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. Introducción

Utilizando la API de Spotify se descargó la información de 2206 pistas de audio. Cada registro tiene variables de features de alto (*audio_features*), bajo nivel (*audio_analysis*) y de metadata, las cuales en su mayoría no van a ser útiles para el análisis.

Ejemplos de los procedimiento para abrir los datos pueden encontrarse en `https://github.com/pabloriera/dmcyt_tp1/blob/master/pre_TP1.ipynb`

2. Dataset *metadata*

- Con los datos de *metadata*, separar las etiquetas que se podrán utilizar para la validación externa (Artista, Álbum, Año, Género), de los campos que no se utilizarán en este TP.

3. Dataset *audio_features*

El dataset *audio_features* contiene 9 atributos globales de alto nivel para cada pista de audio. Ejemplos para realizar los pasos pueden encontrarse en: `https://github.com/pabloriera/dmcyt_tp1/blob/master/Introduccion.ipynb`.

- Con los datos de *audio_features*, 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.

4. Dataset *audio_analysis*

El dataset *audio_analysis* contiene las variables continuas de bajo nivel, estimadas en ventanas temporales, como *timbre* o *pitches*. Al tener canciones de distintas duraciones (distinta cantidad de ventanas), estas variables se guardan por separado. 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.