**UNIVERSIDAD DE SANTIAGO DE CHILE**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA INFORMÁTICA**

# INFORME LABORATORIO 2 ANÁLISIS DE DATOS

# TIC-TAC-TOE ENDGAME DATA SET

**CARLOS CÁCERES**
**DANY RUBIANO**

Profesor: Felipe Bello
Ayudante: Bryan Guzmán

Santiago - Chile
24 de septiembre de 2016

# TABLA DE CONTENIDOS

# ÍNDICE DE FIGURAS

# ÍNDICE DE CUADROS

# CAPÍTULO 1. INTRODUCCIÓN

## 1.1 MOTIVACIÓN Y ANTECEDENTES

Durante la experiencia anterior con base en un análisis de estadística descriptiva e inferencial se pudo ahondar en el dominio del problema presentado en el dataset Tic-Tac-Toe. En este análisis, se llegó a esclarecer desde cuales jugadas tienen mayor probabilidad de ganar, hasta cual es la casilla que se ocupa más en el tablero. Ahora es el turno de utilizar técnicas de clustering o algoritmos de agrupamiento, las cuales son procedimiento de agrupación de una serie de vectores de acuerdo con un criterio ya sea de distancia o similitud, de manera que en acorde a los grupos formados, se pueda obtener una descripción sintética de un conjunto de datos multidimensional complejo que ayuden a obtener mayor conocimiento del problema en cuestión.

## 1.2 OBJETIVOS

El objetivo de esta experiencia es utilizar el algoritmo de clustering K-means con el fin de obtener conocimiento del problema que se esta abordando. A partir de ello, se busca inferir las hipótesis pertinentes y realizar el análisis respectivo, de manera que se pueda comparar los resultados obtenidos con los estudios relacionados.

## 1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento distribuye su contenido de la siguiente forma, en primer lugar se encuentra un capitulo dedicado a un pequeño marco teórico en el cual se incluyen las definiciones de los conceptos y técnicas a utilizar en el desarrollo de la experiencia.

A continuación, se realiza un pre-procesamiento de la base de datos abordada con el fin de modelarla, de manera que todos los registros estén representados de forma correcta. Luego, se presenta el resultado de aplicar los criterios seleccionados dada la base de dato para el desarrollo del clustering.

En lo que sigue, se presenta los análisis de los resultados obtenidos en el desarrollo de los capítulos anteriores.

Por último, con lo desarrollado, se realiza una síntesis total, la cual es presentada en las conclusiones del presente documento.

# CAPÍTULO 2. MARCO TEÓRICO

## 2.1 CLUSTERING

Técnica en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico. El clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional y muchas otras.

De forma general, las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases. *(EcuRed, s.f.).*

## 2.2 ALGORITMO K-MEANS

K-means, o k-medias, es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Su descripción es la siguiente, dado un conjunto de observaciones $(x_1, x_2, \ldots, x_n)$, donde cada observación es un vector real de $d$ dimensiones, k-medias construye una partición de las observaciones en $k$ conjuntos $(k \leq n)$ a fin de minimizar la suma de los cuadrados dentro de cada grupo $(WCSS) : S = \{S_1, S_2, \ldots, S_k\}$.

$$arg_s \quad min \quad \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

donde $\mu_i$ es la media de puntos en $S_i$ . *(Wikipedia, s.f.-b).*

## 2.3 DISTANCIAS

La medida fundamental para el agrupamiento, es la similaridad (asociación, proximidad) o la distancia en $R^n$. Una similaridad debe cumplir las condiciones de una distancia (una distancia corresponde a una disimilaridad):

- No-negatividad: $d(x, y) \geq 0$
- La distancia de una instancia (observación) así misma es cero, $d(x, x) = 0$
- Simetría: $d(x, y) = d(y, x)$
- Desigualdad Triangular: $d(x, y) \leq d(x, z) + d(z, y)$

Las medidas de similaridad más conocidas son las de distancia.
Para dos vectores $\vec{x}$ e $\vec{y}\,\epsilon\, R^n$

$$||x, y|| = \sqrt[p]{\sum_{i=0}^{n} |x_i - y_i|^p}$$

*(Chacón, s.f.).*

En la ecuación anterior se puede observar lo que es la distancia euclídea la cual opera sobre variables cuantitativas, pero la naturaleza de las variables presentes en el dataset abordado son del tipo cualitativas. Es por ello que se recurre a una distancia que se acomode a las característica vistas, y esta es la distancia de Hamming.

