**APRENDIZAJE NO SUPERVISADO**
MÁSTER EN INTELIGENCIA ARTIFICIAL
UNIVERSIDAD INTERNACIONAL DE VALENCIA

# TÉCNICAS DE CLUSTERING


Componentes del grupo:

**Coordinadora:** Carmen Raposo Jiménez
**Secretario:** Víctor Dot Mariano
**Revisor:** Ricardo Sánchez Pastor


7 Abril de 2019

# Índice general

# Apartado 1

# Conjuntos de Datos

Se han escogido dos conjuntos de datos distintos para las dos casuísticas planteadas: uno conociendo la clasificación real de los datos y otro sin saberlo.

## 1.1. Agrupamiento Real Conocido

Como dataset para el problema cuyo agrupamiento real es conocido se ha seleccionado el ya cargado en las librerías de sklearn, iris, el cual categoriza distintos tipos de iris en tres especies. Este conjunto de datos tiene las siguiente características:

| | |
|---|---|
| **Número de instancias** | 150 |
| **Número de atributos** | 4 |
| **Descripción de atributos** | sepal-length. Numérico. Longitud del sépalo.<br>sepal-width. Numérico. Anchura del sépalo.<br>petal-length. Numérico. Longitud del pétalo.<br>petal-width. Numérico. Anchura del pétalo. |
| **Tipo de datos de la clase** | Categórico: iris-setosa, iris-versicolor, iris-virginica |
| **Descripción de la clase** | Nombre: class. Descripción: Especie de la planta. |
| **Valores ausentes** | 0 |

De los cuatro atributos que posee este dataset, nuestro problema se centrará en los tres primeros, por lo que trabajaremos en un espacio de tres dimensiones para nuestros algoritmos.

## 1.2. Agrupamiento Real No Conocido

# Apartado 2

# Métricas de evaluación

Teniendo en cuenta la naturaleza del trabajo, necesitaremos definir dos conjuntos de métricas, unas para el problema con datos de agrupación real conocidos y otras para los no conocidos. De esta forma, las métricas quedan divididas en: **extrínsecas** e **intrínsecas**.

## 2.1. Medidas de evaluación extrínsecas

Las medidas escogidas para evaluar los algoritmos de agrupación de datos reales conocidos han sido tres:

- **Error** : Estudia el número media de casos de errores que se han cometido, representándose por la siguiente fórmula:

$$E = 1 - \frac{1}{n} \max_{\sigma} \sum_{l=1}^{K'} n_{\sigma(l)l}$$

  donde la función sigma asigna a cada clúster original un clúster predicho.

- **Pureza** : Se entiende como la precisión promedio, donde se calcula la precisión media de los clústeres predichos. Para ello, presentemos la siguiente fórmula:

$$Pu = \sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \max_{l \in \{1,\dots,K'\}} P_{kl}$$

  donde $P_{kl}$ representa la precisión, es decir, la proporción de casos sobre un clúster predicho que pertenecen a la mismo clúster que al predicho.

- **F1** : A diferencia de las medidas anteriores esta representa el promedio ponderado de la media armónica de la precisión y el recall, siendo esta última medida equivalente a la precisión pero sobre un clúster real. Su expresión es la siguiente:

$$F1 = \sum_{l=1}^{K'} \frac{n_{\cdot l}}{n} \max_{k \in \{1,\ldots,K\}} \left( \frac{2P_{kl}R_{lk}}{P_{kl} + R_{lk}} \right)$$

## 2.2. Medidas de evaluación intrínsecas

Para este caso se han escogido otras tres métricas, a diferencia de que estas pueden ser aplicadas para ambos problemas. Son las siguientes:

- **RMSSTD** : Es la raíz del cuadrado de la media de la desviación típica, de ahí las siglas en inglés RMSSTD, y su expresión matemática es:

$$RMSSTD = \sqrt{\frac{\sum_{k=1}^{K} \sum_{x_i \in C_k} x_i - c_{k^2}}{v \cdot \sum_{k=1}^{K} (|C_k| - 1)}}$$

  Estudia la homogeneidad de los clústeres del agrupamiento sin tener en cuenta la distancia interclúster.

- **Ancho de silueta** : Esta métrica supone una diferencia normalizada de la distancia interclúster menos la distancia intraclúster, siendo su fórmula:

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x_i \in C_k} \frac{b_k(x_i) - a_k(x_i)}{\max \{a_k(x_i), b_k(x_i)\}}$$

  donde

$$a_k(x_i) = \frac{1}{|C_k| - 1} \sum_{x_{i'} \in C_k : x_{i'} \neq x_i} d(x_i, x_{i'}) \qquad b_k(x_i) = \min_{h \neq k} \frac{1}{|C_h|} \sum_{x_{i'} \in C_h} d(x_i, x_{i'})$$

