Modelamiento paramétrico

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

Paramétricos vs no paramétricos

- Son representados por una formula matemática:
  - Regresión logística
  - **Clasificador bayesiano**
  - Redes neuronales
  - GMM

- Son representados por la data
  - K-medias
  - **K-nn**
  - Arboles de decisión

Clasificador bayesiano

Naive bayes

Naive bayes aprende una distribución condicional de la probabilidad. Dado un punto de datos x, la salida del modelo es la probabilidad que x pertenezca a una clase específica.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- ¿Qué tan probable es comprar una computadora a un estudiante de 23 años con un crédito justo y ingresos medios? **x**={<=30,medium,yes,fair}

- **P(BC=yes|A<=30; I=med; S=yes; CR=fair)∝0.028**
- **P(BC=no|A<=30; I=med; S=yes; CR=fair)∝0.007**

Características

Naive bayes utiliza 3 aspectos clave: probabilidad condicional, teorema bayesiano e independencia condicional.

Probabilidad condicional:

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}, C)}{P(\mathbf{X})} \qquad P(\mathbf{X}|C) = \frac{P(\mathbf{X}, C)}{P(C)}$$

Bayesian theorem:

$$\underset{\text{Posterior}}{P(C|\mathbf{X})} = \frac{\overset{\text{Likelihood}}{P(\mathbf{X}|C)}\overset{\text{Prior}}{P(C)}}{P(\mathbf{X})_{\text{Marginalization}}}$$

Independencia condicional:

Si X1, X2, ..., Xk son independientes dado C entonces

$$P(X_1, X_2, \ldots, X_k|C) = \prod_{i=1}^{k} P(X_i|C)$$

Teorema bayesiano

Dado que:

Meningitis produce torticolis el 50% de las veces => P(T|M)=0.5

La probabilidad de meningitis es 1/50,000 =>P(M)=1/50000

La probabilidad de torticolis es 1/20 => P(T)=1/20

Si un paciente tiene torticolis, ¿cuál es la probabilidad de meningitis?

$$P(M|T) = \frac{P(T|M)P(M)}{P(T)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Funcionamiento

Asuma cada atributo y clase como variables

Dado un conjunto de datos con atributos (X1, X2,...,Xk)

El objetivo es predecir la clase C

En concreto, queremos encontrar el modelo que maximice P(C|X1, X2,...,Xk) dados todos los puntos del conjunto de datos (entrenamiento)

Sin embargo, ¿podemos estimar P(C|X1, X2,...,Xk) a partir de los datos?

$$P(C|X_1, X_2, \ldots, X_k) = \frac{P(X_1, X_2, \ldots, X_k|C)P(C)}{P(X_1, X_2, \ldots, X_k)}$$

Bayes ingenuo

$$P(C|X_1, X_2, \ldots, X_k) \propto P(X_1, X_2, \ldots, X_k|C)P(C)$$

¿Cómo podemos calcular P(X1, X2,…,Xk| C)? Es complicado y necesario.

Truco: **ASUMIR INGENUAMENTE** independencia condicional de X1, X2,…,Xk dado C (aunque esto no es necesariamente cierto).

$$P(C|X_1, X_2, \ldots, X_k) \propto \prod_{i=1}^{k} P(X_i|C)P(C)$$

- Ahora tenemos que calcular cada P (Xi|C) a partir de los datos.
- Un nuevo punto se clasifica como Cj si P(Cj| X) es el máximo entre todos los P(Ci|X)

Aprendizaje

$$P(BC|A,I,S,CR) = \frac{P(A,I,S,CR|BC)P(BC)}{P(A,I,S,CR)}$$
$$= \frac{P(A|BC)P(I|BC)P(S|BC)P(CR|BC)P(BC)}{P(A,I,S,CR)}$$
$$\propto P(A|BC)P(I|BC)P(S|BC)P(CR|BC)P(BC)$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

NBC parameters = CPDs+prior

CPDs: P(A|BC)
P(I|BC)
P(S|BC)
P(CR|BC)
Prior: P(BC)

