Clustering: Auto-associative Multivariate Regression Tree (AAMRT)

Miguel Bernal C
Quantil

12 de diciembre de 2013

Contenido

Análisis no supervisado

Observaciones sin marcar, i.e. no hay variable objetivo.

- Reglas de asociación
- Clustering
- Self organizing maps
- Etc.

Clustering

Segmentar observaciones en grupos para que dentro de éstos sean lo más homogéneos posibles y para que entre ellos sean altamente heterogéneos.

- Combinatorial
- Mixture modelling
- Mode seeking

Mucha influencia del investigador (e.g. Métrica, algoritmo, número de clusters, etc.)

Cosas claves

Matriz de disimilitud

- La disimilitud entre la observación i y la i':

$$D(x_i, x_{i'}) = \sum_{j=1}^{P} \alpha_j d_j(x_{ij}, x_{i'j}); \sum_{j=1}^{P} \alpha_j = 1$$

donde $d_j(x_{ij}, x_{i'j})$ es la disimilitud entre los valores del atributo j, y α_j es el peso del atributo j.

- Note que darle el mismo peso a todos los atributos no implica necesariamente darles la misma influencia.

Cosas claves

Tipos de variables

- Variables cuantitativas: Muchas medidas posibles, se recomienda primero normalizarlas.
- Variables ordinales

Se recomienda recodificar las variables entre 0 y 1, y luego utilizar la medida de las cuantitativas.
e.g. $\frac{i-\frac{1}{2}}{M}, i = 1, ..., M$ para M valores.

- Variable nominales

Existen varias medidas, la más popular es 0 si son la misma categoría y 1 si son de categoría diferente.
Datos faltantes: Depende del caso particular.

Cosas claves

¡Más importante seleccionar correctamente la medida de disimilitud y los pesos que el algoritmo!

Contenido

Combinatorial

Utiliza directamente los datos sin usar ningún modelo de probabilidad subyacente.

- K-Medias
- K-Medoids
- Jerárquicos
  1. Agglomerativos
  2. Divisivos
- Muchos otros (e.g. Spectral Clustering, DBSCAN, Affinity Propagation, etc.)

