

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. También se presenta la comparación con otros métodos de clasificación

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

I. INTRODUCCIÓN

Los métodos de predicción consisten en 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 industria 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 contemporáneas. Entre los metadatos disponibles se encuentra 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 numérica 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 dividida en doce segmentos. De los cuales se recuperan los promedios y covarianzas de todos ellos.

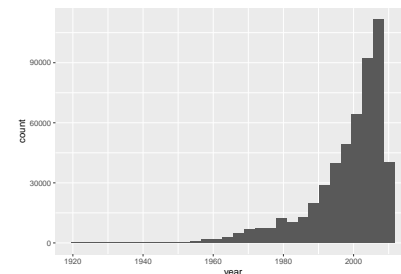


Figura 1. Distribución de las canciones por año

B. Regresión múltiple

1) Regresión ordinaria (mejor subconjunto):

- Forward stepwise selection

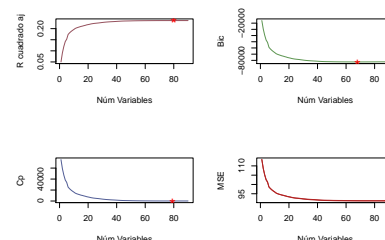


Figura 2. Forward stepwise selection

- Forward stepwise selection

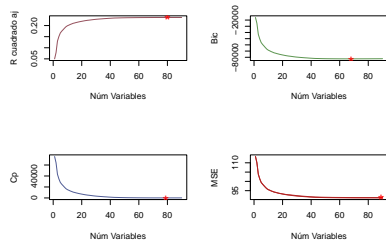


Figura 3. Backward stepwise selection

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 probaremos ahora con algunos clasificadores y también comparar desempeños

1) *Downsampling:* Nuestro dataset tiene algunas características que pueden dificultar la tarea. Primero se encuentra altamente desbalanceado, es decir pocas categorías (años de lanzamiento) agrupan un volumen importante de los datos, lo cual estorba el aprendizaje para categorías de menor volumen. Por otra parte existen muchas categorías, 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.

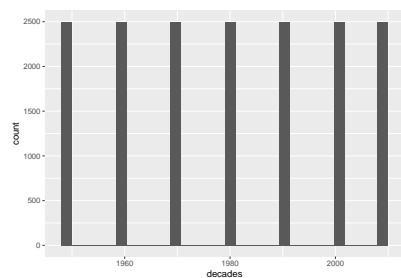


Figura 4. Dataset balanceado

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

Tabla I
MATRIZ DE CONFUSIÓN MODELO MULTINOMIAL

	1950	1960	1970	1980	1990	2000	2010
1950	668	175	54	51	19	24	14
1960	225	381	158	74	45	60	27
1970	103	163	396	177	63	66	46
1980	91	91	148	435	90	64	57
1990	99	97	93	200	216	148	144
2000	56	47	55	54	106	294	369
2010	33	50	61	55	79	248	531

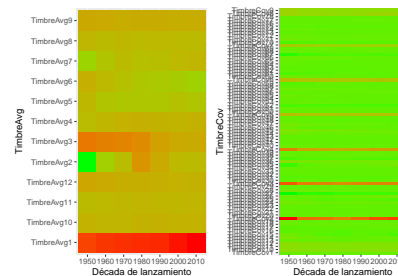


Figura 5. Dispersión de los décadas sobre *TimbreAvg* y *TimbreCov*

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:

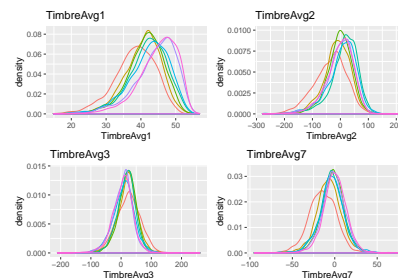


Figura 6. Distribución de los datos

2) *Regresión multinomial con Lasso:* Usamos un modelo multinomial con restricción Lasso para disminuir la cantidad de variables

En vista, que el dataset balanceado tiene 7 clases, ajustamos un modelo de regresión multiclase, (glm de familia multinomial). Se *cross-validation* con 10 folds y se ajusta λ que minimice el error de clasificación.

• Predicción

Se presenta la matriz de confusión del modelo con el conjunto de prueba, se tiene un *accuracy* de 0.41

3) *Regresión multinomial con componentes principales:* De nuevo para obtener un modelo reducido, probamos un modelo basado en componentes principales. Para escoger la cantidad de componentes evaluamos el *accuracy* de modelos con diferente cantidad de componentes

• Predicción

Tabla II
MATRIZ DE CONFUSIÓN PCR

	1950	1960	1970	1980	1990	2000	2010
1950	770	62	52	75	9	11	26
1960	435	149	159	118	28	30	51
1970	274	64	353	174	35	25	89
1980	185	41	172	415	45	35	83
1990	207	44	119	245	121	102	159
2000	172	37	123	160	67	143	279
2010	163	43	136	143	68	100	404

Se muestra en el gráfico, el cambio de *accuracy* con el número de componentes y la matriz de confusión. El *accuracy* es de 0.34

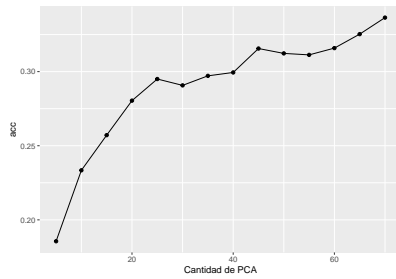


Figura 7. Cantidad de PCA y acc

4) *Otros métodos de clasificación:*

III. CONCLUSIONES