Ejemplo

age	income	student	credit	buys
<=30	high	no	fair	**no**
<=30	high	no	excellent	**no**
31…40	high	no	fair	**yes**
>40	medium	no	fair	**yes**
>40	low	yes	fair	**yes**
>40	low	yes	excellent	**no**
31…40	low	yes	excellent	**yes**
<=30	medium	no	fair	**no**
<=30	low	yes	fair	**yes**
>40	medium	yes	fair	**yes**
<=30	medium	yes	excellent	**yes**
31…40	medium	no	excellent	**yes**
31…40	high	yes	fair	**yes**
>40	medium	no	excellent	**no**

Estimación prior P(BC) y distribuciones de probabilidad condicional P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independientemente con la estimación de máxima probabilidad

P(BC)

BC	
YES	
NO	

Ejemplo

age	income	student	credit	buys	
<=30	high	no	fair	**no**	**1**
<=30	high	no	excellent	**no**	**2**
31…40	high	no	fair	**yes**	**3**
>40	medium	no	fair	**yes**	**4**
>40	low	yes	fair	**yes**	**5**
>40	low	yes	excellent	**no**	**6**
31…40	low	yes	excellent	**yes**	**7**
<=30	medium	no	fair	**no**	**8**
<=30	low	yes	fair	**yes**	**9**
>40	medium	yes	fair	**yes**	**10**
<=30	medium	yes	excellent	**yes**	**11**
31…40	medium	no	excellent	**yes**	**12**
31…40	high	yes	fair	**yes**	**13**
>40	medium	no	excellent	**no**	**14**

Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

n=14

P(BC)

BC	
YES	
NO	

Ejemplo

age	income	student	credit	buys	
<=30	high	no	fair	**no**	
<=30	high	no	excellent	**no**	
31…40	high	no	fair	**yes**	**1**
>40	medium	no	fair	**yes**	**2**
>40	low	yes	fair	**yes**	**3**
>40	low	yes	excellent	**no**	
31…40	low	yes	excellent	**yes**	**4**
<=30	medium	no	fair	**no**	
<=30	low	yes	fair	**yes**	**5**
>40	medium	yes	fair	**yes**	**6**
<=30	medium	yes	excellent	**yes**	**7**
31…40	medium	no	excellent	**yes**	**8**
31…40	high	yes	fair	**yes**	**9**
>40	medium	no	excellent	**no**	

Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

BC=yes => 9

P(BC)

BC	
YES	**9/14**
NO	

Ejemplo

age	income	student	credit	buys	
<=30	high	no	fair	no	1
<=30	high	no	excellent	no	2
31…40	high	no	fair	yes	
>40	medium	no	fair	yes	
>40	low	yes	fair	yes	
>40	low	yes	excellent	no	3
31…40	low	yes	excellent	yes	
<=30	medium	no	fair	no	4
<=30	low	yes	fair	yes	
>40	medium	yes	fair	yes	
<=30	medium	yes	excellent	yes	
31…40	medium	no	excellent	yes	
31…40	high	yes	fair	yes	
>40	medium	no	excellent	no	5

Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

BC=no => 5

P(BC)

BC	
YES	9/14
NO	5/14

Ejemplo

age	income	student	credit	buys
<=30	high	no	fair	**no**
<=30	high	no	excellent	**no**
31…40	high	no	fair	**yes**
>40	medium	no	fair	**yes**
>40	low	yes	fair	**yes**
>40	low	yes	excellent	**no**
31…40	low	yes	excellent	**yes**
<=30	medium	no	fair	**no**
<=30	low	yes	fair	**yes**
>40	medium	yes	fair	**yes**
<=30	medium	yes	excellent	**yes**
31…40	medium	no	excellent	**yes**
31…40	high	yes	fair	**yes**
>40	medium	no	excellent	**no**

Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

P(A|BC)

BC	A	
YES	<=30	**2/9**
	31..40	**4/9**
	>40	**3/9**
NO	<=30	
	31..40	
	>40	

Ejemplo

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

P(A | BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	4/9
	low	3/9
NO	high	2/5
	med	2/5
	low	1/5

P(S | BC)

BC	S	
YES	yes	6/9
	no	3/9
NO	yes	1/5
	no	4/5

P(CR | BC)

BC	CR	
YES	exc	3/9
	fair	6/9
NO	exc	4/5
	fair	2/5

P(BC)

BC	
YES	9/14
NO	5/14

Ejemplo

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

PREDICTION: How likely is to buy a computer a 23 years old student with a fair credit and medium income?