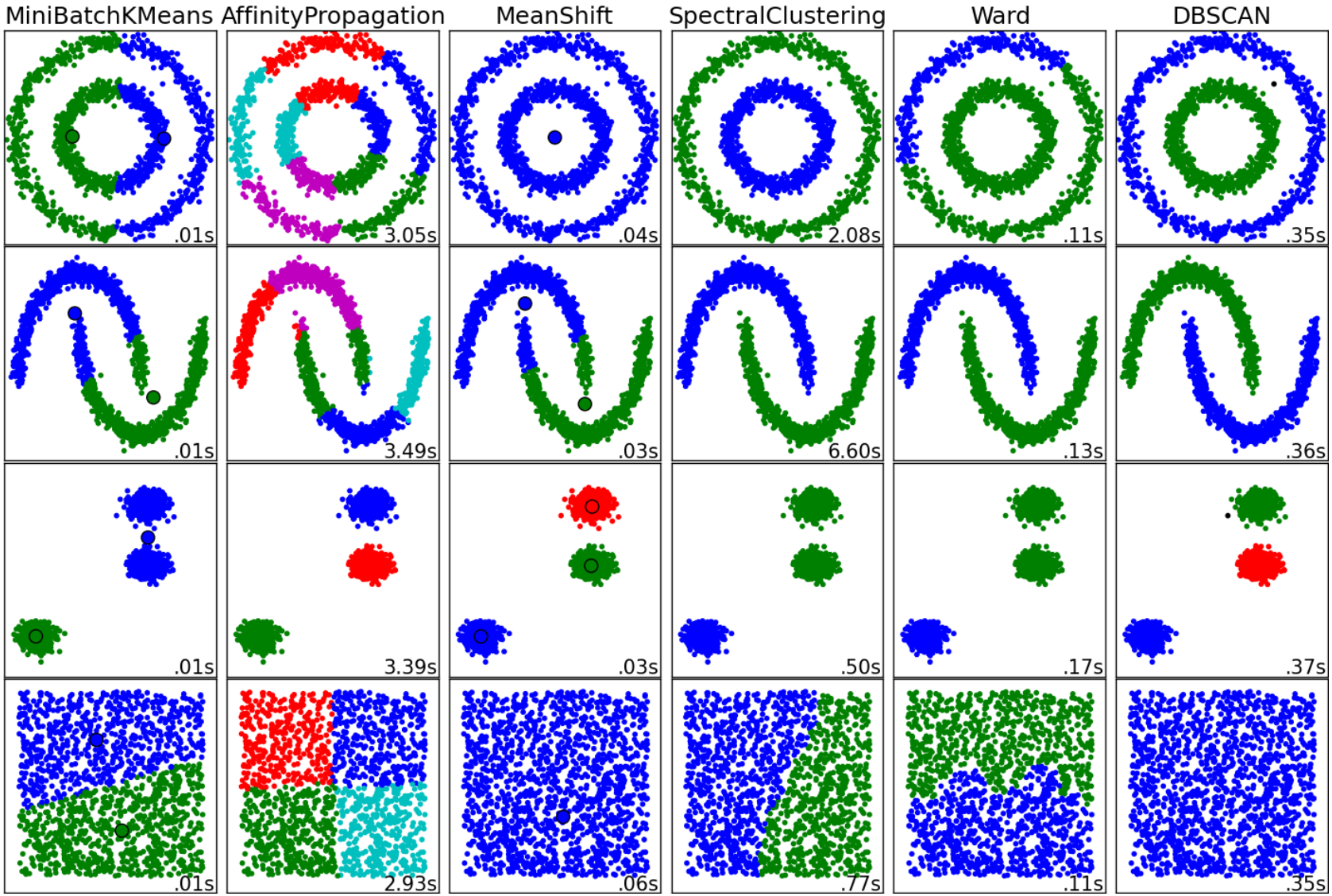
Mixture modelling

- Supone que los datos son una muestra i.i.d de una población descrita por una función de densidad de probabilidad.
- La función de densidad está caracterizada por un modelo paramétrico de mixtura de funciones de densidad.
- Permite estimar la probabilidad que una observación i pertenezca a un grupo m.
- e.g. Mixturas Gaussianas
- $f(x) = \sum_{m=1}^{M} \alpha_m \phi(x)$

Mode seeking

- Trata de estimar directamente modas diferentes de la función de densidad de probabilidad.
- Aproximación no paramétrica
- Las observaciones más cercanas a la respectiva moda hacen parte de un clúster.

Ejemplo


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
DBSCAN
.01s
3.05s
.04s
2.08s
.11s
.35s
.01s
3.49s
.03s
6.60s
.13s
.36s
.01s
3.39s
.03s
.50s
.17s
.37s
.01s
2.93s
.06s
.77s
.11s
.35s

Contenido

Validación

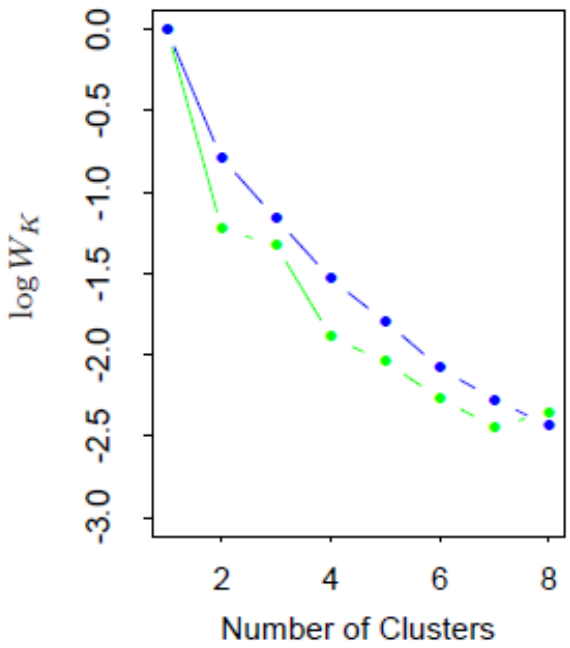
- Número de clusters
- Selección de algoritmos
- Comparar clusters
- Comparar grupos de clusters
- Evitar encontrar patrones en ruido

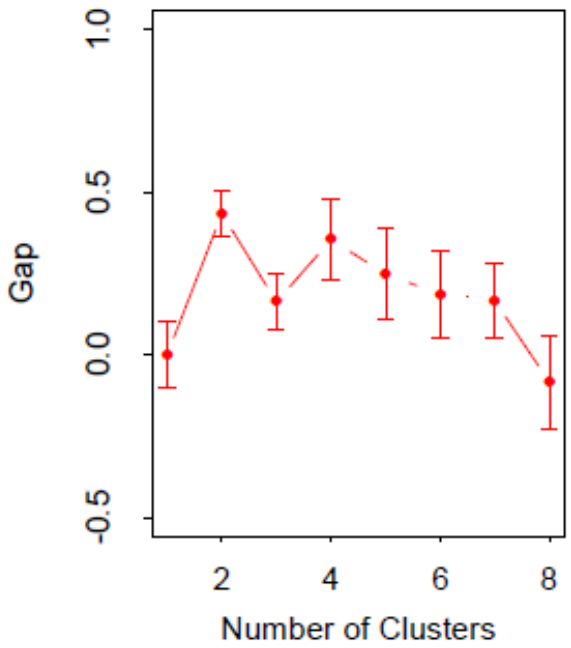
Medidas

Muchísimas

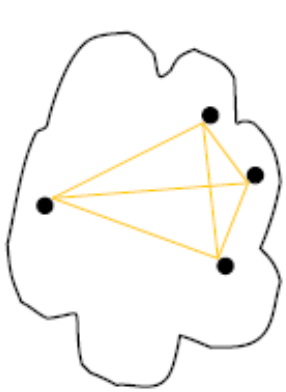
- Externas
- Internas
  1. Cohesión
  2. Separabilidad
  3. Mixtas
  4. e.g. Hubert, Calinski-Harabasz, Dunn, I, SSE, Gap, etc.

