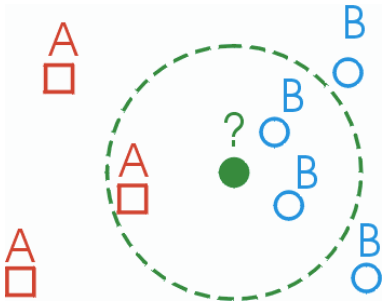
Aprendizaje con vecinos cercanos

Karim Pichara
Associate Professor
Computer Science Department
Pontificia Universidad Católica de Chile

Aprendizaje en base a casos

- Idea: Clasificar o predecir datos nuevos en base a los casos similares que ya conocemos.

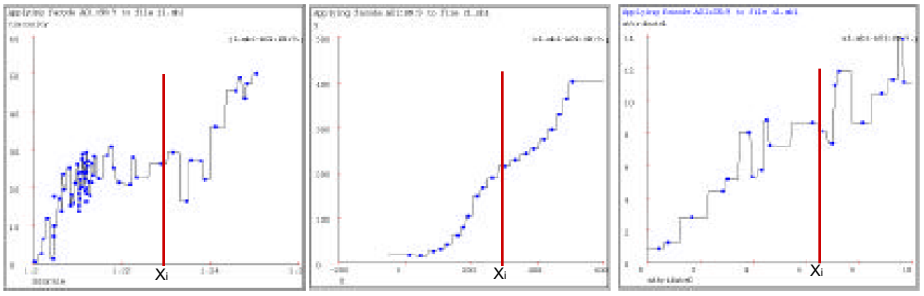


1- Vecino Cercano

- Clasificación o Predicción: Se utiliza la clase o el valor del que más se "parece".
- Medida de similitud = Algún tipo de distancia, ej. distancia Euclideana
- Sensible a errores o ruido en los datos

Karim Pichara B.

Ej: 1- Vecino Cercano (Regresión)



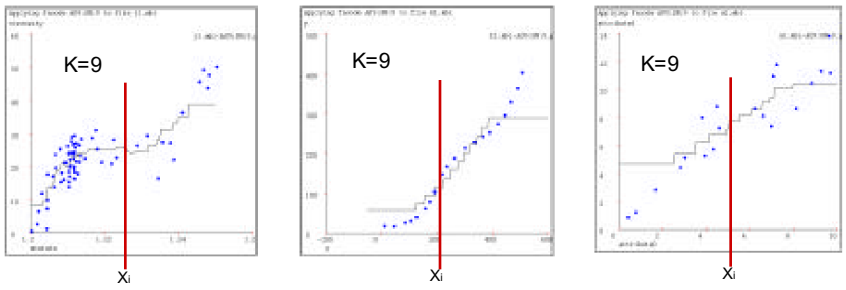
Karim Pichara B.

K- Vecinos Cercanos

- Clasificación: Cada vez que se quiera clasificar un ejemplo nuevo se utiliza la clase de la mayoría de los K que más se "parecen".
- Predicción: Se utiliza el valor del promedio de los k elementos más cercanos.

Karim Pichara B.

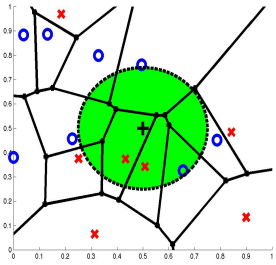
Ej: K- Vecinos Cercanos (Regresión)



Karim Pichara B.

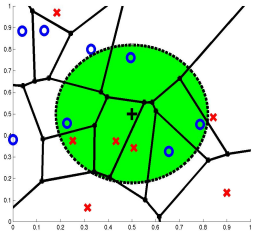
K- Vecinos Cercanos

Clasificación



3-NN

Diagramas de Voronoi



7-NN

Los polígonos representan la región más cercana para cada instancia, es decir,
Si llega una instancia nueva y queremos clasificarla usando 1-NN. Bastaría ver en qué
Polígono cae y clasificarla como la instancia que estaba dentro del polígono

Ventajas y Desventajas

- Ventajas
  - No se gasta tiempo en entrenamiento previo.
  - Permite aprender funciones complejas
- Desventajas
  - Lento al momento de consultas.
  - Problemas con atributos irrelevantes.
  - Problemas en grandes dimensiones.

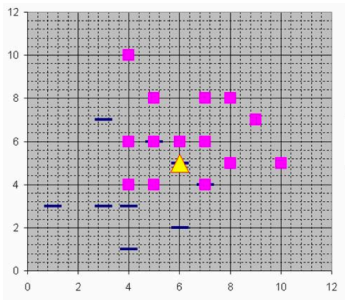
Karim Pichara B.

Ejemplo

X1	X2	Y
4	3	+
1	3	+
3	3	+
3	7	+
7	4	+
4	1	+
6	5	+
5	6	+
3	7	+
6	2	+
4	6	-
4	4	-
5	8	-
7	8	-
5	6	-
10	5	-
7	6	-
4	10	-
9	7	-
5	4	-
8	5	-
6	6	-
7	4	-
8	8	-
6	5	?

training data

Classification

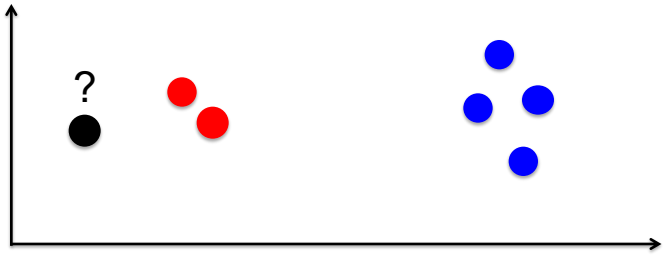


\+
\-
?