P(BC=yes|A<=30; I=med; S=yes; CR=fair)

$\propto$P(A<=30|BC=y)P(I=med|BC=y)
P(S=y|BC=y)P(CR=fair|BC=y)P(BC=y)
= 2/9 * 4/9 * 6/9 * 6/9 * 9/14 ≈ 0.028

P(A | BC)

BC	A	
YES	**<=30**	**2/9**
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	**4/9**
	low	3/9
NO	high	2/5
	med	2/5
	low	1/5

P(S | BC)

BC	S	
YES	**yes**	**6/9**
	no	3/9
NO	yes	1/5
	no	4/5

P(CR | BC)

BC	CR	
YES	exc	3/9
	fair	**6/9**
NO	exc	3/5
	fair	2/5

P(BC)

BC	
YES	**9/14**
NO	5/14

Ejemplo

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

PREDICTION: How likely is to NOT buy a computer a 23 years old student with a fair credit and medium income?

P(BC=no|A<=30; I=med; S=yes; CR=fair)

$\propto$P(A<=30|BC=n)P(I=med|BC=n)
P(S=y|BC=n)P(CR=fair|BC=n)P(BC=n)
= 3/5 * 2/5 * 1/5 * 2/5 * 5/14 ≈ 0.007

P(A | BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	**<=30**	**3/5**
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	4/9
	low	3/9
NO	high	2/5
	med	**2/5**
	low	1/5

P(S | BC)

BC	S	
YES	yes	6/9
	no	3/9
NO	**yes**	**1/5**
	no	4/5

P(CR | BC)

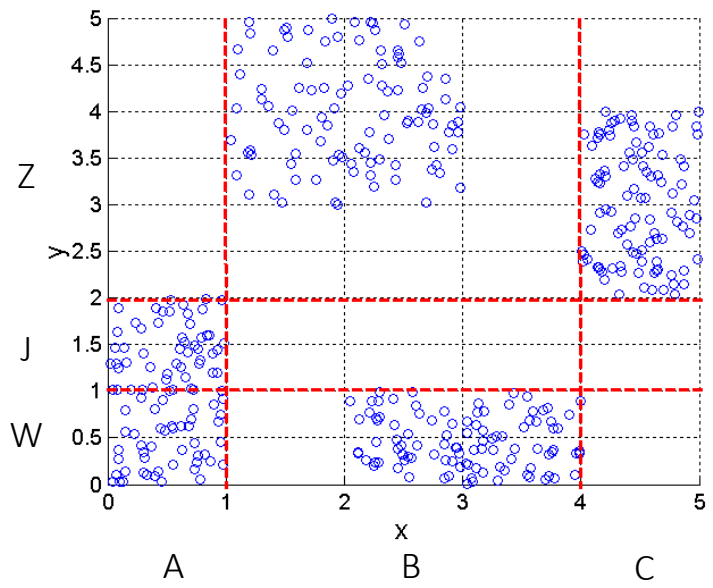
BC	CR	
YES	exc	3/9
	fair	6/9
NO	exc	3/5
	fair	**2/5**

P(BC)

