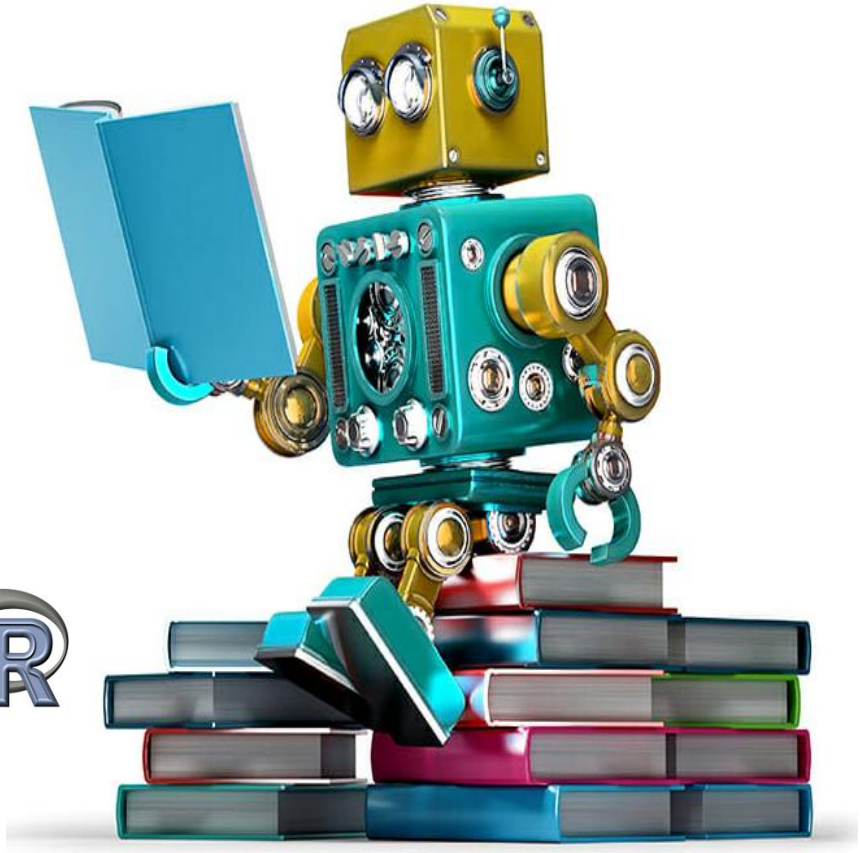
ALGORITMOS DE CLASIFICACIÓN : REGRESIÓN LOGÍSTICA – K-NN – NAIVE BAYES



«Lo llaman suerte, pero es constancia. Lo llaman casualidad, pero es disciplina. Lo llaman genética pero es sacrificio. Ellos hablan , tú estudia.»»

San Marcos Data Science Community

Mentores

José Antonio Cárdenas Garro
ESTADÍSTICA
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricard Palma

Predictive Modelling Specialist Scotiabank

Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Risk Modeling | Data Science

André Omar Chávez Panduro
ESTADÍSTICA
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricard Palma

Data Scientist Interbank

Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Business Analytics| Data Science

Aldo Ray Chávez Panduro
INGENIERÍA DE SISTEMAS
UNMSM
Student of MSc in Data Science
Universidad Ricard Palma

Risk Management Specialist BBVA Continental

Especialidades : Big Data| Machine Learning | Programming | Risk Specialist – IFRS9 | Data Science

SMDSC

Agenda

- Introducción a los Problemas de Clasificación o Aprendizaje Supervisado.
- Regresión Logística Binaria.
- K – Vecinos más cercanos.
- Elección del K – óptimo.
- Clasificación Bayesiana : Naive Bayes.

DEFINICIONES BÀSICAS

- **Conjunto de Datos (Data Set):** El total del conjunto de datos sobre los que queremos desarrollar un algoritmo de Machine Learning con el fin de obtener un modelo que lo represente lo mejor posible. Contendrá variables independientes y dependientes.

- **Variables Independientes (Features), (VI):** Aquellas columnas del Data Set que serán usadas por el algoritmo para generar un modelo que prediga lo mejor posible las variables dependientes.

- **Variables dependientes (Labels,Target), (VD):** Columna del data set que responde a una correlación de VI y que debe ser predicha por el futuro modelo

- **Conjunto de Datos de Entrenamiento (Training Set):** Subconjunto del Data Set que será utilizado para entrenar el modelo que se pretende generar.

- **Conjunto de Datos de Test (Test Set):** Subconjunto del data set que se le pasará al modelo una vez haya sido entrenado para comprobar, mediante el uso de diferentes métricas, sus indicadores más importantes de calidad.

CLASIFICACIÓN: DEFINICIÓN

- Dada una colección de registros (Conjunto de Entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con una variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