### 2.3.1 Distancia de Hamming

Se denomina distancia de Hamming a la efectividad de los códigos de bloque y depende de la diferencia entre una palabra de código válida y otra. Cuanto mayor sea esta diferencia, menor es la posibilidad de que un código válido se transforme en otro código válido por una serie de errores. A esta diferencia se le llama distancia de Hamming, y se define como el número de bits que tienen que cambiarse para transformar una palabra de código válida en otra palabra de código válida. Si dos palabras de código difieren en una distancia d, se necesitan d errores para convertir una en la otra.

Por ejemplo:

La distancia Hamming entre 1011101 y 1001001 es 2.

La distancia Hamming entre 2143896 y 2233796 es 3.

La distancia Hamming entre ”tener” y ”reses” es 3.

*(Wikipedia, s.f.-a).*

# CAPÍTULO 3. PRE-PROCESAMIENTO

Para poder aplicar cualquier algoritmo de agrupamiento, es necesario realizar un pre-procesamiento de la base de datos con el fin de eliminar aquellas variables que contengan datos perdidos o que no aporten información relevante para el dominio del problema. Dado el tipo de dataset con el que se está trabajando solo muestra los estados finales de los *tableros* de partidas de tic tac toe, los datos que contiene son netamente cualitativos. Por esta misma razón, como los atributos representan a cada casilla, no existen datos irrelevantes pues solo existen 3 posibles valores *x*, *o* ó *b* para las casillas que queden en blanco; no tiene sentido modificar estos valores pues se estaría modificando un tablero a gusto y no sería representativo. Por lo anterior tampoco se identifican outliers ni se puede aplicar ningún otro tipo de proceso que solo se utilizan cuando los datos son cuantitativos.

Sin embargo para aplicar el algoritmo de las k-medias es necesario tener alguna medida cuantitativa para poder agrupar. Un método para pasar estos datos cualitativos a cuantitativos es calcular la distancia de Gower para cada individuo, pero al ser solo variables cualitativas y no mixtas para cada individuo (tanto cualitativas como cuantitativas) se descarta esta opción. Otro método disponible, es codificar los datos de manera que se representen mediante datos cualitativos asignándole un numero a los posibles valores de *x*, *o* ó *b*, pero esto restringiría el agrupamiento según el símbolo al que se le asigne un mayor valor. Sabiendo esto, y por ser cada elemento un vector de igual tamaño se usará el código de Hamming (vease Marco Teórico - Sección 3). Utilizando R se aplicará el algoritmo de las k-medias sin problemas. Se debe considerar que al ser 9 atributos por elemento, la distancia de Hamming máxima que se podrá encontrar es 9.

# CAPÍTULO 4. OBTENCIÓN DEL CLUSTER

Como el algoritmo de las k-medias necesita de un valor de k predeterminado para generar los grupos. Para un análisis más objetivo, el valor que se da a k en este caso será determinado por una medida externa al algoritmo en sí; esto se logra utilizando el índice de Silueta que arroja la función *pam* de R. Cabe destacar que en esta función se aplica, mediante la modificación de un parámetro, el código de Hamming al dataset. En la figura 4.1 se presenta un gráfico que muestra el índice de silueta para distintos valores de k que varía entre el 2 y el 10. En el gráfico se aprecia que el número de clusters óptimo dado el intervalo escogido para evaluar (entre 2 y 10 grupos) es de $k = 5$.