Gap

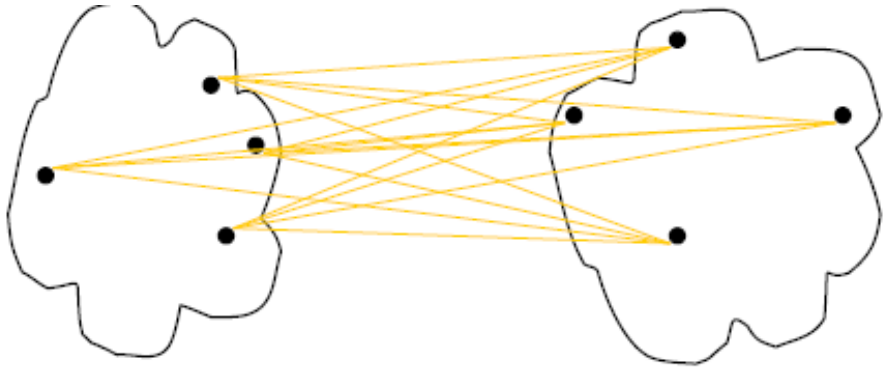

$\log W_K$
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
2
4
6
8
Number of Clusters


Gap
1.0
0.5
0.0
-0.5
2
4
6
8
Number of Clusters

Medidas



cohesion



separation

«The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.»

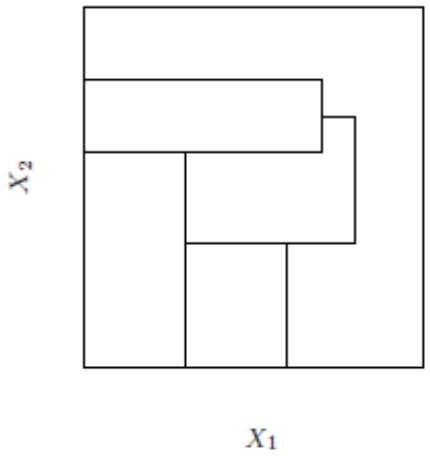
Motivación

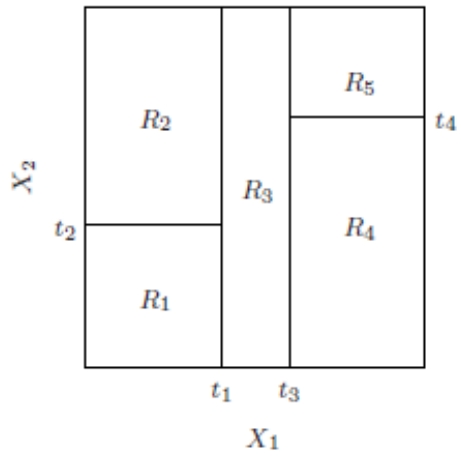
- Carga computacional
- Datos nuevos
- Implementación en sistemas en tiempo real.
- Selección de variables

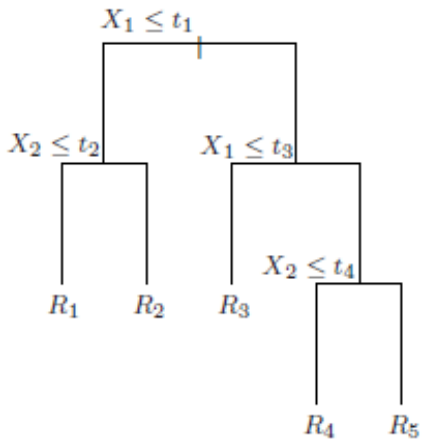
Contenido

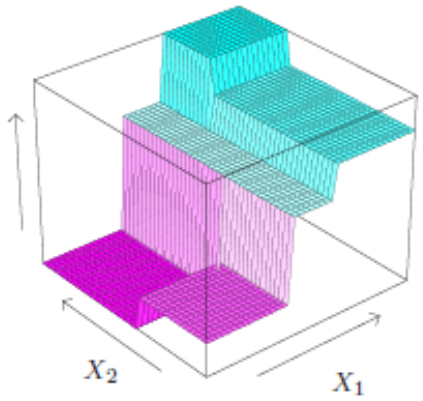
Árboles de clasificación

Partir el espacio de manera binaria en rectángulos.





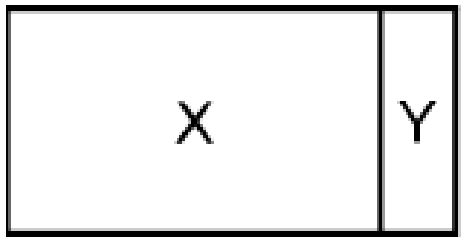




Una mirada a todos

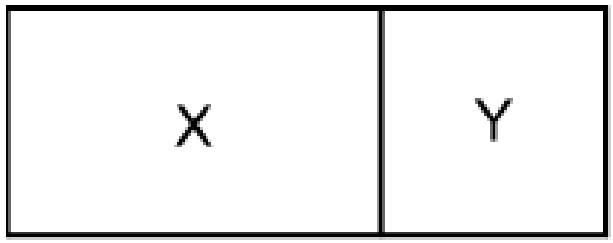
(a) CART (supervised)

explanatory variables response variable



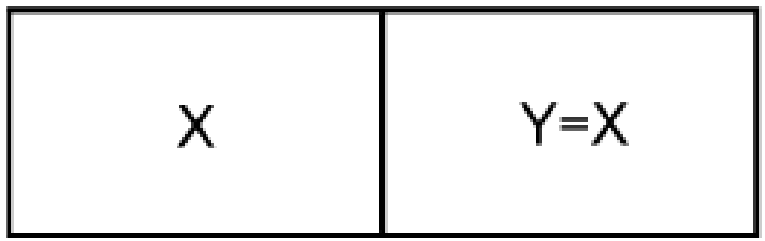
(b) MRT (supervised)

explanatory variables response variables



(c) AAMRT (unsupervised)

explanatory variables response variables



CART

Rápidamente:

1. Todos los objetos se asignan a un nodo.
2. Partir la variable objetivo usando todas las variables explicativas en todas sus posibles particiones.
3. Seleccionar la variable y la partición que más reduce la impureza. Partir el nodo padre de acuerdo a esta partición.
4. Repetir los pasos 2 y 3 hasta que se cumpla un criterio (e.g. tamaño del árbol, impureza de los nodos, etc.)
5. Podar el árbol utilizando cross-validation y así escoger el tamaño óptimo del árbol.