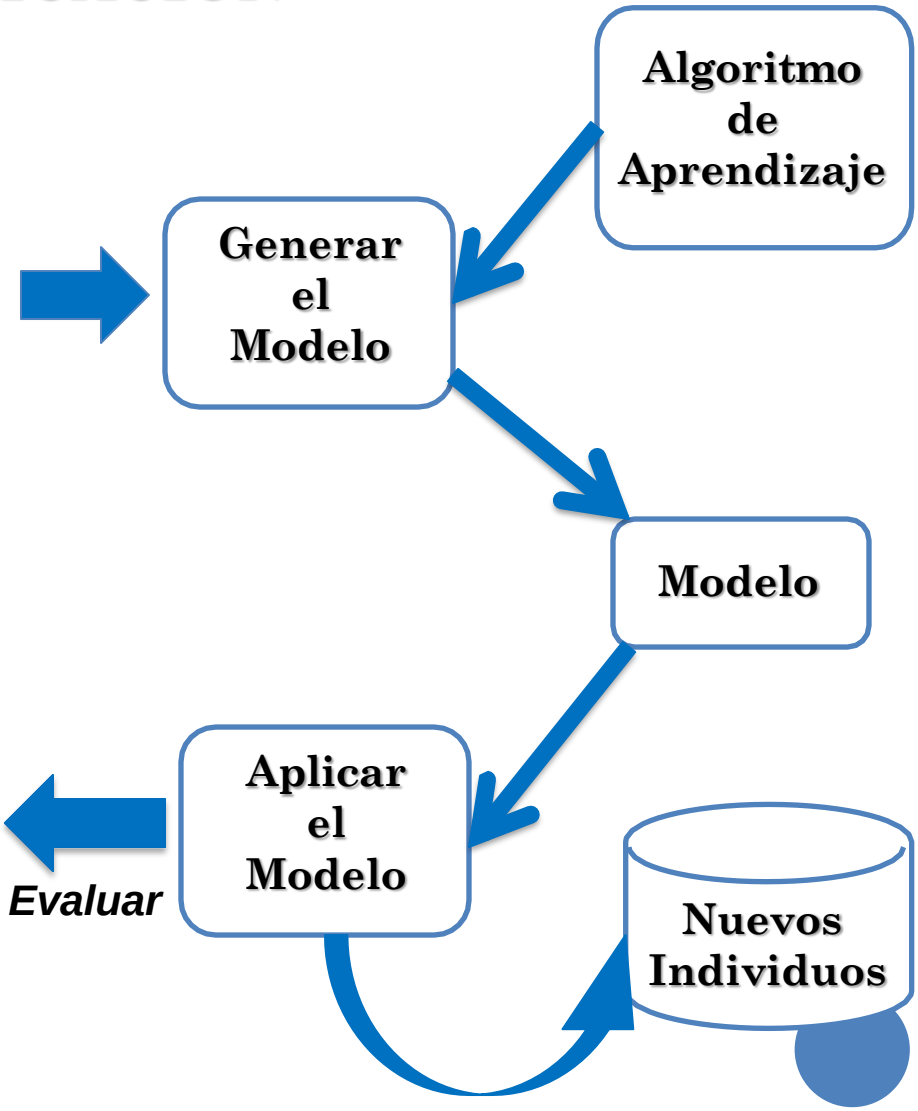
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing

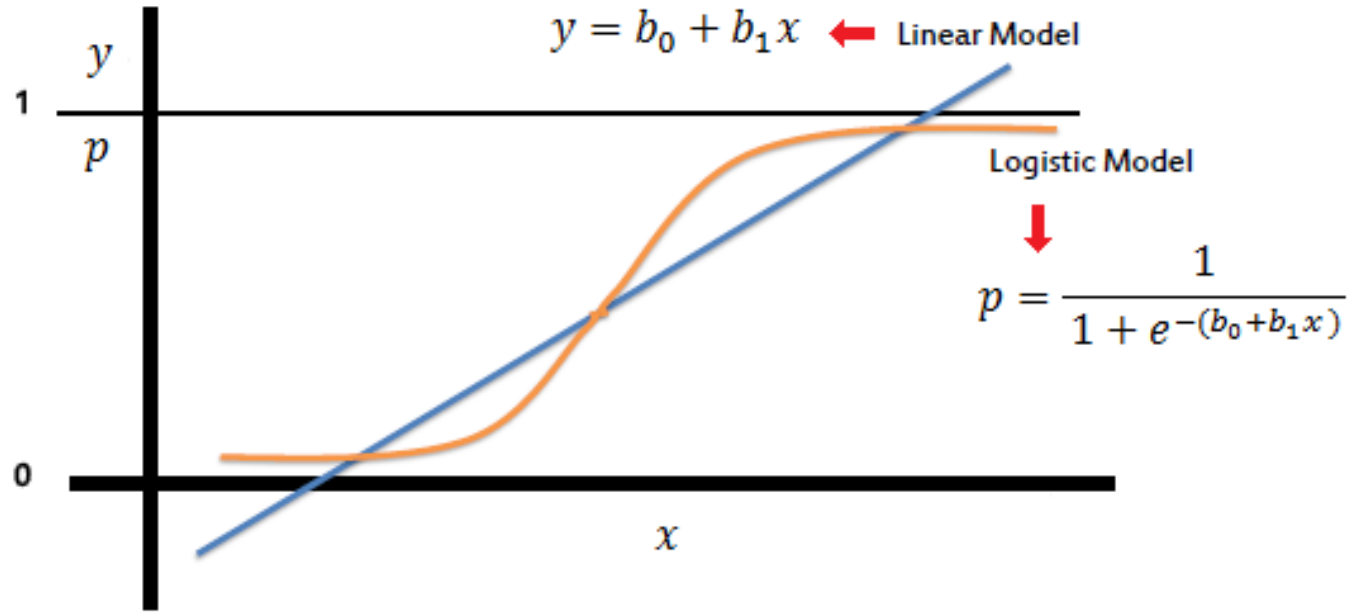


DEFINICIÓN DE CLASIFICACIÓN

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j.

- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Beyesiana.

Regresión Logística



DEFINICIÓN

- Es un modelo predictivo supervisado.
- La regresión logística es un modelo de elección discreta en el que la variable dependiente es cualitativa.
- Es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas.