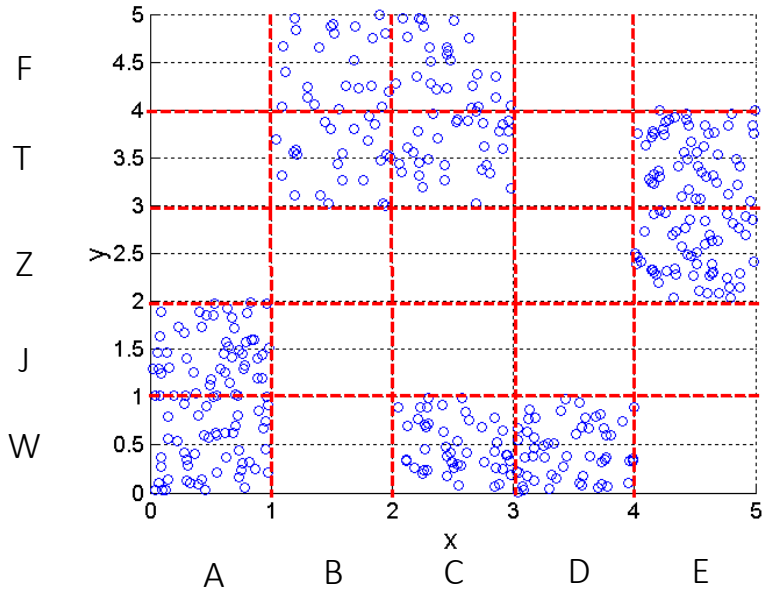
BC	
YES	9/14
NO	**5/14**

Discretización

La discretización es el proceso de poner valores en buckets para que haya un número limitado de estados posibles.
Discretización: cambie una variable continua a una variable categórica.



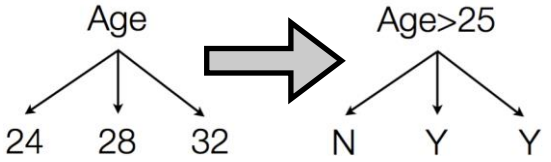
3 categories for both x and y (custom sized bin)



5 categories for both x and y (equal-sized bins)

**(equal Nº of entities per bin)

1D example:



Variables de salida continuas

refund	status	income	affair
yes	single	125K	**no**
no	married	100K	**no**
no	single	70K	**no**
yes	married	120K	**no**
no	Divorced	95K	**yes**
no	married	60K	**no**
yes	Divorced	220K	**no**
no	single	85K	**yes**
no	married	75K	**no**
no	single	90K	**yes**

Para calcular P(income|affair) tenemos dos opciones

1.

1. Discretizar los ingresos y tratado como un parámetro discreto

2. Estimar una serie de distribuciones normales iguales al número de clases

$$P(\text{income} = 120K|\text{affair} = no) = \frac{1}{\sqrt{2\pi}55}\exp\left(-\frac{(120-110)^2}{2*55^2}\right) \approx 0.0072$$

Resumen

Fortalezas:

- Fácil de implementar y se puede aprender de forma incremental
- A menudo funciona bien incluso cuando se viola la suposición
- Se puede aprender gradualmente
- Los valores que faltan se ignoran en el proceso de aprendizaje
- Modelo robusto con respecto a valores atípicos y datos irrelevantes

Debilidades:

- La suposición condicional de clase produce estimaciones de probabilidad sesgadas
- Las dependencias entre variables no se pueden modelar

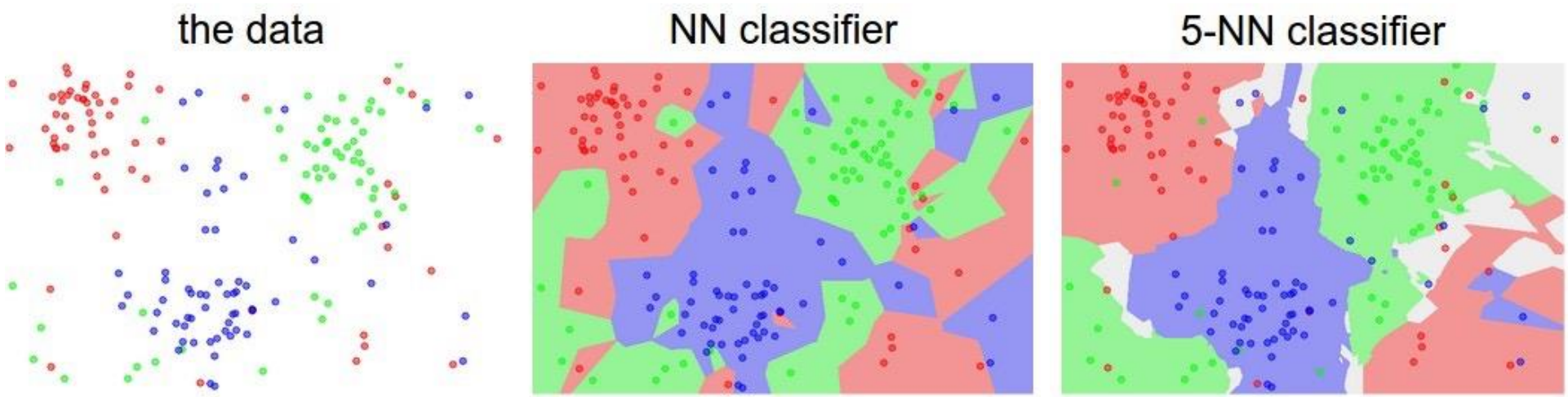
K vecino más cercano (k-NN)