Todo esto se puede ver formalmente.

Continuas vs Categóricas

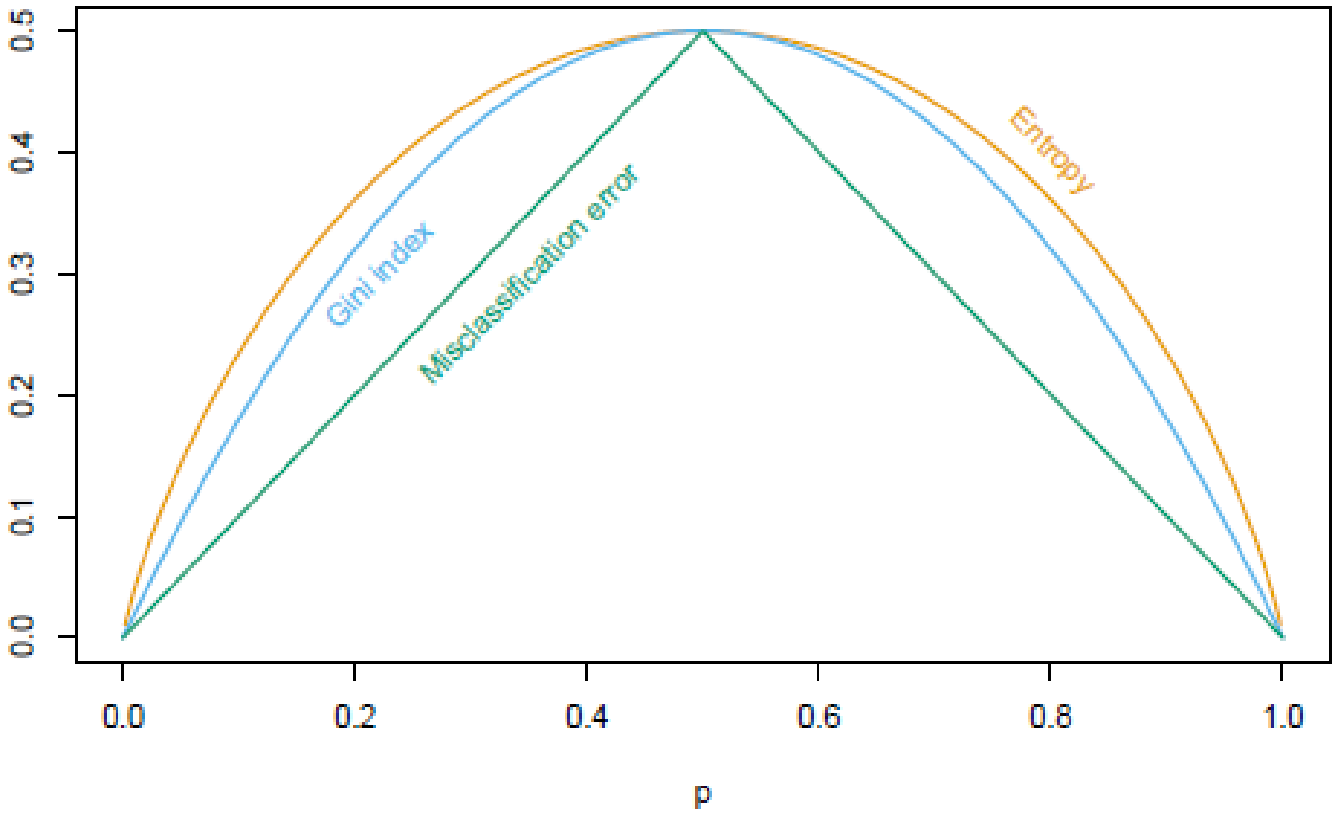
Para un nodo m en una región R_m con N_m observaciones, la proporción de la clase k en el nodo m es:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Se clasifica un nodo m a la clase $k(m) = arg\,\text{máx}_k\, \hat{p}_{mk}$

1. Cuadrático: $Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \bar{y})^2$
2. Misclassification: $1 - \hat{p}_{mk(m)}$
3. Gini: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
4. Cross-entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Impureza


0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
p
Entropy
Gini index
Misclassification error

Partición

- Partición:

Variable j división s. Así se definen 2 subconjuntos:
$R_1(j,s) = \{X|X_j \leq s\}$ y $R_2(j,s) = \{X|X_j > s\}$

$$\text{mín}_{j,s}\left[\text{mín}\sum_{x_i \in R_1}(y_i - \bar{y})^2 + \text{mín}\sum_{x_i \in R_2}(y_i - \bar{y})^2\right]$$

¿Hasta dónde?