- **Índice Calinski-Harabasz** : Evalúa la bondad de un agrupamiento basado en la suma promedio de distancias interclúster e intraclúster al cuadrado, más concretamente de la siguiente forma:

$$iCH = \frac{(n - K) \sum_{k=1}^{K} |C_k| \cdot d(c_k, c)^2}{(K - 1) \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, c_k)^2}$$

# Apartado 3

# Algoritmos de clustering

En el ámbito del clustering, existen infinidad de técnicas y algoritmos de agrupamiento, pudiéndose hacer una amplia clasificación de los mismos en cinco grupos:

- Algoritmos de agrupamiento basados en particiones
- Agrupamiento jerárquico
- Agrupamiento espectral
- Agrupamiento basado en densidad
- Agrupamiento basado en modelos probabilísticos

En este trabajo se aplican cinco algoritmos de agrupación en total, perteneciendo cada uno a un grupo de los antes mencionados.

## 3.1. KNN

## 3.2. Aproximación aglomerativa

## 3.3. Espectral

## 3.4. DBSCAN

## 3.5. Modelos de mixturas

# Apartado 4

# Pruebas realizadas

Se ha llevado a cabo sobre ambos problemas la iteración de los cinco algoritmos mencionados anteriormente, estableciendo dos partes:

- ¿Cuál es la mejor configuración para cada algoritmo?
- Partiendo de la mejor configuración de cada algoritmo, ¿cuál es el mejor algoritmo para cada problema?

Sabiendo la estructura de las pruebas realizadas, podemos pasar a la exposición de los resultados y conclusiones obtenidos.

## 4.1. Dataset Iris

El Dataset Iris se corresponde con el problema cuyos datos de agrupamiento real son conocidos. A continuación, su representación visual:

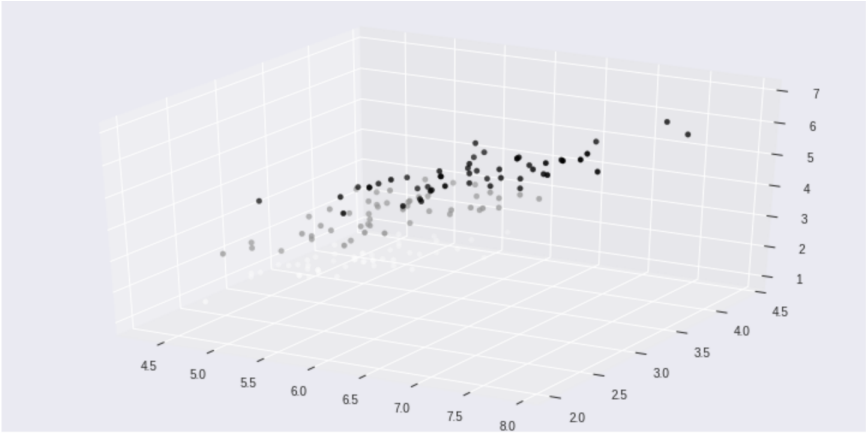


Figura 4.1: Dataset Iris

Como se ha expuesto antes, lo primero es conocer cuál es la mejor configuración posible entre todas las escogidas para cada algoritmo, por lo que veamos uno por uno los resultados obtenidos.

### 4.1.1. KNN

Este algoritmo sólo depende de un parámetro, $K$, por lo que las distintas configuraciones se basan en distintos valores para este parámetro.

Los distintos valores escogidos han sido $\{2, 3, 4, 5\}$ centros, siendo la mejor elección del valor para cada métrica la mostrada en la imagen 4.2.

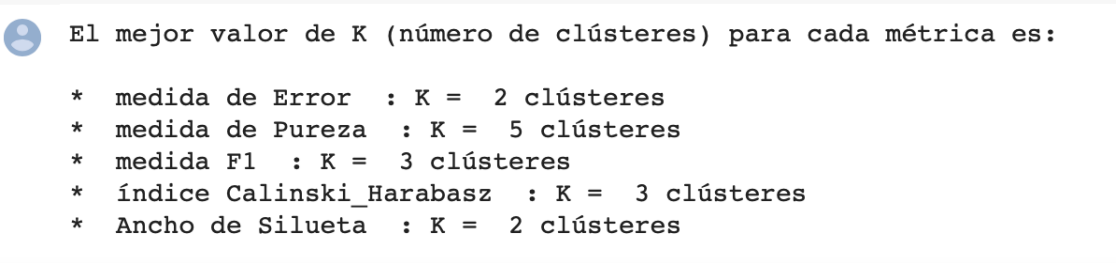


Figura 4.2: Mejor valor del parámetro $K$ para cada medida de evaluación

Como se puede apreciar, el mejor valor del parámetro se halla entre 2 y 3 clústeres, pero optaremos por establecer como mejor configuración para este algoritmos el valor de 3 clústeres ya que es lo más aproximado a la agrupación real del dataset, como queda reflejado en la figura 4.1.

