Crust		Filling		Class	# differences
		Size	Shade	Size	Shade		
x	Square	Thick	Gray	Thin	White	?	–
ex_1	Circle	Thick	Gray	Thick	Dark	pos	3
ex_2	Circle	Thick	White	Thick	Dark	pos	4
ex_3	Triangle	Thick	Dark	Thick	Gray	pos	4
ex_4	Circle	Thin	White	Thin	Dark	pos	4
ex_5	Square	Thick	Dark	Thin	White	pos	1
ex_6	Circle	Thick	White	Thin	Dark	pos	3
ex_7	Circle	Thick	Gray	Thick	White	neg	2
ex_8	Square	Thick	White	Thick	Gray	neg	3
ex_9	Triangle	Thin	Gray	Thin	Dark	neg	3
ex_{10}	Circle	Thick	Dark	Thick	White	neg	3
ex_{11}	Square	Thick	White	Thick	Dark	neg	3
ex_{12}	Triangle	Thick	White	Thick	Gray	neg	4

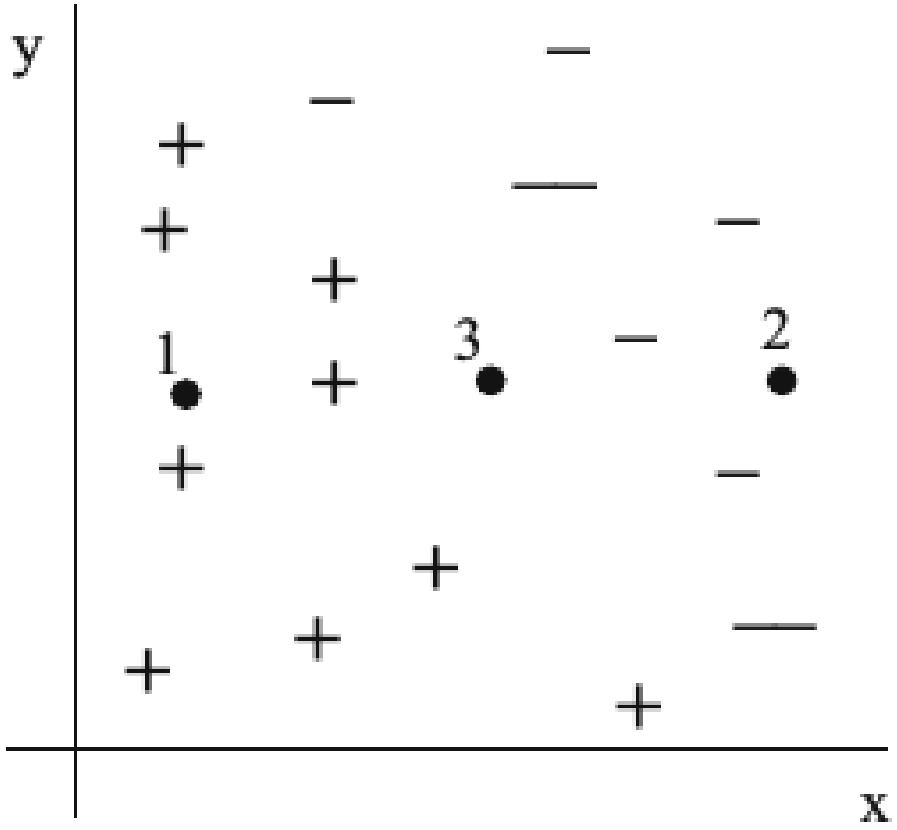
El vecino más cercano

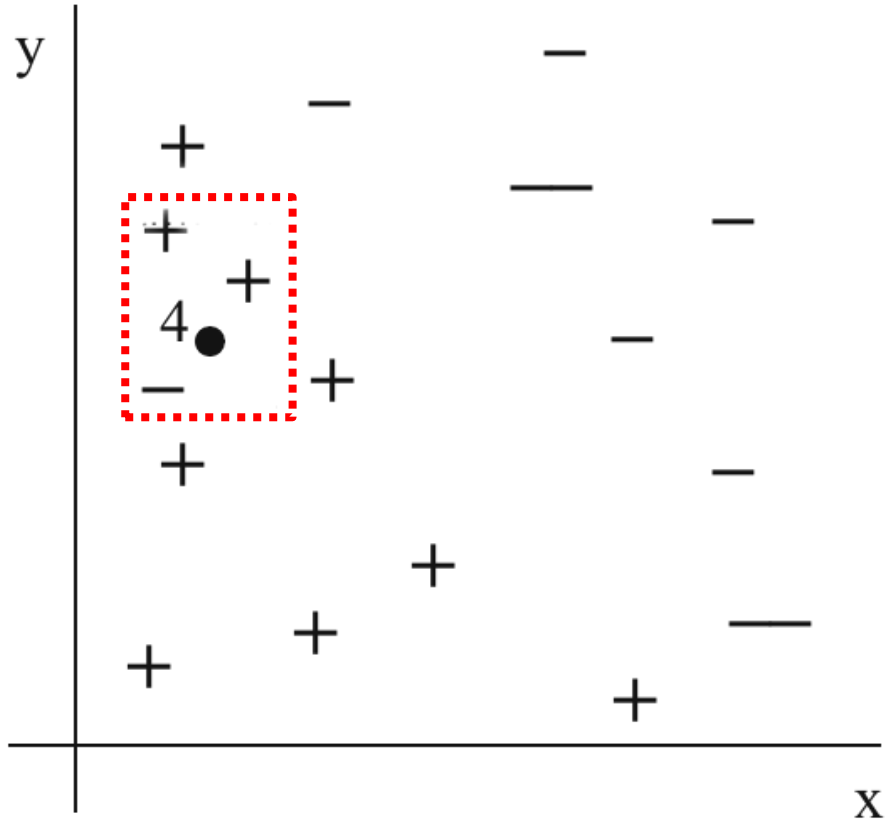
Cuando los objetos se representan como un punto en el espacio n-dimensional se puede calcular la distancia geométrica entre cualquier par de instancias, por ejemplo, mediante la distancia Euclidiana

Para estos casos también se puede establecer que entre más cercanas estén unas instancias de otras en el espacio, mayor será la similitud mutua

De este principio es de donde el clasificador del *vecino más cercano* toma su nombre: el ejemplo de entrenamiento con la menor distancia de x en el espacio de instancias es, desde el punto de vista geométrico, el vecino más cercano de x

Ejemplos





Los k vecinos más cercanos

En dominios más ruidosos, la cercanía a un solo vecino puede no ser tan confiable, dado que la etiqueta de clase de dicho vecino puede ser incorrecta

En un enfoque más robusto se identificarían no uno si no varios vecinos cercanos y se utiliza el voto de la mayoría para determinar la clase de la instancia a predecir

Esta idea es la base del algoritmo llamado Clasificador k vecinos más cercanos (k-NN, k Nearest Neighbors), donde k es el número de vecinos que votan

Algoritmo simple de los k vecinos más cercanos

Suppose we have a mechanism to evaluate the similarly between attribute vectors. Let **x** denote the object whose class we want to determine.

1. Among the training examples, identify the k nearest neighbors of **x** (examples most similar to **x**).
2. Let c_i be the class most frequently found among these k nearest neighbors.
3. Label **x** with c_i.

Métrica de similitud

Una forma de encontrar los vecinos más cercanos al objeto *x* es comparar las distancias geométricas individuales de cada ejemplo de entrenamiento con respecto a *x*