1. Ganancia... No siempre (Mejores divisiones más adelante)
2. Número de nodos
3. Mínimo número de elementos en el nodo
4. Podar un árbol grande T_0

Poda

Minimizar el criterio de costo complejidad, es decir, para cada α encontrar el árbol $T_\alpha \subseteq T_0$ que minimice:

$$C_{(\alpha)}(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

Donde $|T|$ es el número de nodos del árbol y α es el parámetro que determina el trade off entre el tamaño del árbol y el ajuste a los datos. Note que $\alpha = 0$ sería tomar todo el árbol T_0.

Se puede mostrar que para cada α existe un único árbol T_α que minimiza $C_{(\alpha)}$. Luego se escoge $\hat{\alpha}$ utilizando cross validation.

Comentarios

1. Importante notar que los árboles son muy inestables. Por eso se puede usar Boosting o Random Forests para reducir su varianza.
2. Ya está muy estudiado teóricamente y programado muy eficientemente.
    1. e.g. Un predictor categórico con q valores diferentes se podría partir en $2^{q-1} - 1$ grupos distintos. Un q grande haría complicado evaluar las particiones pero hay resultados teóricos interesantes que permiten hacerlo rápidamente.

MRT

Extensión con respuesta multivariada a CART. Utilizando impureza error cuadrático:

- En CART:

$$\sum_{m=1}^{M} (y_i - \bar{y})^2$$

- En MRT:

$$\sum_{m=1}^{M} \sum_{j=1}^{P} (y_{i,j} - \bar{y})^2$$

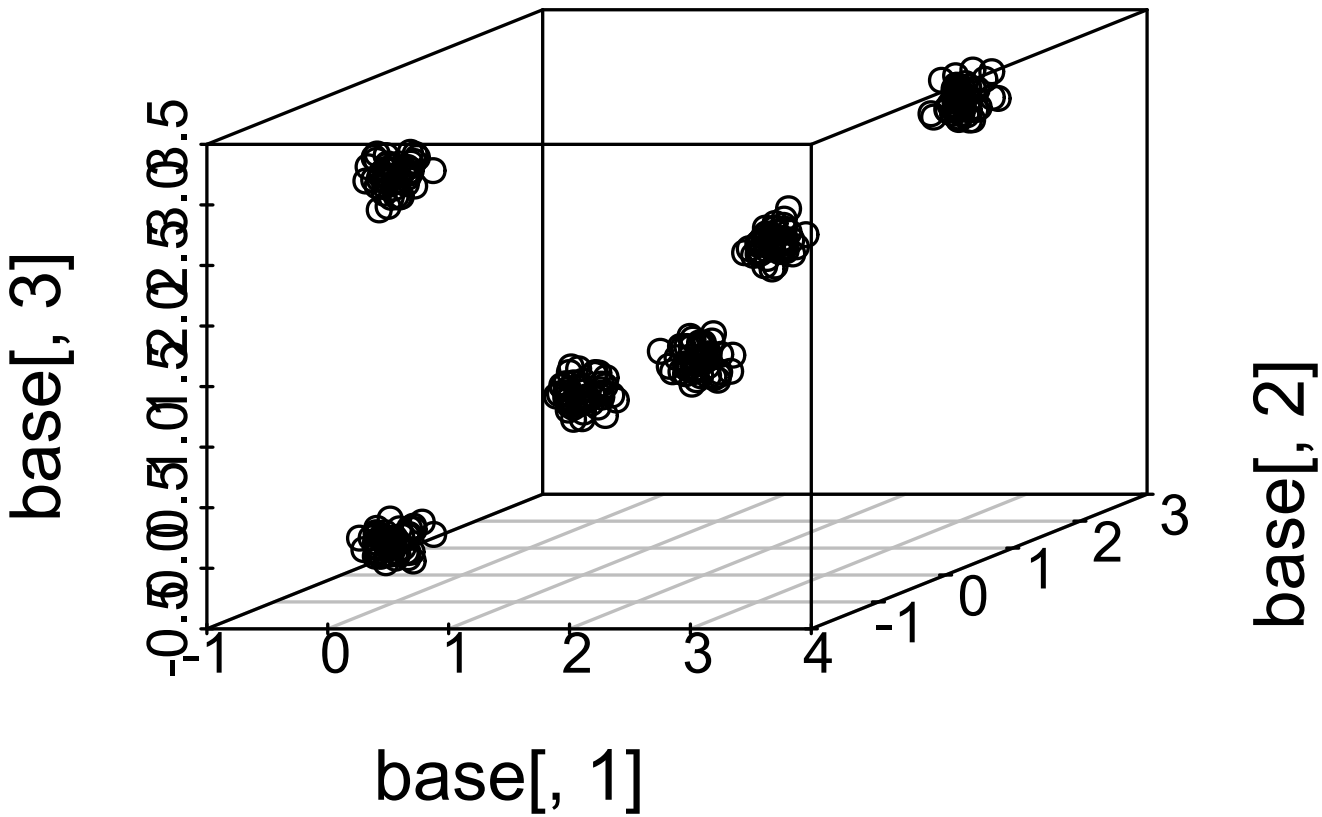
Para P variables objetivo

Todo lo demás de CART se extiende de forma natural a MRT.