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

- Para este modelo se considera que la variable respuesta, es una variable dicotómica que tomas dos valores.
- Para estos modelos dicotómicos, las dos categorías deben de ser mutuamente excluyentes.
- La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \Pr ob(Y_i = 1) = P_i \\ 0, \Pr ob(Y_i = 0) = 1 - P_i \end{cases}$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

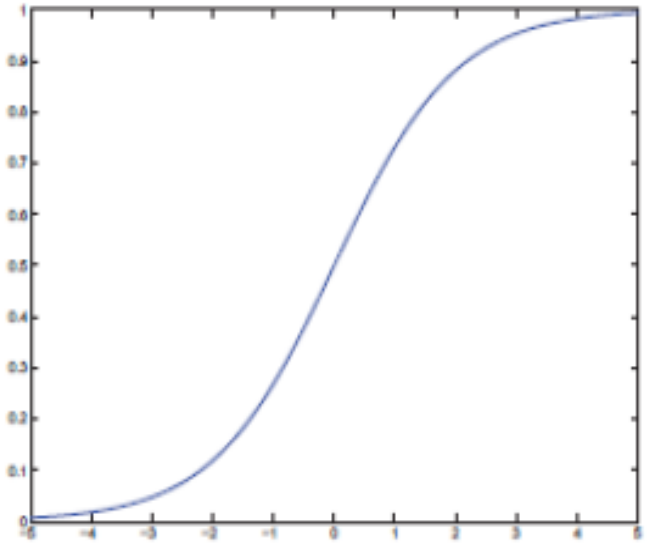
Ejemplo:

La variable Morosidad toma los siguientes valores:

- "1" si el cliente es moroso.
- "0" si el cliente es no moroso.

¿Es dicotómica?
¿Es cualitativa?
¿Es mutuamente excluyente?

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



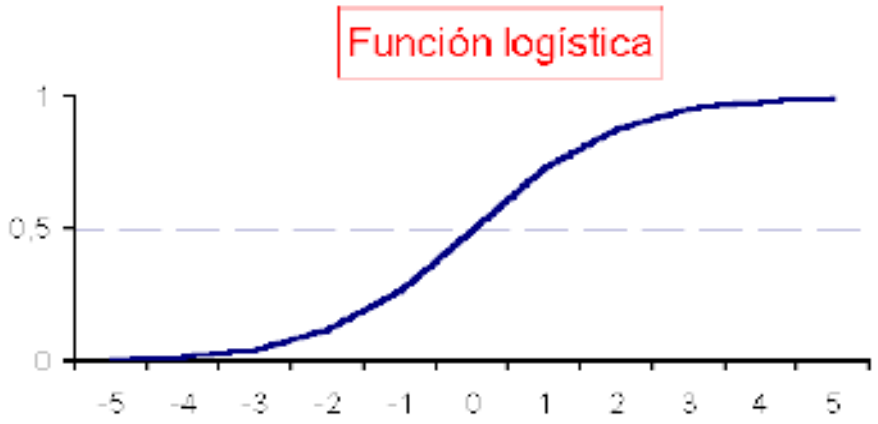
Se basa en la función logística:

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\frac{1}{e^{z}}} = \frac{e^{z}}{1+e^{z}}$$

Está acotada entre 0 y 1:

$$\lim_{z\to-\infty} f(z) = 0, \quad \lim_{z\to\infty} f(z) = 1.$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



La representación matemática del modelo es la siguiente:

$$z_i = \log \frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

z_i : Variable dependiente del modelo: "Moroso" y " No Moroso"

p_i : Probabilidad de que el cliente sea "Moroso"

β_i : Coeficientes del modelo (parámetros a estimar)

x_i : Variables explicativas del modelo

SMDSC

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

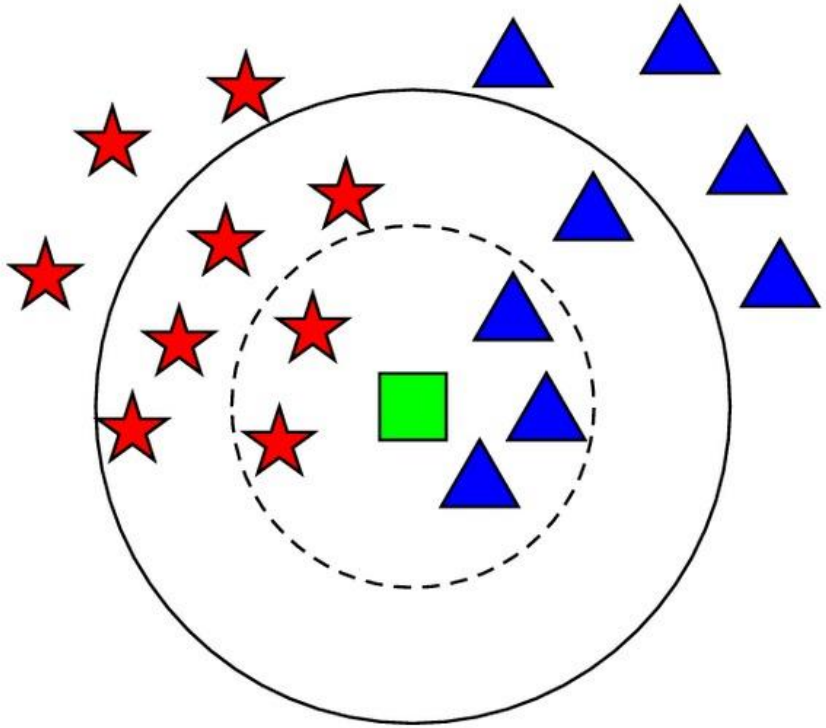
Odds Ratio

Es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso.