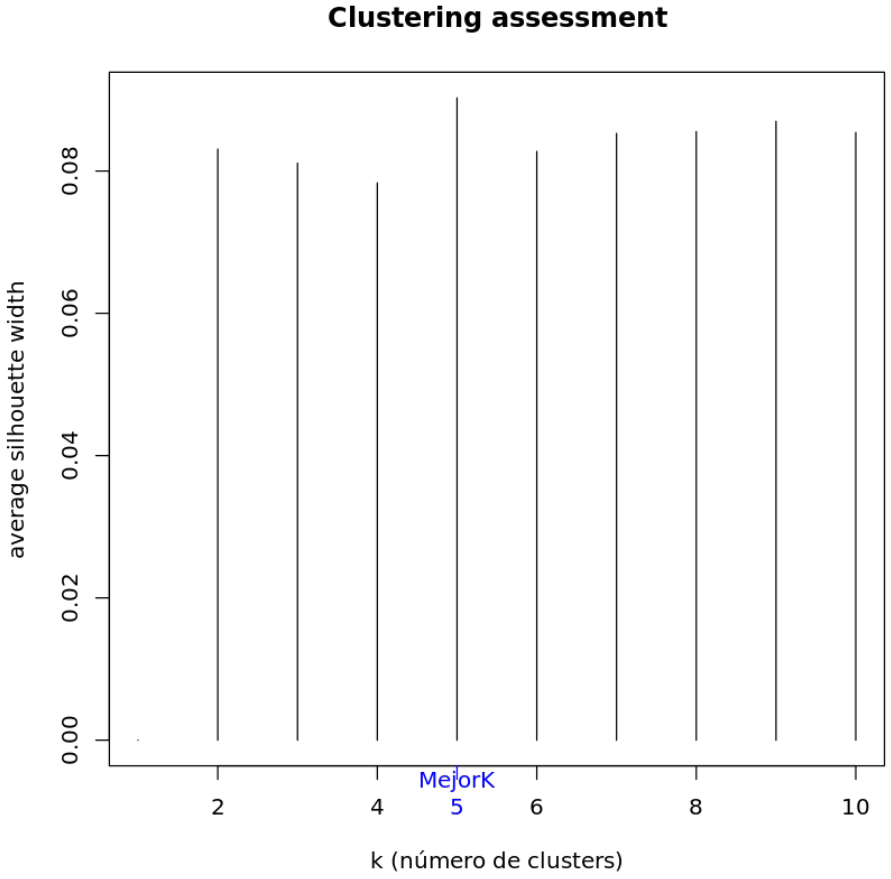


Figura 4.1: Mejor valor de k

Sabiendo el número de clusters que necesitamos visualizar se procede a graficar lo arrojado por el algoritmo de las k-medias con un valor de k igual a 5 (nuevamente con la función *pam* y dando como parámetro el dataset original más la modificación de la métrica para que utilice Hamming). En la Figura 4.2 se presentan los 5 clusters arrojados por R; como se pensó en un comienzo, como la disimilitud máxima que se puede alcanzar entre elementos es de 9, se puede encontrar una suerte de 9 lineas verticales representando cada disimilitud y los grupos, en general, compuestos por más de una de estas.