k-NN

k-NN es un tipo de aprendizaje basado en instancias o aprendizaje perezoso (lazy).

k-NN almacena los datos de entrenamiento p-dimensionales y retrasa el proceso de aprendizaje hasta que se debe clasificar una nueva instancia.

Para predecir un nuevo punto, los vecinos k más cercanos se calculan utilizando la distancia euclidiana y la clasificación final se realiza en función de las etiquetas de clase de estos vecinos.



1-NN

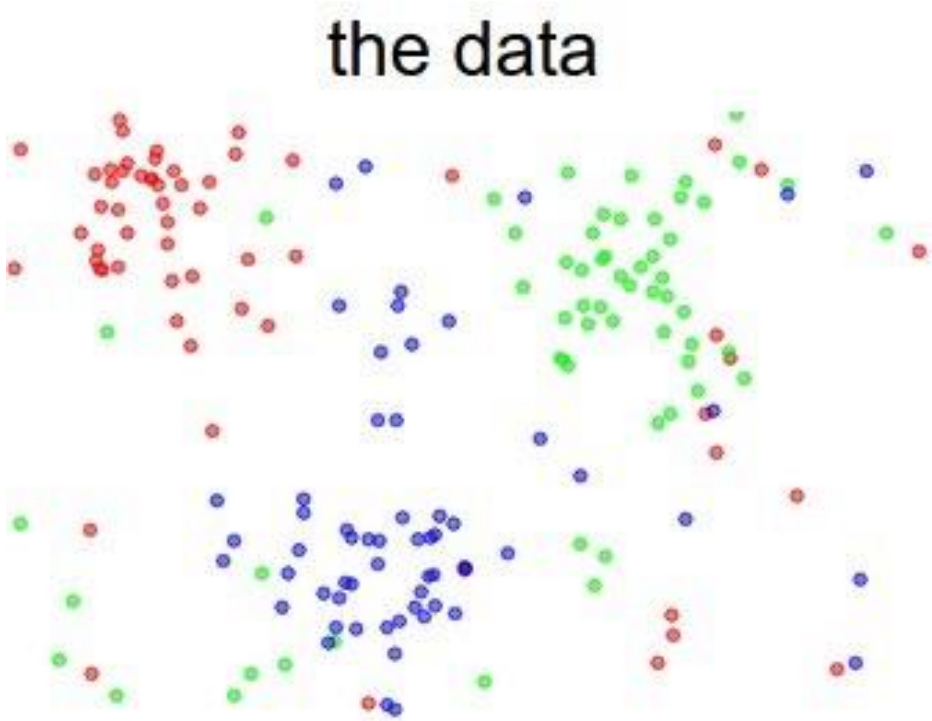
Idea clave: busque instancias "similares" a la nueva instancia y utilice sus etiquetas de clase para realizar predicciones para la nueva instancia