$$z_i = \log\frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

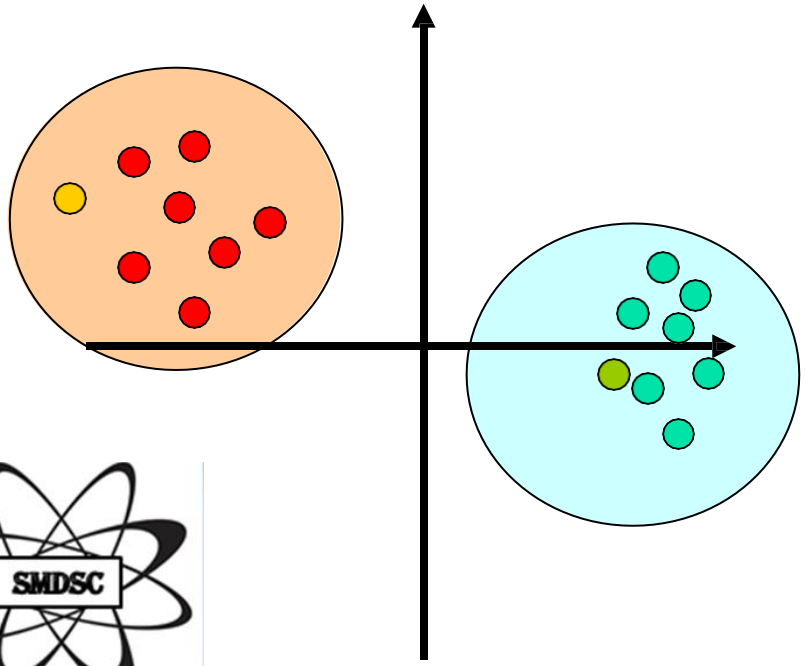
Indica cuánto más probable es ser un cliente "Moroso" que "No Moroso"

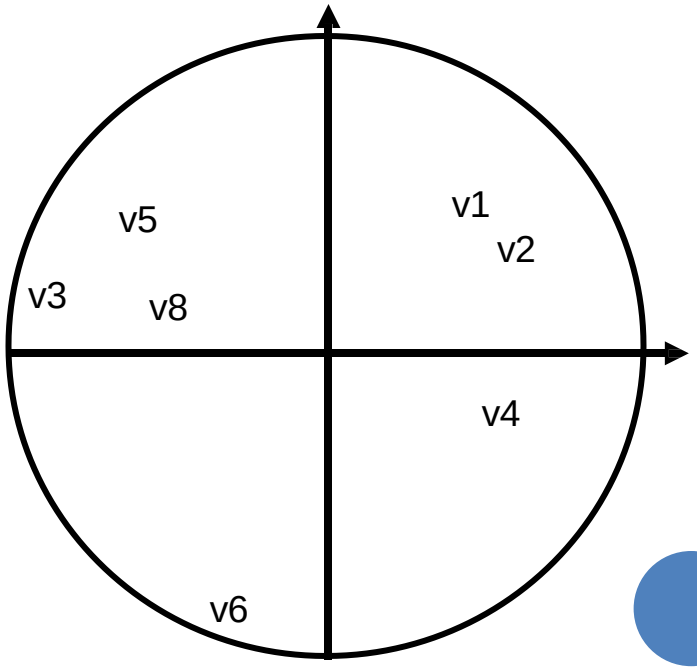
K – Vecinos más cercanos



Clasificación mediante k-vecinos más cercanos

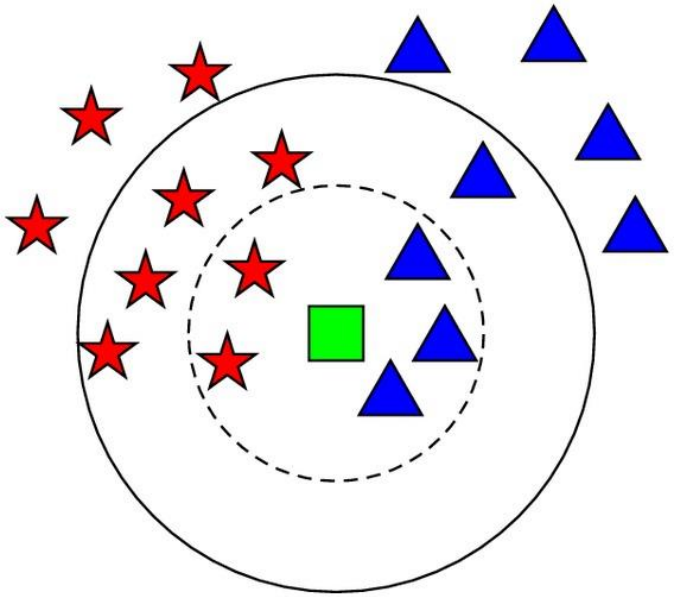
- Análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En aprendizaje de máquinas, se ha desarrollado como una forma de reconocer patrones de datos sin requerir una coincidencia exacta con patrones o casos almacenados.