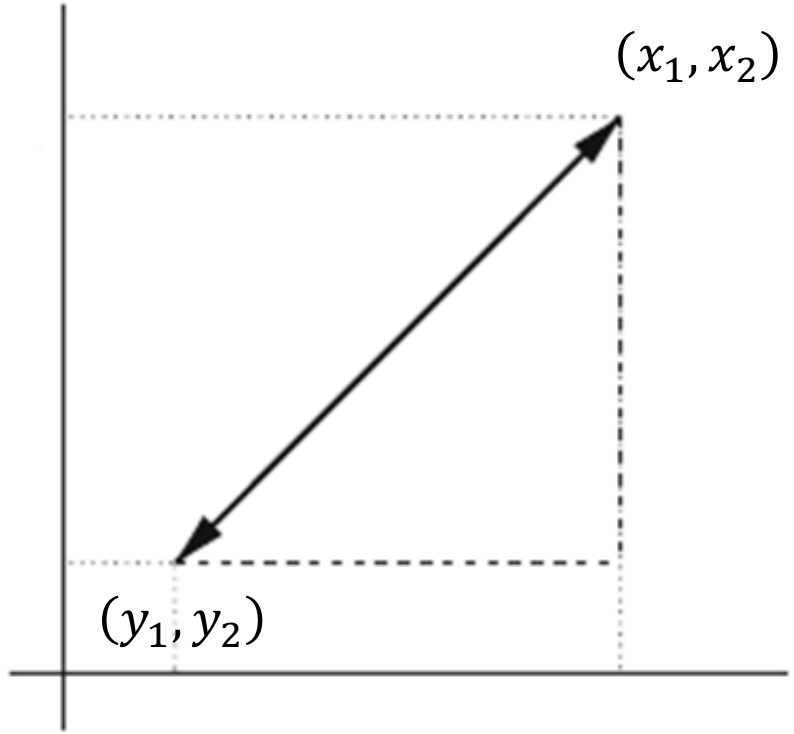
Como ya se había mencionado, la distancia Euclidiana es adecuada para usarla como métrica de similitud

La distancia Euclidiana entre dos puntos en un espacio de dos dimensiones es igual a la longitud de la hipotenusa del triángulo y se calcula mediante la siguiente fórmula

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Esta fórmula se puede generalizar para usarla con n atributos mediante la siguiente fórmula

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2}$$



Ejemplo

Dada la siguiente instancia con tres atributos numéricos x = [2, 4, 2], determine su clase usando 1 (1-NN) y 3 (3-NN) vecinos más cercanos

Distance between ex_i and $[2, 4, 2]$		
ex_1	$\{[1, 3, 1], \text{pos}\}$	$\sqrt{(2-1)^2 + (4-3)^2 + (2-1)^2} = \sqrt{3}$
ex_2	$\{[3, 5, 2], \text{pos}\}$	$\sqrt{(2-3)^2 + (4-5)^2 + (2-2)^2} = \sqrt{2}$
ex_3	$\{[3, 2, 2], \text{neg}\}$	$\sqrt{(2-3)^2 + (4-2)^2 + (2-2)^2} = \sqrt{5}$
ex_4	$\{[5, 2, 3], \text{neg}\}$	$\sqrt{(2-5)^2 + (4-2)^2 + (2-3)^2} = \sqrt{4}$

Atributos con valores continuos y con valores discretos

Los valores de los atributos pueden ser continuos o categóricos

Esto se puede generalizar mediante la siguiente fórmula

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{\Sigma_{i=1}^{n} d(x_i, y_i)}$$

El subíndice M en la fórmula indica que los atributos continuos y discretos pueden estar mezclados, pero la aplicación de esta fórmula debe considerar los siguientes dos casos:

- Para valores continuos

$$d(x_i, y_i) = (x_i - y_i)^2$$

- Para valores categóricos

$$d(x_i, y_i) = 0 \text{ if } x_i = y_i$$

$$d(x_i, y_i) = 1 \text{ if } x_i \neq y_i$$

Otras medidas de similitud

El algoritmo de k vecinos más cercanos puede utilizar distintas métricas de similitud dependiendo del objetivo que se busque

Algunas de las métricas más utilizadas son

- Minkowski
- Manhattan
- Mahalanobis
- Coseno

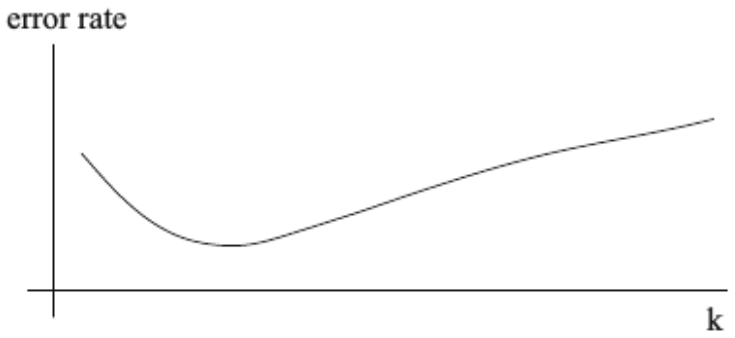
Aspectos a considerar

Escala de los valores de los atributos

$$\mathbf{x} = (t, 0.2, 254) \qquad \mathbf{y} = (f, 0.1, 194)$$

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(1-0)^2 + (0.2-0.1)^2 + (254-194)^2}$$