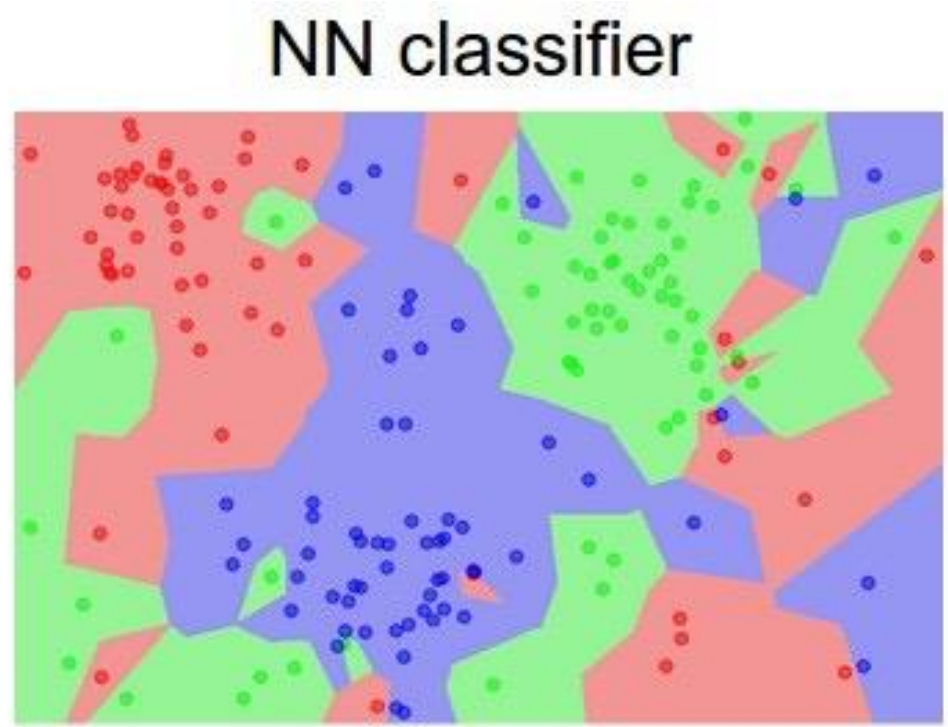
Algoritmo 1-NN:

- Deje que xj sea un nuevo punto p-dimensional sin etiquetar.
- Calcular d(xi, xj) para i={1,...,n}
- Encuentra el punto xi como d(xi, xj) se minimiza.
- Etiqueta xj con la misma etiqueta que xi

1-NN se generaliza a k-NN cuando se consideran más vecinos

Fronteras de decisión





Algoritmo k-NN:

- Deje que xj sea un nuevo punto pdimensional sin etiquetar.
- Calcular d(xi, xj) para i={1,...,n}
- Encuentra los vecinos k más cercanos de xj (d(xi, xj) se minimiza).
- Etiqueta xj basada en los vecinos más cercanos k

¿Cuál es la mejor relación calidad-precio para k? Por lo general, se utiliza un valor pequeño, por ejemplo, k<10.

¿Qué medida de distancia d( ) utilizar? A menudo se utiliza la distancia euclidiana (L2).

¿Cómo puedo etiquetar xj en función de los vecinos k más cercanos? A menudo se utiliza el voto mayoritario.

Características

Parámetros del modelo:

- k (número de vecinos)
- cualquier parámetro de medida de distancia (por ejemplo, pesos en las entidades)

Fortalezas:

- Modelo sencillo, fácil de implementar
- Aprendizaje muy eficiente: O(1)

Debilidades:

- Inferencia ineficiente: tiempo y espacio O(n)
- Maldición de la dimensionalidad: A medida que aumenta el número de entidades, necesita un aumento exponencial en el tamaño de los datos para asegurarse de que tiene ejemplos cercanos para cualquier dato dado

Ejemplo

El siguiente conjunto de datos corresponde a 10 puntos etiquetados como + y -. Clasifique cada punto en función de sus otros datos aplicando K-NN con K igual a 1 y 3 (utilice la distancia euclidiana).
Para cada valor de K calcule el error de clasificación (número de puntos clasificados erróneamente). Por último, teniendo en cuenta todos los puntos de datos, dibuje el límite de decisión utilizando 1-NN.