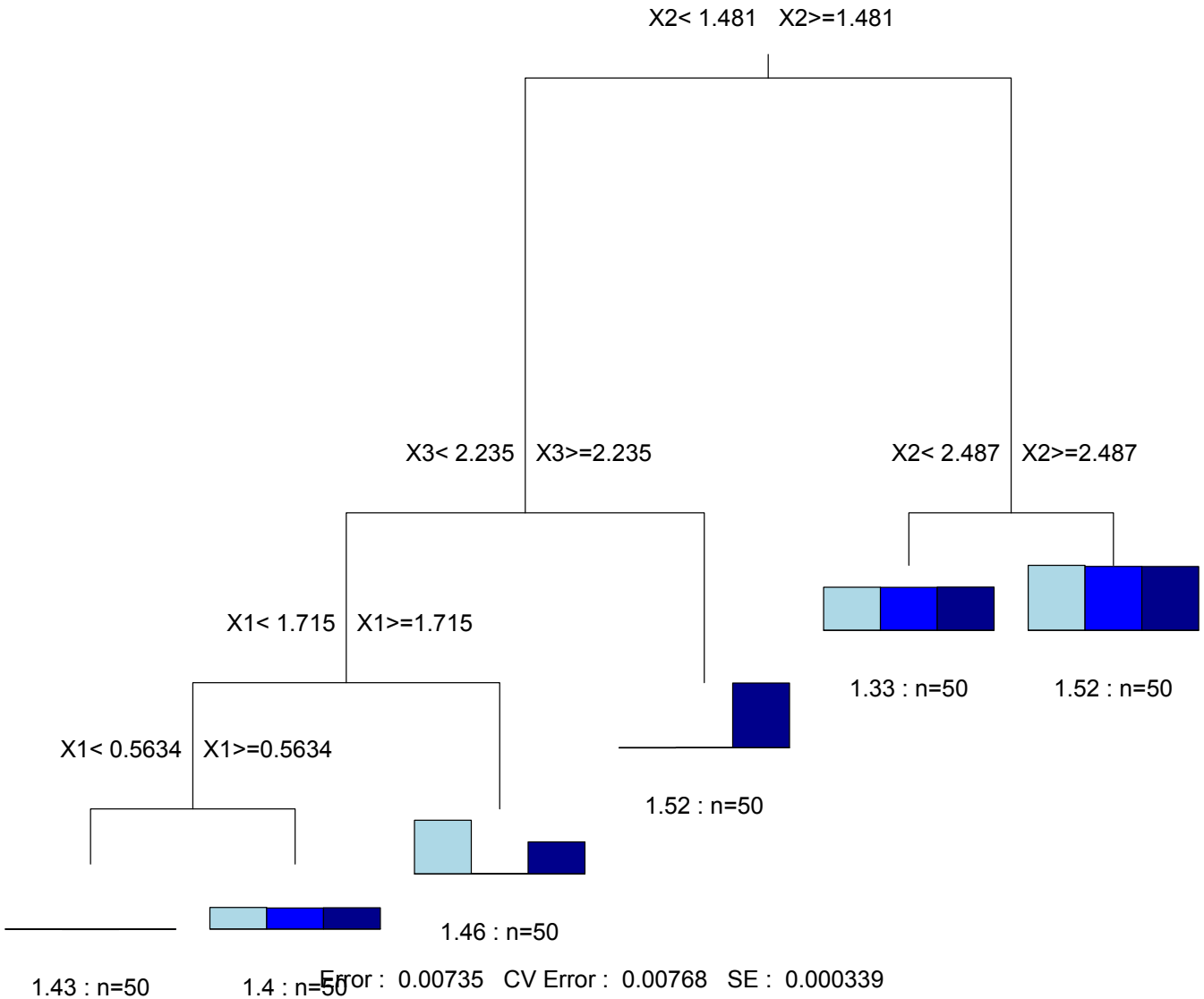
AAMRT

- Variables explicativas son las objetivo.
- Resultados iguales o mejores que K-medias.
- Ayuda a escoger las variables relevantes.
- Estable frente a ruido.
- Muy fácil de comprender e implementar.
- Malo cuando los clusters no son convexos