Computation	
Distance	Nearest Neighbor sign
8	
29	
13	
13	
2	+
20	
0	+
2	+
13	
9	
5	
5	
10	
10	
2	-
16	
2	-
29	
13	
2	-
4	
1	-
2	-
13	

Variaciones

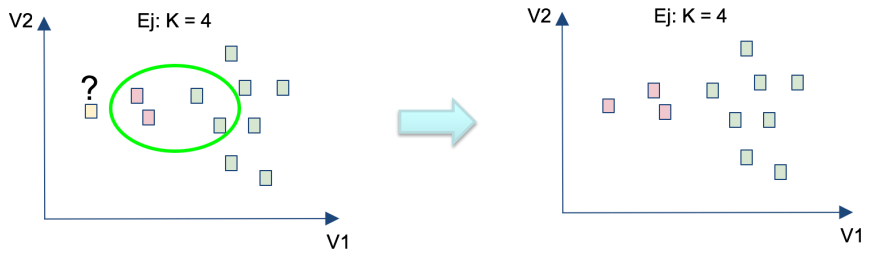
- Se pueden utilizar distancias pesadas, para das más énfasis a los elementos más cercanos y menos importancia a los vecinos no tan cercanos.



Karim Pichara B.

Variaciones

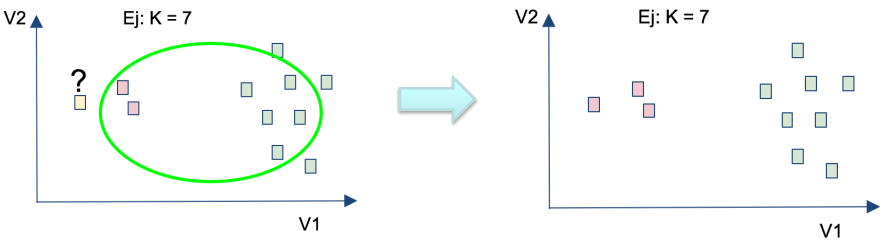
- Se pueden utilizar distancias pesadas, para das más énfasis a los elementos más cercanos y menos importancia a los vecinos no tan cercanos.



Karim Pichara B.

Variaciones

- Se pueden utilizar distancias pesadas, para das más énfasis a los elementos más cercanos y menos importancia a los vecinos no tan cercanos.



Karim Pichara B.

Distance Weighted k-NN

- Pesos más comunes para ponderar a los vecinos:

- w = $\frac{1}{d}$
- w = $\frac{1}{\sqrt{2\pi}} e^{\frac{-d^2}{2}}$

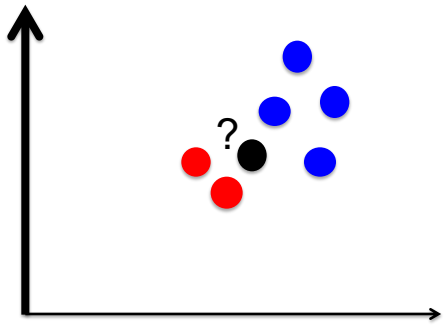
$$Clase(x_q) = \sum_{v=1}^{K} \hat{w}_v * Clase(x_v)$$

$$\hat{w}_v = \frac{w_v}{\sum_{i=1}^{k} w_i}$$

Karim Pichara B.

Variaciones

- Se pueden utilizar pesos para las dimensiones, de tal forma de dar pesos menores a las dimensiones menos relevantes.



Karim Pichara B.

Variaciones

- Se pueden utilizar pesos para las dimensiones, de tal forma de dar pesos menores a las dimensiones menos relevantes.
- Ej: Pesos en la distancia Euclidiana

2 dimensiones

$$d(P,Q) = \sqrt{w_1 * (P_1 - Q_1)^2 + w_2 * (P_2 - Q_2)^2}$$

D dimensiones

$$d(P,Q) = \sqrt{\sum_{i=1}^{D} w_i * (P_i - Q_i)^2}$$

Karim Pichara B.

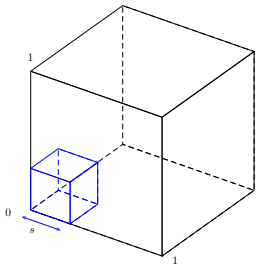
Decisiones relevantes

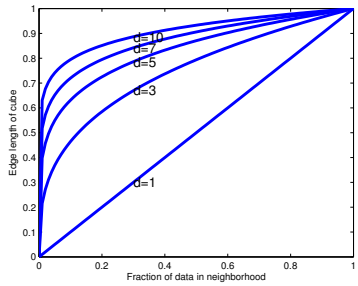
- Cuantos vecinos considerar (valor de K)
- Una métrica de distancia.
- Posibles variantes (pesos)
- Tiempos de respuesta a consultas

Karim Pichara B.

KNN and the curse of dimensionality

- Si aplicamos KNN a datos uniformes en un cubo de dimensión D
- Para alcanzar el "f %" de los datos, cada arista del cubo debe alcanzar $e_D(f) = f^{1/D}$ datos en cada dimensión
- Ej: D=10, queremos alcanzar el 10% de los datos. En cada dimensión, la arista debe alcanzar $e_{10}(0.1) = 0.8$





Karim Pichara B.

Resumen

•K-NN puede modelar funciones complejas.

•Problemas con la métrica de distancia cuando se tienen muchos atributos y sólo algunos influyen en la clasificación

•1-NN es más sensible al ruido que K-NN

Karim Pichara B.

Resumen

- Éstos predictores no ocupan tiempo de entrenamiento (Lazy Learner)
- El tiempo se utiliza cada vez que llega un nuevo elemento
- K-NN es un modelo no paramétrico

Karim Pichara B.