Incremento del número de vecinos



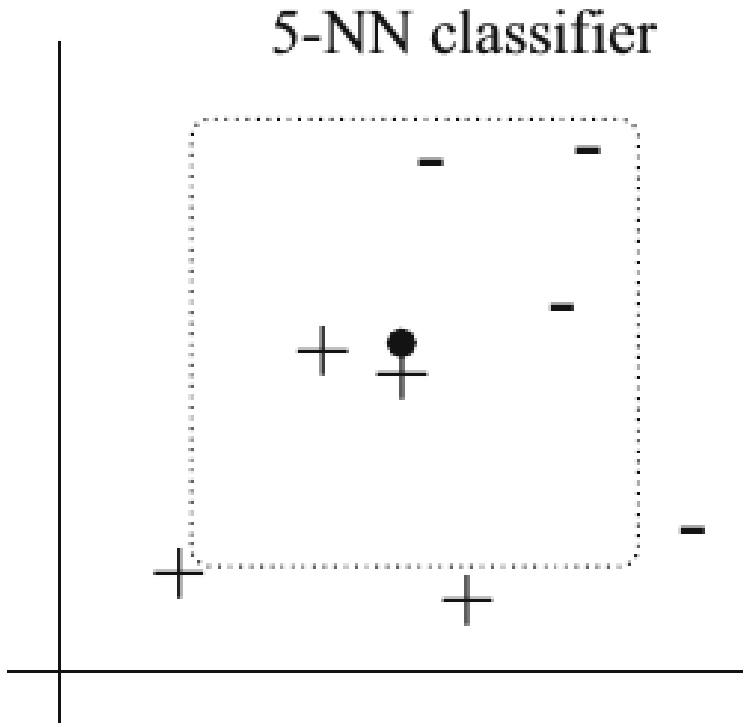
Vecinos más cercanos con peso

El algoritmo revisado considera que los vecinos más cercanos "votan" de igual manera por la clase que se debe asignar a la instancia a predecir

Sin embargo, en ocasiones esto puede no ser lo ideal para algunos casos

En este ejemplo se puede apreciar que la clase seleccionada sería la negativa, a pesar de que estas instancias están más alejadas del ejemplo que las dos instancias positivas

Una forma de solucionar esto es hacer que el peso de los vecinos sea proporcional a su distancia con la instancia a clasificar



Ejemplo

Dadas las distancias entre el ejemplo x y sus cinco vecinos más cercanos $d_1 = 1, d_2 = 3, d_3 = 4, d_4 = 5, d_5 = 8$ calcula el peso de cada vecino con la siguiente formula

$$w_i = \begin{cases} \dfrac{d_{max} - d_i}{d_{max} - d_{min}} & , \quad d_{max} \neq d_{min} \\ 1 & , \quad d_{max} = d_{min} \end{cases}$$

$w_1 = \frac{8-d_i}{7} = \frac{8-1}{7} = 1$

$w_2 = \frac{8-d_i}{7} = \frac{8-3}{7} = \frac{5}{7}$

$w_3 = \frac{8-d_i}{7} = \frac{8-4}{7} = \frac{4}{7}$

$w_4 = \frac{8-d_i}{7} = \frac{8-5}{7} = \frac{3}{7}$

$w_5 = \frac{8-d_i}{7} = \frac{8-8}{7} = 0$

Ejemplo

Si las primeras dos instancias pertenecen a la clase + y las últimas 3 a la clase negativa determina a que clase pertenece el ejemplo x

$$w_1 = 1, w_2 = \frac{5}{7}, w_3 = \frac{4}{7}, w_4 = \frac{3}{7}, w_5 = 0$$

$$\Sigma^+ = 1 + \frac{5}{7}$$

$$\Sigma^- = \frac{4}{7} + \frac{3}{7} + 0 = 1$$

$$\Sigma^+ > \Sigma^- \quad \therefore x \in +$$