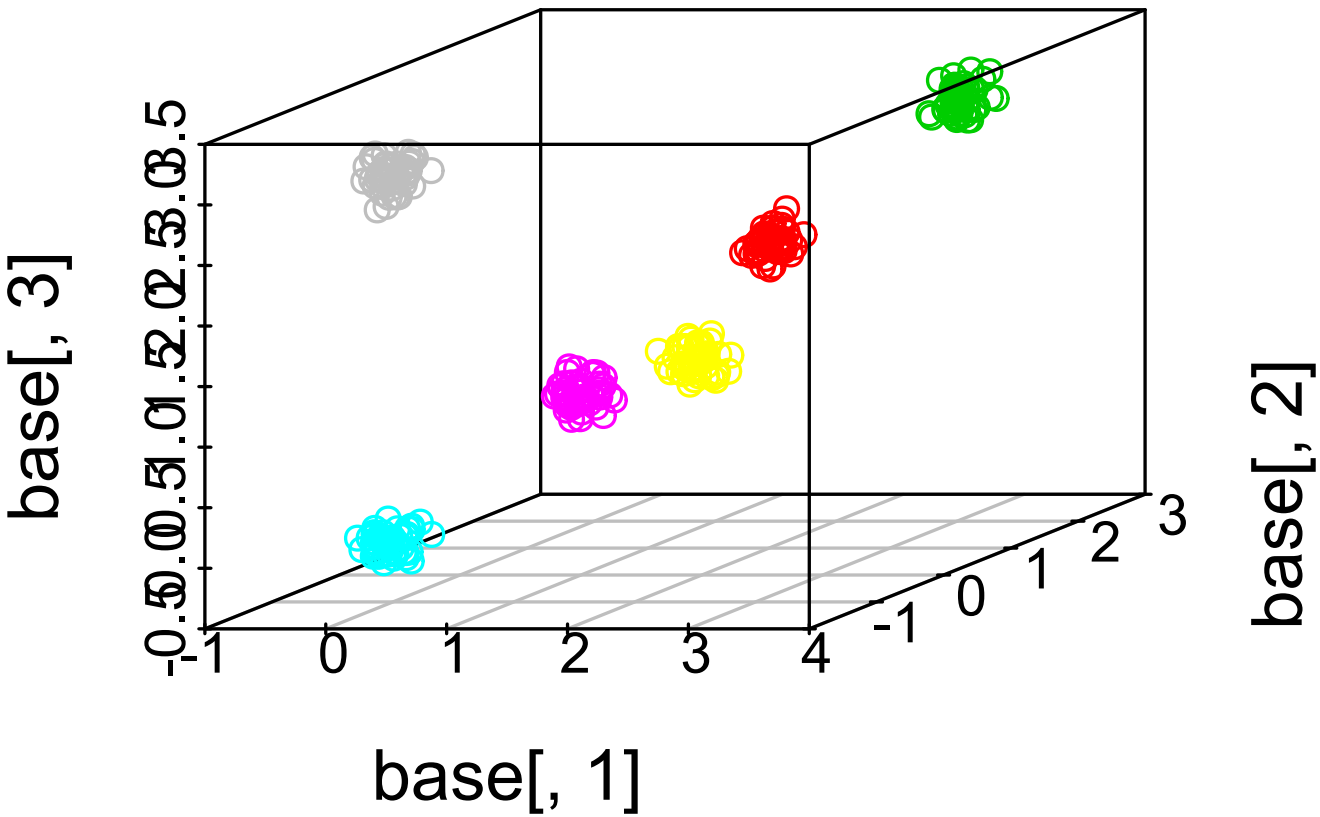
Ejemplos



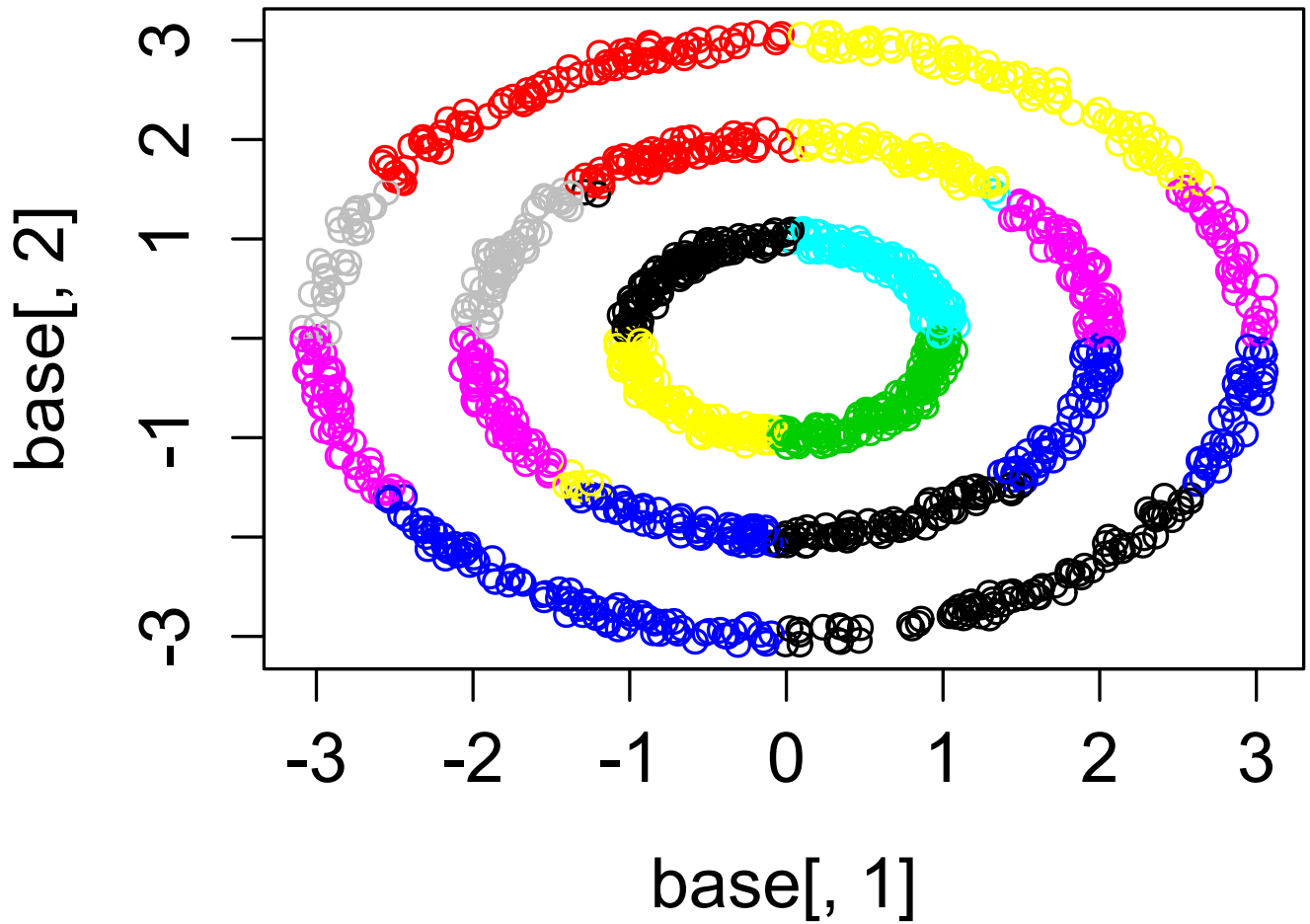
Ejemplos


X2< 1.481
X2>=1.481
X3< 2.235
X3>=2.235
X2< 2.487
X2>=2.487
X1< 1.715
X1>=1.715
X1< 0.5634
X1>=0.5634
1.33 : n=50