Clasificación mediante k-vecinos más cercanos

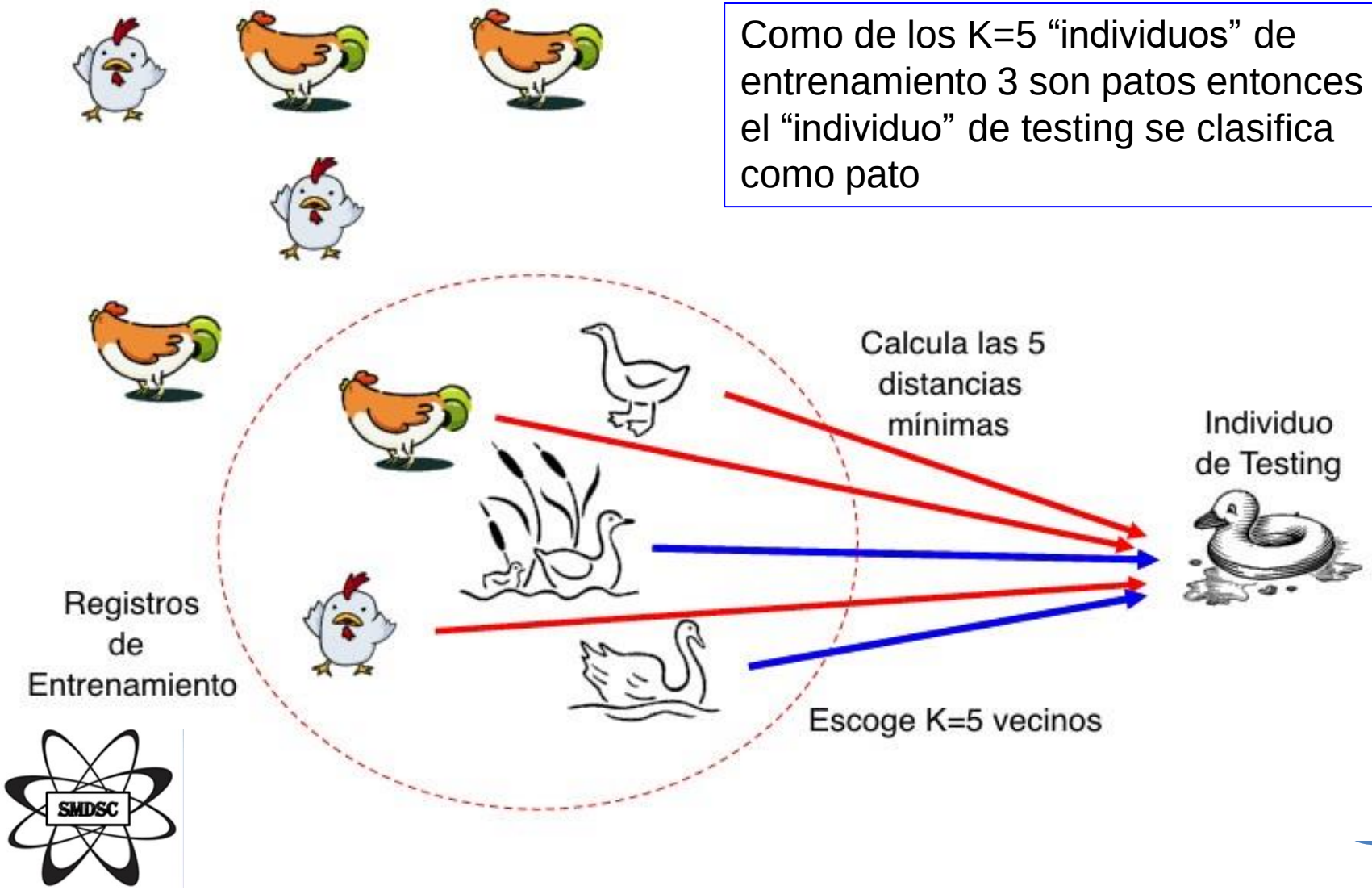
- Los casos similares están cercanos entre sí y los casos no similares están distantes entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.



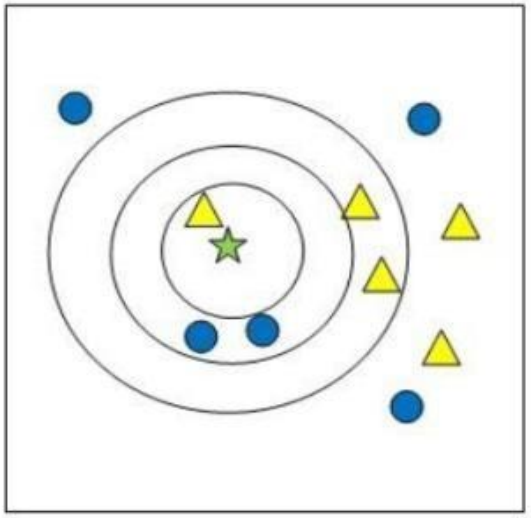
SMDSC

Clasificación mediante k-vecinos más cercanos : Idea Intuitiva

Como de los K=5 “individuos” de entrenamiento 3 son patos entonces el “individuo” de testing se clasifica como pato



Clasificación mediante k-vecinos más cercanos



Para K=1 (círculo más pequeño), la clase de la nueva instancia sería la Clase 1, ya que es la clase de su vecino más cercano, mientras que para K=3 la clase de la nueva instancia sería la Clase 2 pues habrían dos vecinos de la Clase 2 y solo 1 de la Clase 1

SMDSC

Clasificación mediante k-vecinos más cercanos : Algoritmo

COMIENZO

Entrada: $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$

$\mathbf{x} = (x_1, \ldots, x_n)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (x_i, c_i)

calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$

Ordenar $d_i (i = 1, \ldots, N)$ en orden ascendente

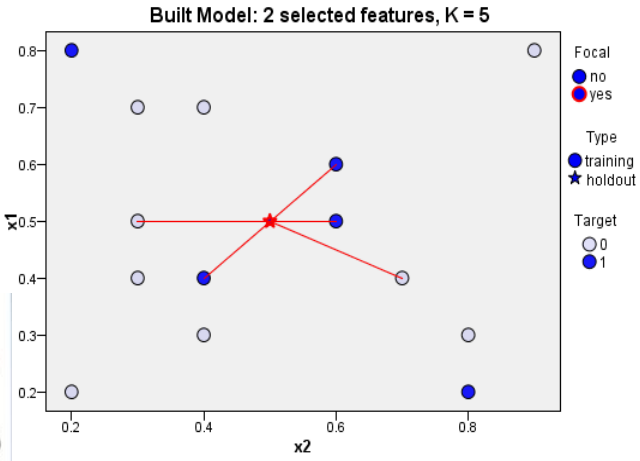
Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$

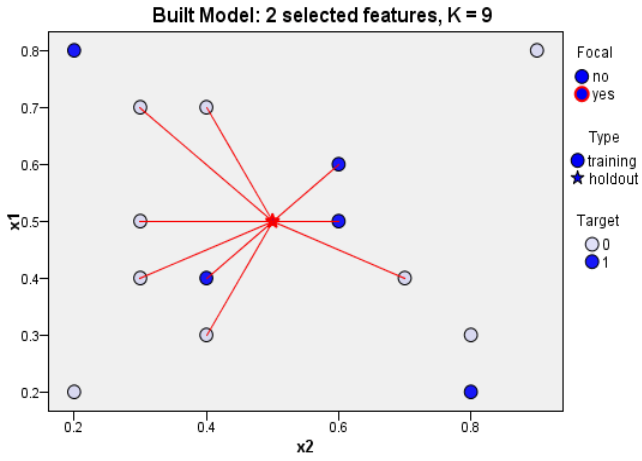
Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$

