Segundo bimestre de 2022 TP 3: Modelos mixtos y splines penalizados

Herramientas de Modelado Estadístico

Reglas del TP:

- Este trabajo debe hacerse de forma individual o de a dos.
- Deben enviarse el script de R y un informe que contenga las respuestas a todas las preguntas y los gráficos pedidos, en lo posible en RMarkdown. No hace falta explicar en el informe que es lo que hace cada una de las funciones del script.
- El script debe estar prolijo. Esto en particular implica que las variables tienen que tener un nombre descriptivo (es decir, no llamar `a`, `b`, `c` a las variables).

En el archivo `titles_train.csv` se presentan 4000 títulos de una plataforma de streaming. El archivo `credits_train.csv` contiene los actores y directores para estas películas y series. La idea de este trabajo es poder predecir la calificación de IMDB a partir de otras covariables para cada título. Siempre, a lo largo de este trabajo, se va a considerar la pérdida cuadrática como forma de evaluar modelos.

1. Hacer un análisis exploratorio de estos datos. Algunas ideas (no es obligatorio seguir esta lista, pueden ignorar algunos de estos items e inventar nuevos):
   (a) ¿Hay algún género que parezca estar más asociado con el puntaje del título?
   (b) ¿Cómo fue evolucionando este puntaje a lo largo del tiempo?
   (c) ¿Hay algún actor o director asociado con mayores o menores puntajes?
   (d) ¿Las películas más populares son las mejor puntuadas?
   (e) (Más complicada) Hay palabras de la descripción o del nombre del título que estén asociadas con un mayor/menor puntaje?
2. (a) Plantear un modelo de efectos fijos para predecir el puntaje de IMDB únicamente en función del país de origen.

(b) Plantear un modelo de efectos aleatorios para predecir el puntaje de IMDB únicamente en función del país de origen.

(c) Mostrar las estimaciones de los efectos de ambos modelos en un mismo gráfico e interpretar cómo se diferencian.

3. (a) Usando el modelo de efectos aleatorios del item anterior, decidir, usando la función `anova`, si agregaría la variable `release_year`.

(b) Usando el modelo de efectos aleatorios del item anterior, decidir si agregaría la variable `release_year` separando la data en dos: entrenamiento y testeo (estimar los coeficientes usando la data de entrenamiento y evualuarlo usando la de testeo).

(c) Comparar ambos items anteriores.

4. (a) Usando únicamente la variable `release_year`, predecir la popularidad de cada título (usando un tipo de modelo que crea adecuado) con una curva de splines penalizados. Usar $k = 1, 2, 3, 5, 10, 20, 50$ nodos y comparar todas las curvas estimadas en un mismo gráfico.

(b) Usando una partición entrenamiento / testeo, decidir cúal es el número óptimo de nodos a utilizar.

5. Dividir al conjunto de datos en entrenamiento y testeo (tambien puede usar otra técnica, como validación cruzada). Con todas las variables que tiene disponibles, probar al menos 10 modelos diferentes y elegir el que minimice el error cuadrático médio de predicción para el rating de IMDB.

6. En los archivos `titles_test.csv` y `credits_test.csv` aparecen 1806 nuevos títulos, para los cuales no aparece el rating de IMDB (pero yo sí los tengo). A partir del modelo elegido en el item anterior, producir un archivo `predicciones.csv` que tenga una sola columna que contenga, en la fila i, la predicción del rating de IMDB para el título de la fila i. (tiene que tener 1806 filas). A partir de estas predicciones, yo voy a computar el error cuadrático medio de predicción. El equipo que tenga el menor error cuadrático medio gana un premio sorpresa.