1.52 : n=50
1.52 : n=50
1.46 : n=50
1.43 : n=50
1.4 : n=50
Error : 0.00735 CV Error : 0.00768 SE : 0.000339

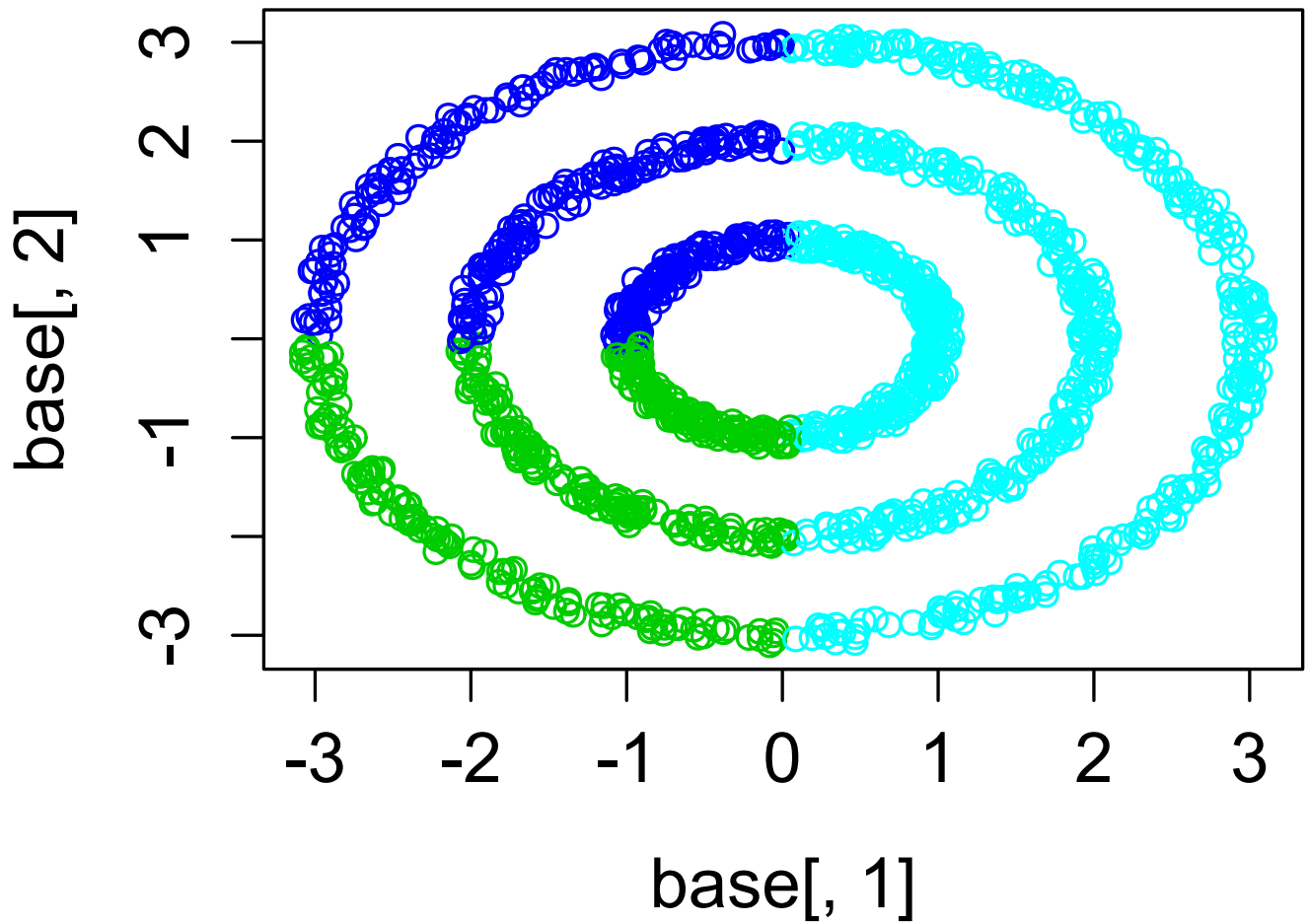
Ejemplos



Ejemplos



Ejemplos



Fin

¿Preguntas?

¡Gracias!