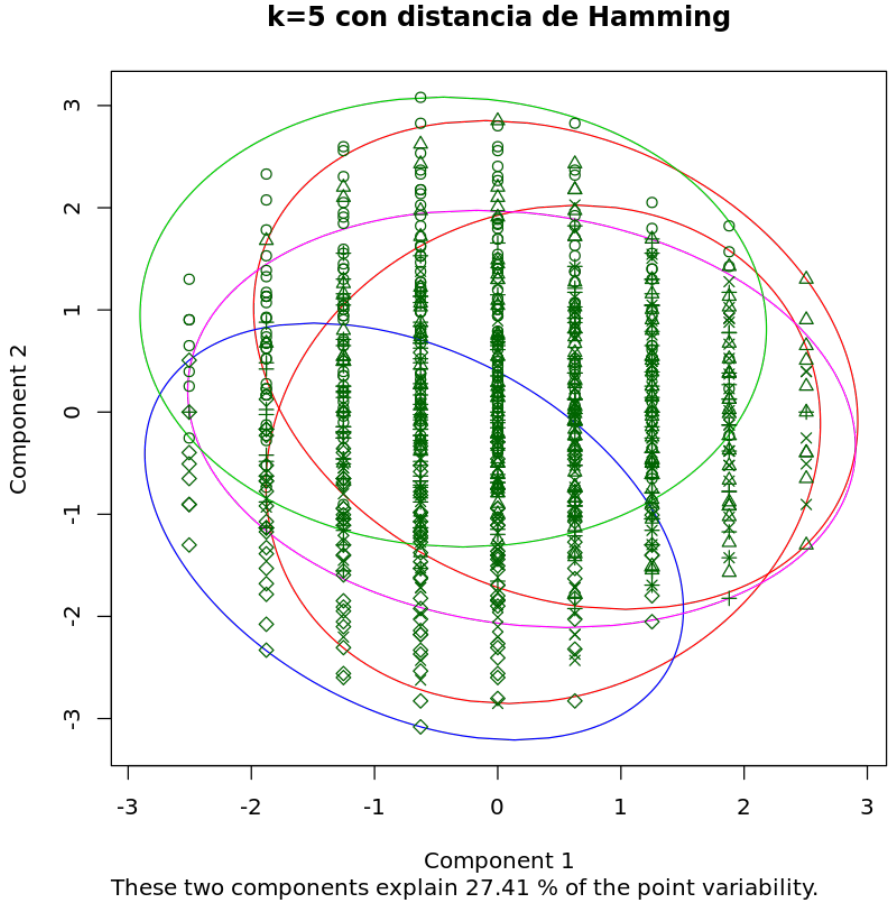


Figura 4.2: Clúster con k=5

Dado el agrupamiento obtenido, a continuación en la figura 4.3 se presenta la distribución de los elementos en los 5 grupos generados por el algoritmo de las k-medias al utilizar la función *pam*.