FIN

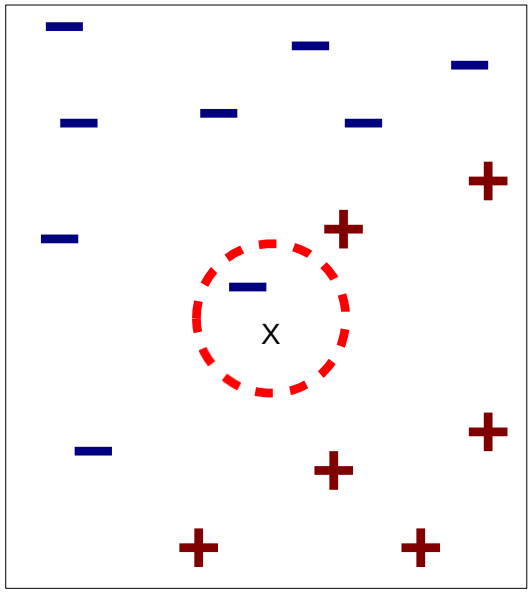
Clasificación mediante k-vecinos más cercanos : Algoritmo

- Los casos muy cercanos a otros se denominan “vecinos”. Cuando se presenta un nuevo caso (reserva), se calcula su distancia desde cada caso del modelo. Las clasificaciones de la mayoría de casos similares (los vecinos más próximos) se anotan y el nuevo caso se coloca en la categoría que contiene el mayor número de vecinos más próximos.
- Puede especificar el número de vecinos más próximos que se van a examinar; este valor se denomina *k*.

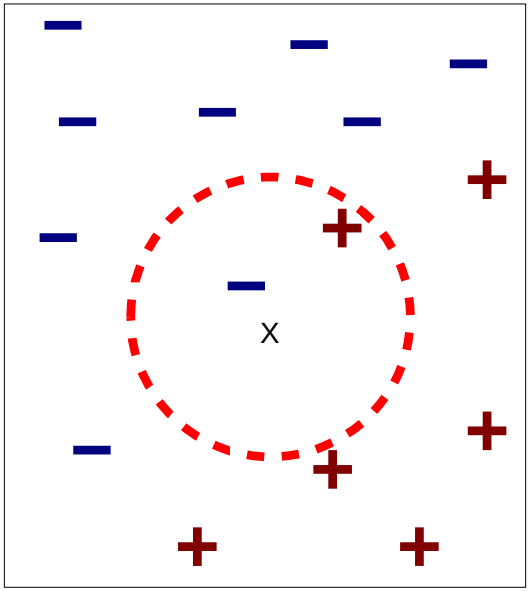




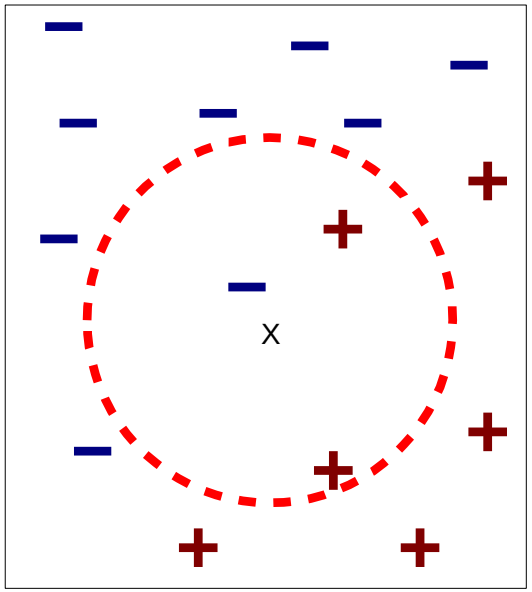
Clasificación mediante k-vecinos más cercanos : ¿Elección del k óptimo?



(a) 1 - Vecino más cercano.



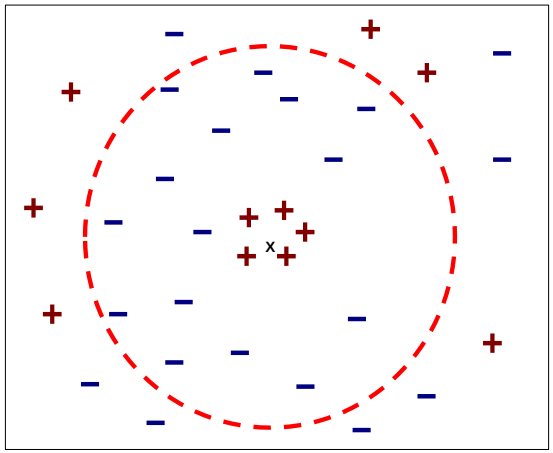
(b) 2 - Vecinos más cercanos.



(c) 3- Vecinos más cercanos.

Clasificación mediante k-vecinos más cercanos : ¿Elección del k óptimo?

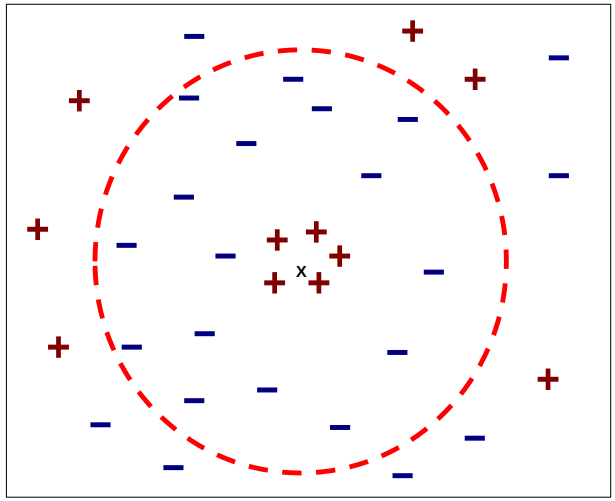
- Escogiendo el valor de K:
  - ✓ Si K es muy pequeño el modelo será muy sentitivo a puntos que son atípicos o que son ruido (datos corruptos)
  - ✓ Si K es muy grande, el modelo tiende a asignar siempre a la clase más grande.



