Pre-TP1: Data mining en Música: Preparación de los datos

Pablo Riera, Juan E Kamienkowski
Data Mining en Ciencia y Tecnología

3 de septiembre de 2019

1. Dataset *metadata* / *audio_features*

Utilizando la API de Spotify se descargó la información de 2206 canciones. Cada registro tiene 17 variables, las cuales en su mayoría no van a ser útiles para el análisis. Entonces, como primer paso, deberán:

- Separar en distintas tablas, las variables continuas que se utilizarán para el clustering, de las etiquetas que se podrán utilizar para la validación externa (Músico, Album y Género), de los códigos que no se utilizarán en este TP.
- Generar un gráfico tipo "*scatter matrix*".
- Identificar variables más o menos informativas *a priori* y variables que requieran, además de la estandarización, alguna corrección para asimilar la distribución a una normal.
- Estandarizar y volver a generar un gráfico tipo "*scatter matrix*".
- Identificar, si es que hay, valores extremos que sea necesario descartar.

El dataset *metadata* contiene las etiquetas casi depuradas.
El dataset *audio_features* contiene las variables continuas de alto nivel ya separadas.
Ejemplos de este procedimiento y una guía para abrir los datos pueden encotrarse en `https://github.com/pabloriera/dmcyt_tp1`.

2. Dataset *audio_analysis*

El dataset *audio_analysis* contiene las variables continuas de bajo nivel, estimadas en ventanas temporales, como 'timbre' o '*pitch*'. Al tener canciones de distintas duraciones (distinta cantidad de ventanas), estas variables se guardan por separado y en `https://github.com/pabloriera/dmcyt_tp1` se encuentran algunas indicaciones de cómo cargalas. Entonces, como primer paso, deberán:

- Resumir estas variables en valores por canción. Por ejemplo, tomar el promedio o el desvío estándar del timbre entre todas las ventanas, obteniendo 12 valores de timbre promedio y 12 valores de desvío estándar del timbre por canción.
- Contruir un *data frame* con estos valores.
- Generar un gráfico tipo "*scatter matrix*".
- Identificar variables más o menos informativas *a priori* y variables que requieran, además de la estandarización, alguna corrección para asimilar la distribución a una normal.
- Estandarizar y volver a generar un gráfico tipo "*scatter matrix*".
- Identificar, si es que hay, valores extremos que sea necesario descartar.