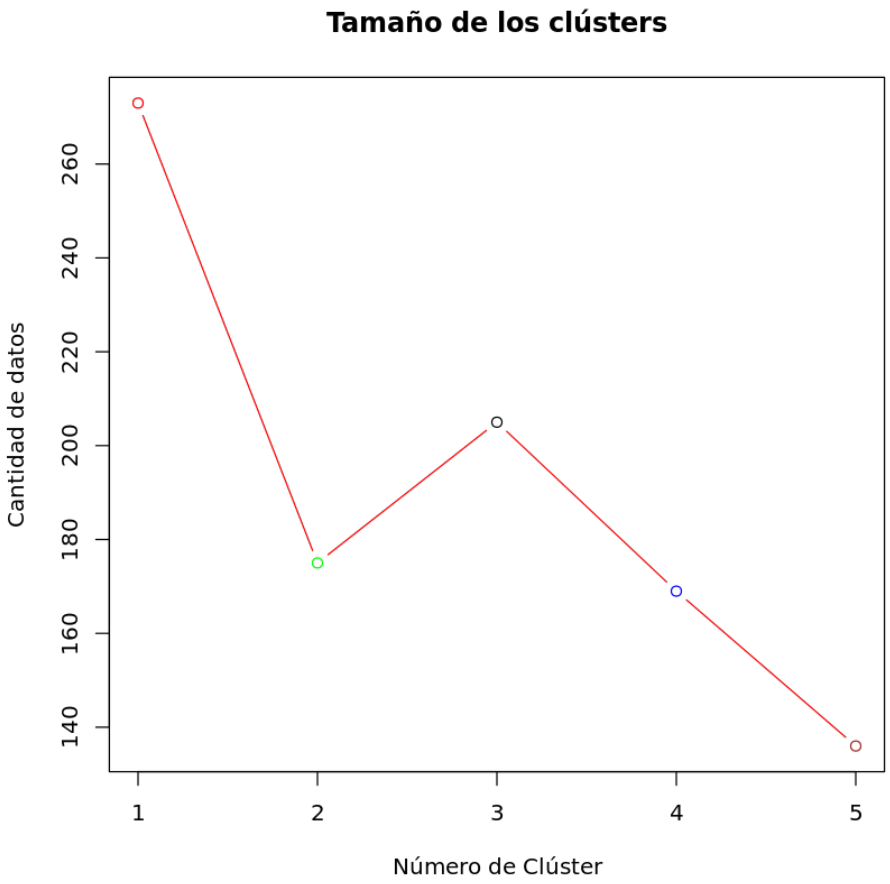


Figura 4.3: Distribución de datos por clúster

Así mismo, en la tabla 4.1 se puede observar la información tanto de la disimilitud máxima, la disimilitud media, el diámetro, la separación y el ancho medio de cada clúster.

Cuadro 4.1: Información de cada clúster

| **Clúster** | **size** | **max_diss** | **av_diss** | **diameter** | **separation** | **avg_width** |
|---|---|---|---|---|---|---|
| 1 | 273 | 3.316.625 | 2.272.883 | 4.898.979 | 1.414.214 | 0.06060563 |
| 2 | 175 | 3.316.625 | 2.319.935 | 4.795.832 | 1.414.214 | 0.08732214 |
| 3 | 205 | 3.316.625 | 2.231.297 | 4.472.136 | 1.414.214 | 0.10389148 |
| 4 | 169 | 3.162.278 | 2.279.002 | 4.582.576 | 1.414.214 | 0.09859000 |
| 5 | 136 | 3.162.278 | 2.182.342 | 4.582.576 | 1.414.214 | 0.12282680 |

# CAPÍTULO 5. ANÁLISIS DE RESULTADOS

Para comenzar es necesario dejar en claro que por la naturaleza del dataset escogido, si bien el procesamiento fue simplificado gracias a R, la información relevante que se puede obtener respecto a este, utilizando el algoritmo de las k-medias, deja con gusto a poco. Teniendo en consideración que cada elemento del dataset corresponde a un tablero de tic tac toe en su estado final y que cada variable corresponde a cada una de las 9 casillas contenidas en este, al utilizar el código de hamming se pasan por alto cosas; por ejemplo en un caso hipotético, un tablero vacio en comparación con uno lleno tendrá una medida de disimilitud de 9, mismo valor que se obtendría en la comparación entre 2 tableros llenos en donde uno tuviera *x* donde el otro tenga *o* y *o* donde el otro tenga *x*.

Debido a la configuración del dataset, en donde se presentan dos clases (”negative” y ”positive”) que indican los registros en los que gana o pierde el símbolo *x*, se pensó en algún momento que en la aplicación del algoritmo k-means el número de grupos recomendado fuera 2, tal como el número de clases. La idea anterior, fue descartada en primer lugar por las incongruencias que presentaba el agrupamiento con la inclusión de la variable de clase y también con la ayuda de el método de las siluetas, que permitió encontrar el número de grupos óptimo, representado por el tamaño mayor de las siluetas en la relación entre el tamaño de la muestra y el número de grupos.

Teniendo lo anterior en cuenta, en el gráfico de agrupamientos con $k = 5$ (ver figura 4.2), en donde se muestran las 9 lineas verticales (que representan cada una de las disimilitudes entre los elementos), se puede observar con claridad que todos los clusters contienen elementos de otro (esto puede deberse a lo señalado en el párrafo anterior). Esto no debería darse en ningún caso, pero como se explicó, el dataset analizado no cuenta con características necesarias para hacer un estudio utilizando el algoritmo requerido por lo que se espera extraer información realmente relevante en próximas experiencias.
Con respecto a la información que se detalla en la tabla 4.1 se puede observar que la máxima disimilitud es muy parecida entre los diferentes clusters presentes, lo mismo ocurre

con la disimilitud promedio, el diámetro, y la separación o aislamiento de los datos por cada clúster, todo esto se observa en el gráfico del agrupamiento obtenido. Lamentablemente debido a las características de las variables, las cuales son todas de tipo cualitativas y a su representación de tres símbolos solamente, no se puede tener un análisis más profundo, ni obtener mayor conocimiento que permita encontrar relaciones entre las variables descriptivas.