SMDSC

Clasificación mediante k-vecinos más cercanos : ¿Elección del k óptimo?

- Escogiendo el valor de K:
  - ✓ Mediante la Tabla de Aprendizaje el modelo escogerá el valor de K que mejor clasificación logre en esta tabla, es decir, prueba con K=1, K=2, ….
  - ✓ Esto pude ser muy caro computacionalmente.

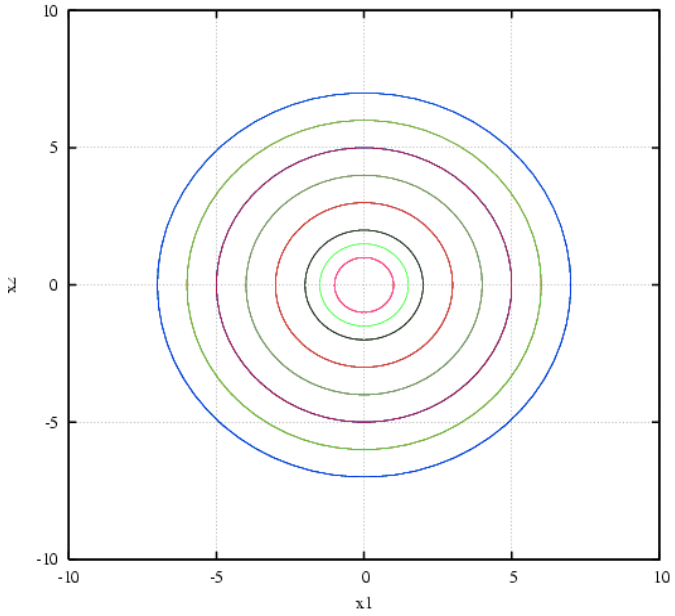


SMDSC

Clasificación mediante k-vecinos más cercanos : Elección de la distancia

$$d(A,B) \equiv \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} = \sqrt{(A-B)^T (A-B)}$$

Distancia Euclídea



Clasificación mediante k-vecinos más cercanos : Distancia Euclídea

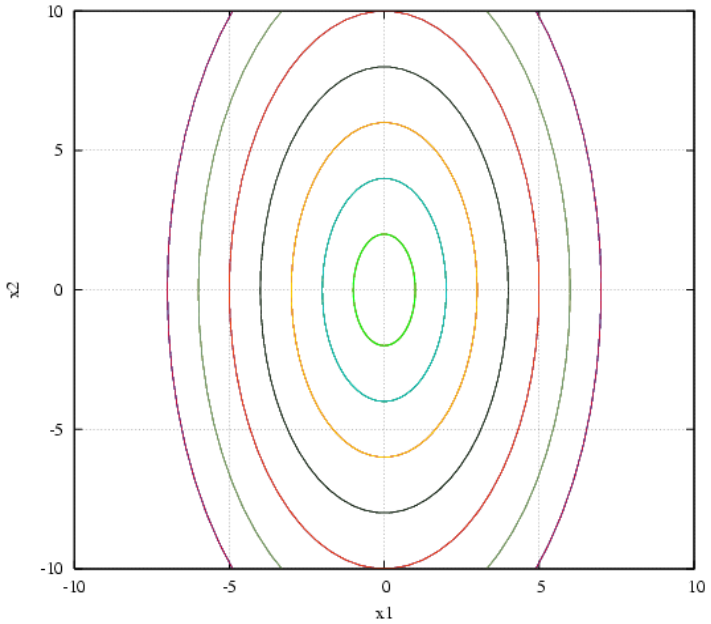
$$d(A,B) \equiv \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} = \sqrt{(A-B)^T(A-B)}$$

- Algunas consideraciones importantes para el uso de ésta ditancia, es que no está acotada, es sensible a cambios de escalas y considera las T variables estocásticamente independientes.

Clasificación mediante k-vecinos más cercanos : Elección de la distancia

$$d(A,B) \equiv \sqrt{(A-B)^T M^T M (A-B)}$$

Distancia Euclídea Ponderada

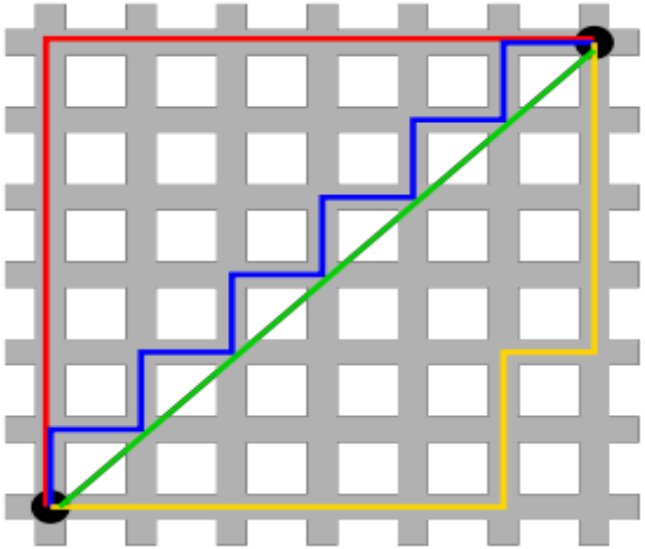


SMDSC

Clasificación mediante k-vecinos más cercanos : Elección de métrica de distancia

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Distancia Manhattan (city-block)



