Aplicación de modelos de regresión para predecir el año de lanzamiento de una canción a partir de las características del audio

Aguayo M, Ester, Castillo C, Román
CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS AC
Estudiantes de la maestría en cómputo estadístico
Unidad Monterrey

***Abstract*—Se presenta el trabajo como reporte final de la materia 'Cómputo Estadístico', se explora la eficacia de métodos lineales y lineales generalizados para la predicción del año de lanzamiento de una canción. Tambien se presenta la comparación con otros métodos de clasificación**

I. Introducción

Los métodos de predicción consiste una herramienta fundamental en esta época de la abundancia de datos. Muchos negocios, día con día producen enormes cantidades de nueva información, por ejemplo en la industría musical, horas de nueva música. Una tarea de interés, es poder determinar el año de lanzamiento. Como parte de los recursos para la solución de esta tarea, se cuenta con el *Million Song Dataset* (MSD) que es un *data* de acceso gratuito de características de audio y metadatos de un millón de piezas musicales contemporaneas. Entre los metadatos disponibles se encuetra el año de lanzamiento de cada pieza.

A. Objetivos

Usando el **MSD** se plantean los siguientes objetivos

1. Considerando al año de lanzamiento como variable númerica continua, ajustar un modelo de regresión, usando como predictores características del audio
2. Ajustar un clasificador basado en regresión multiclase para predecir el año de lanzamiento de una canción
3. Comparar los resultados con otros clasificadores tales como SVM y NN

II. Métodos y resultados

A. Análisis exploratorio

YearPredictionMSD es un *subset* del **MSD**, destinado a la tarea de predicción del año de lanzamiento de un track. Esta formado por 515345 registros de canciones correspondientes a 90 características tomadas de cada composición. Las características fueron obtenidas del API de **Echo Nest**. Donde cada composición fue divida en doce segmentos. De los cuales se recuperan los promedios y covarianzas de todos ellos.

Con respecto a la cantidad de registros por año de lanzamiento, el conjunto de datos se encuetra altamente desbalanceado. Más del 80% correponde a canciones a partir de la primera decada del año 2000.

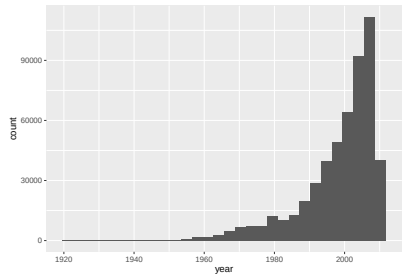


Figura 1. Distribución de las canciones por año

B. Regresión múltiple

1) Regresión ordinaria (mejor subconjunto):

- Forward stepwise selection

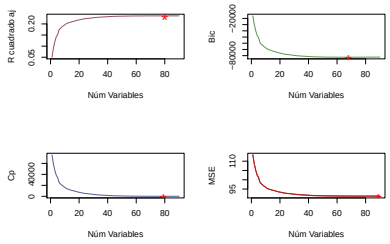


Figura 2. Forward stepwise selection

- Forward stepwise selection

2) Regresión con contracción Lasso:

3) Regresión con componentes principales:

C. Enfoque problema de la clasificación

Un mejor enfoque para el problema, es considerarlo como un problema de clasificación multiclase, por tanto probare-

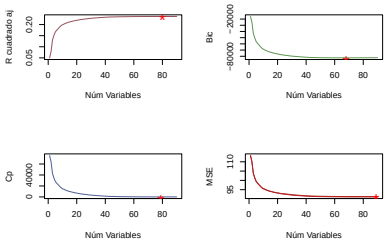


Figura 3. Backward stepwise selection

mos ahora con algunos clasificadores y tambien comparar desempeños

1) Downsampling: Nuestro dataset tiene algunas características que pueden dificultar la tarea. Primero se encuentra altamente desbalanceado, es decir pocas categorias (años de lanzamiento) agrupan un volumen imporante de los datos, lo cual estorba el aprendizaje para categorías de menor volumen. Por otra parte existen muchas categorias, es decir muchos años por predecir, esto aumenta el costo computacional al ajustar un modelo adecuado. Así se realizan los siguientes ajustes en el dataset

1. Agrupamos las piezas musicales por década del lanzamiento
2. Seleccionamos aleatoriamente una muestra de 1000 ejemplos por década.

La gráfica muestra la frecuencia por clases en el conjunto balanceado.

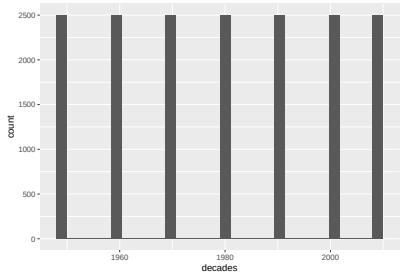


Figura 4. Dataset balanceado

Como parte de la exploración del conjunto de datos, tratamos de identificar alguna variable que pueda ser buen discriminante de las clases. En los gráficos de muestra la variabilidad de la media de los tracks agrupados por décadas tanto en las variables `TimbreAvg`y `TimbreCov`. Se muestra la dispersión

Los gráficos muestran un fuerte sobreempalme de las distribuciones de los datos sobre las variables de interés incluso en las que tienen una 'mayor variabilidad', tal y como se observa:

2) Regresión multinomial con Lasso: hjklñ

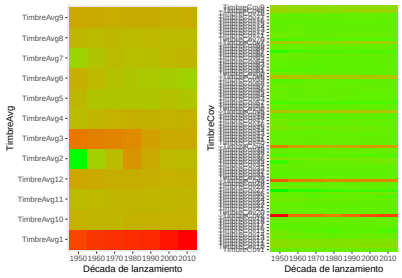


Figura 5. Dispersión de los décadas sobre TimbreAvg y TimbreCov

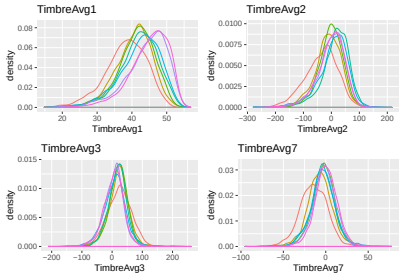
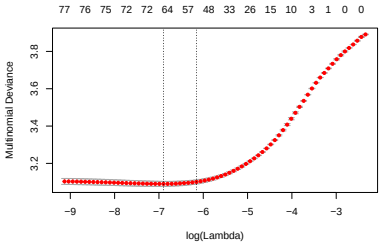


Figura 6. Distribución de los datos

Tabla I
MATRIZ DE CONFUSION MODELO MULTINOMIAL

	1950	1960	1970	1980	1990	2000	2010
1950	675	171	53	50	18	24	14
1960	225	384	157	72	46	60	26
1970	102	164	395	177	65	63	48
1980	91	91	146	436	89	65	58
1990	99	97	91	202	213	148	147
2000	56	48	53	55	104	290	375
2010	34	50	62	55	75	249	532



- **Predicción**

3) Regresión multinomial con componentes principales:
4) Otros métodos de clasificación:

III. CONCLUSIONES