# CAPÍTULO 6. CONCLUSIONES

El principal objetivo de esta etapa del estudio del dataset tenía como fin obtener conocimiento del problema en sí utilizando el algoritmo de las k-medias; ya finalizando la experiencia se puede decir que este objetivo no fue posible lograrlo por distintas razones:

1. La naturaleza del dataset.
2. El tipo de dato con el que se trabaja.
3. Funcionamiento del algoritmo k-medias.

[1] La naturaleza del dataset, como se ha dicho ya reiteradas veces, es muy particular pues cada elemento representa el estado final de un tablero de tic-tac-toe y solo eso. No se cuenta con información adicional como podría ser el orden en el que se jugaron las casillas, dato que podría ser de mucha importancia para un análisis más detallado.

[2] Adicional a esto se tiene que los datos contenidos en el dataset son todos de tipo char; y no solo eso sino que para cada variable existían los mismos posibles valores (*x*, *o* o *b*) lo que limita a cualquier investigador pues si se quiere hacer una comparación general de los distintos elementos del problema se encontraría con que no es posible analizarlo aplicando cualquier método pues algunos están enfocados solo en variables numéricas.

[3] Siguiendo con la idea del punto 2, uno de esos algoritmos es justamente el requerido en esta experiencia: k-means. Lo que considerando el punto 1, dificulta enormemente su uso para un posterior análisis.

Sin embargo estos factores no impidieron la creación de los clusters. Utilizando la función *pam* de R utilizando el código de Hamming se logró este objetivo; pero el agrupamiento obtenido no aporta conocimiento y esto puede deberse a lo señalado en el capítulo de análisis y es que al aplicar este código y medir las disimilitudes entre elementos se dejan de lado aspectos importantes en una partida de tic-tac-toe, como la ubicación de los elementos en el tablero pues al aplicar Hamming, cada elemento se estaría transformando solo en un número, es decir, se pasa de 9 atributos a un valor numérico que puede ser el mismo en muchos casos, pero los tableros totalmente distintos. Pero de todos modos, el no utilizar esta transformación hubiera impedido concluir bien la experiencia.

# CAPÍTULO 7. REFERENCIAS


Aha, D. W. (1991, Agosto). Index of /ml/machine-learning-databases/tic-tac-toe. https://archive.ics.uci.edu/ml/machine-learning-databases/tic-tac-toe/.

Chacón, M. (s.f.). Análisis de agrupamientos. http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=\%2F115775\%2Fmod_resource\%2Fcontent\%2F1\%2FCapitulo\%20III\%20An\%C3\%A1lisis\%20de\%20Datos_AA\%202016.pdf.

EcuRed. (s.f.). Clustering. https://www.ecured.cu/Clustering.

Maechler, M. (2016, Abril). Package 'cluster'. https://cran.r-project.org/web/packages/cluster/cluster.pdf.

Stackoverflow, U. (2013, marzo). Cluster analysis in r: determine the optimal number of clusters. http://stackoverflow.com/questions/15376075/cluster-analysis-in-r-determine-the-optimal-number-of-clusters.

UCI, M. L. R. (s.f.). Tic-tac-toe endgame data set. https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

Wikipedia. (s.f.-a). Distancia de hamming. https://es.wikipedia.org/wiki/Distancia_de_Hamming.

Wikipedia. (s.f.-b). K-means. https://es.wikipedia.org/wiki/K-means.