Los valores de las métricas para esta elección son:

- **Error**: 0.12
- **Pureza**: 0.8799999999999999
- **F1**: 0.8792270531400966
- **Ancho de silueta**: 0.5498955810221876
- **Índice Calinski-Harabasz**: 556.1703613698259

y un ejemplo de solución que se obtiene es el siguiente:

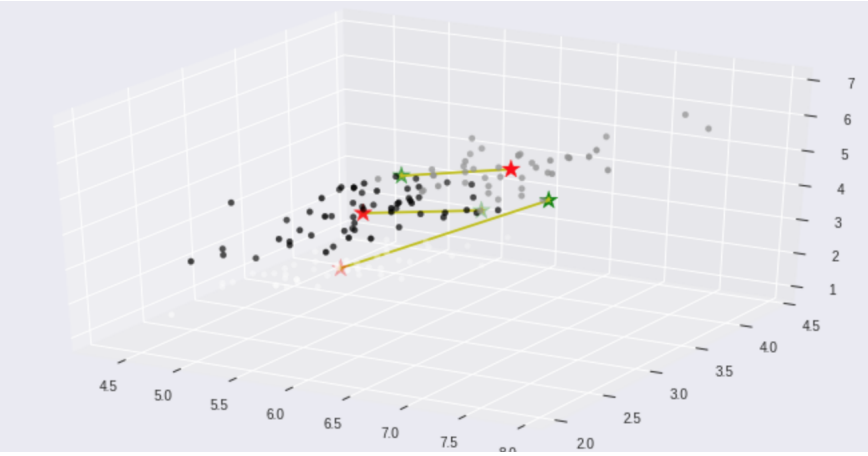


Figura 4.3: Resolución del algoritmo para 3 clústeres

### 4.1.2. Aproximación aglomerativa

### 4.1.3. Espectral

### 4.1.4. DBSCAN

### 4.1.5. Modelos de mixturas

Por último el algoritmo de Mixtura Gaussiana donde, al igual que en el algoritmo KNN, sólo variamos el valor de un parámetro, $K$, en función del número de clústeres que deseemos predecir.

Los valores escogidos para este parámetro son $\{2, 3, 4, 5\}$ centros, y el mejor valor de dicho parámetro para cada medida de evaluación es:

```
El mejor valor de K (número de clústeres) para cada métrica es:

*  medida de Error  : K =  2 clústeres
*  medida de Pureza  : K =  3 clústeres
*  medida F1  : K =  3 clústeres
*  índice Calinski_Harabasz  : K =  2 clústeres
*  Ancho de Silueta  : K =  2 clústeres
```

Figura 4.4: Mejor valor del parámetro $K$ para cada medida de evaluación

Para el algoritmo de mixturas Gaussianas parece evidente que la mejor configuración es la de 2 clústeres siendo los valores de las métricas para estas características las siguientes:

- **Error**: 0.0

- **Pureza**: 0.6666666666666666
- **F1**: 0.7777777777777778
- **Ancho de silueta**: 0.6885194944858716
- **Índice Calinski-Harabasz**: 496.72125954418374

Siendo un ejemplo de solución para el algoritmo de mixturas Gaussianas con dos clústeres el mostrado a continuación:

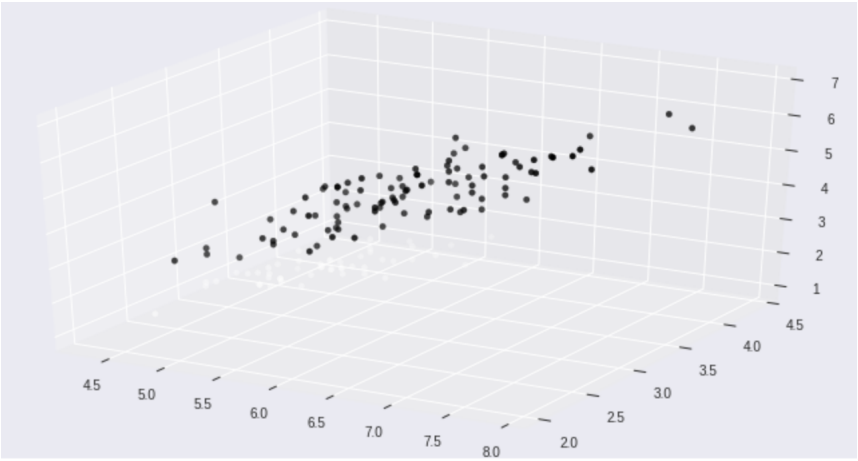


Figura 4.5: Resolución del algoritmos para 2 clústeres

### 4.1.6. Mejor algoritmo

## 4.2. Dataset Wine

### 4.2.1. KNN

### 4.2.2. Aproximación aglomerativa

### 4.2.3. Espectral

### 4.2.4. DBSCAN

### 4.2.5. Modelos de mixturas

### 4.2.6. Mejor algoritmo