Clasificación mediante k-vecinos más cercanos : Distancia de Manhattan

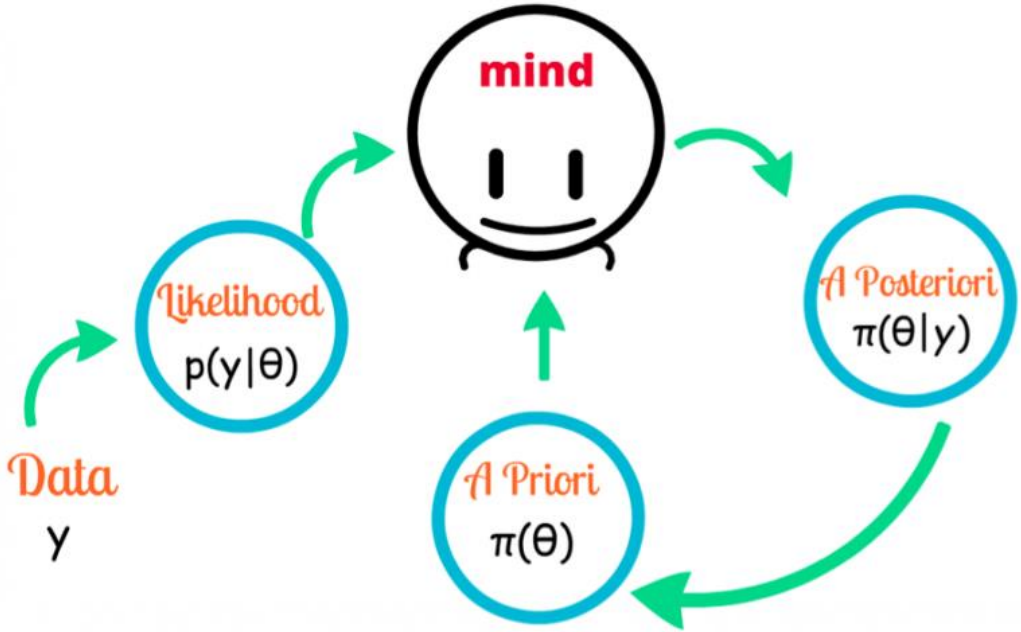
- La distancia de Manhattan o métrica city-block o ciudad, calculada como la suma de las diferencias absolutas entre unidades para cada variable, es un caso particular de la distancia de **Minkowski**. Es menos sensible a valores muy grandes o aberrantes, ya que es función de diferencias absolutas en lugar de diferencias al cuadrado, adicionalmente cada variable puede ser estandarizada por su rango.

$$d(x, y) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Clasificación mediante k-vecinos más cercanos : Elección de la distancia

Distancias y disimilaridades		Métrica	Euclídea
Euclídea	$\sqrt{\sum_{i=1}^{T}(x_{ii}-x_{ji})^2}$	Si	Si
Manhattan	$\sum_{i=1}^{T}\lvert x_{ii}-x_{ji}\rvert$	Si	No
Bray-Curtis	$\frac{\sum_{i=1}^{T}\lvert x_{ii}-x_{ji}\rvert}{\sum_{i=1}^{T}(x_{ii}+x_{ji})}$	Si	No
Canberra	$\sum_{i=1}^{T}\frac{\lvert x_{ii}-x_{ji}\rvert}{(x_{ii}+x_{ji})}$	Si	No
Minkowski	$\sqrt[q]{\sum_{i=1}^{T}\lvert x_{ii}-x_{ji}\rvert^q}$	Si	Si
Mahalanobis	$\sqrt{\sum_{i=1}^{T}\sum_{i=1}^{T}(x_{ii}-x_{ji})\sigma_{ii}^{-1}(x_{ii}-x_{ji})}$	Si	Si

Modelos Bayesianos : Naive Bayes



Introducción

- Estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados.
- El teorema que lleva su nombre se refiere a la probabilidad de un suceso condicionado por la ocurrencia de otro suceso.

Thomas Bayes

Definición

- Es un método importante no sólo porque ofrece un análisis cualitativo de las atributos y valores que pueden intervenir en el problema, sino porque da cuenta también de la importancia cuantitativa de esos atributos.

- En el aspecto cualitativo podemos representar cómo se relacionan esos atributos ya sea en una forma causal, o señalando simplemente de la correlación que existe entre esas variables (o atributos). Cuantitativamente (y ésta es la gran aportación de los métodos bayesianos).

Definición

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Ejemplo para Variables Cualitativas

Ejemplo 1.(atributos discretos solamente)

X1	X2	X3	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
0	0	1	1
1	1	0	1

Ejemplo para Variables Cuantitativas

Clasificador Naïve Bayes (cont)

En este caso

$$P[X_j = a_j / C_i] = \frac{1}{s_j\sqrt{2\pi}} \exp[\frac{-(a_i - \bar{x}_j)^2}{2s_j^2}]$$

Donde, $\bar{x}_j$ y s_j son la media y la varianza de los valores de la variable X_j en la clase C_i.

La libreria e1071 de R contiene una funcion **naiveBayes** que calcula el clasificador naïve Bayes, tanto para datos discretos como continuos.

¿Qué es el Machine Learning (ML)?

- Rama de la inteligencia artificial que pretende que una máquina sea capaz de mejorar su actuación a la hora de resolver un problema mediante la adquisición de experiencia en una tarea determinada.

- **Multitud de algoritmos** con finalidades específicas.

- Ramas de Machine Learning:

  - ✓ Supervised Learning
  - ✓ Unsupervised Learning
  - ✓ Reinforcement Learning
  - ✓ Deep Learning

J

¡Gracias!

San Marcos Data Science Community

Auspicio : Escuela Académica Profesional de Estadística
San Marcos Data Science Community.

¿Preguntas?
Realicemos el taller