data	class
3.2	-
4.9	+
1.0	+
3.5	-
4.1	+
2.5	+
-0.1	-
1.6	-
2.0	+
0.7	+

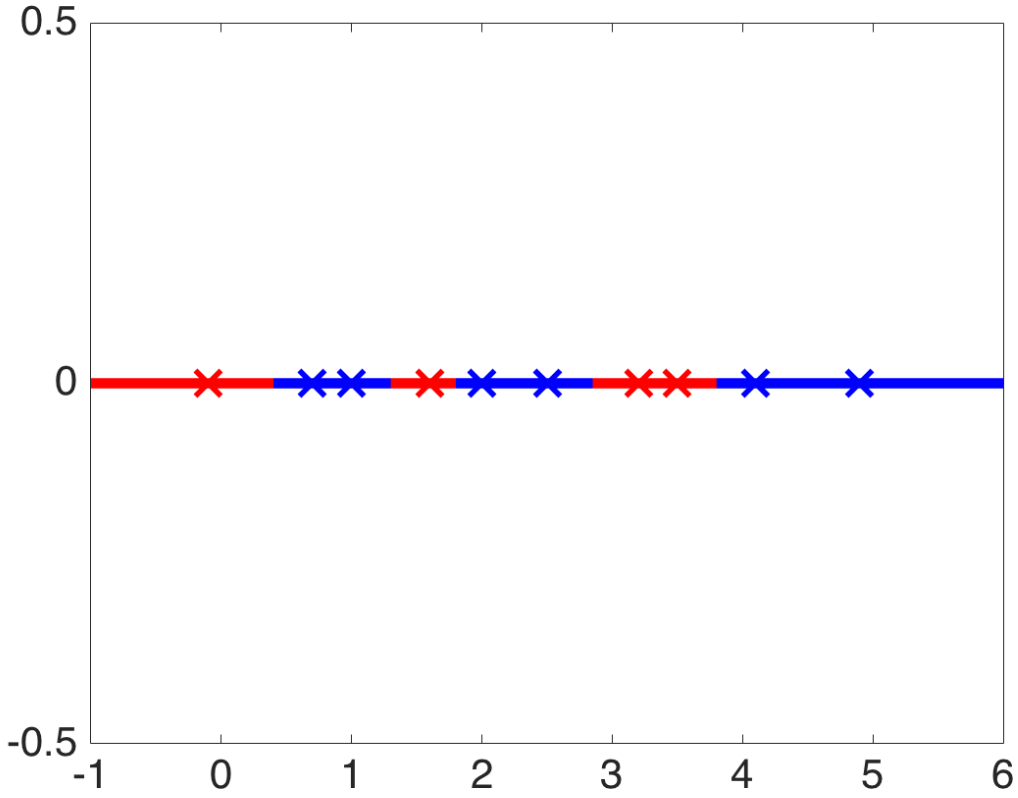
data	class
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

data	class	1-NN	3-NN
-0.1	-	+	+
0.7	+	+	-
1.0	+	+	+
1.6	-	+	+
2.0	+	-	+
2.5	+	+	-
3.2	-	-	+
3.5	-	-	+
4.1	+	-	-
4.9	+	+	-

Ejemplo

El siguiente conjunto de datos corresponde a 10 puntos etiquetados como + y -. Clasifique cada punto en función de sus otros datos aplicando K-NN con K igual a 1 y 3 (utilice la distancia euclidiana).
Para cada valor de K calcule el error de clasificación (número de puntos clasificados erróneamente). Por último, teniendo en cuenta todos los puntos de datos, dibuje el límite de decisión utilizando 1-NN.

data	class	1-NN	3-NN
-0.1	-	+	+
0.7	+	+	-
1.0	+	+	+
1.6	-	+	+
2.0	+	-	+
2.5	+	+	-
3.2	-	-	+
3.5	-	-	+
4.1	+	-	-
4.9	+	+	-



Modelamiento paramétrico

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2