Análisis Avanzado de Datos

Taller 3

Instrucciones

- La calificación se dará sobre 100 puntos y el trabajo se desarrolla a lo más en grupos de 3 personas.
- El trabajo podrá ser recibido de manera posterior a la fecha acordada con una penalización en la nota como sigue:
  1. Entregar el trabajo 1 día después: la calificación obtenida se dará sobre 80 de los 100 puntos totales. Nota máxima: 80 puntos.
  2. Entregar el trabajo 2 días después: la calificación obtenida se dará sobre 60 de los 100 puntos totales. Nota máxima: 60 puntos.
  3. Entregar el trabajo 3 días después: la calificación obtenida se dará sobre 40 de los 100 puntos totales. Nota máxima: 40 puntos.
  4. Entregar el trabajo 4 días después: la calificación obtenida se dará sobre 20 de los 100 puntos totales. Nota máxima: 20 puntos.
- Una persona de la pareja debe enviar el informe o la URL de su repositorio publico junto a la página web de su trabajo al correo andresn.lopez@urosario.edu.co **el día 22 de Mayo antes de la media noche**. Un minuto posterior a media noche de la fecha acordada es considerado como el día siguiente a la entrega y será calificado de acuerdo al punto anterior.

Consideraciones Estas consideraciones no se verán reflejadas en la nota en forma de bonificación:

- El formato de entrega del trabajo puede ser un archivo word guardado en formato pdf, pero se recomienda a los estudiantes a realizar un cuaderno en rmd: vea un ejemplo: https://anlopezl.github.io/AED/NB1_E.html
- Se invita a los estudiantes a versionar su trabajo en git y trabajar en equipo usando la plataforma. Si se trabaja individualmente, este punto no será considerado en la bonificación.
- El estudiante puede publicar su trabajo mediante github pages y enviar la URL correspondiente.
- Para este taller **no podremos extender la fecha de entrega dada la finalización del semestre**.

Problema 1 - 20 pts (teórico). Una familia de distribuciones P_θ con $\theta \in \Theta$ pertenece a la familia exponencial de distribucciones si su fmp/fdp puede escribirse como:

$$p(x|\eta) = h(x)\exp(\eta(\theta)t(x) - a(\theta))$$

Para funciones reales $h(x)$, $a(\theta)$ y $t(x)$. Muestre que tanto la distribución bernoulli (utilizada para la regresión logística), la distribución normal (utilizada en la regresión lineal) y la distribución Poisson (utilizada en la regresión Poisson sobre conteos) pertenecen a esta familia de distribuciones.

Problema 2 - 50 pts (práctico). La Universidad de California Irvine (UCI) tiene un repositorio de datos de ejemplo para el uso de machine learning y aprendizaje estadístico. Uno de los conjuntos de datos es el denominado *Heart Disease*, su descripción detallada se encuentra en la URL a continuación:

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Como podrá ver, estos datos se encuentran disponibles para su uso en la siguiente URL:

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Utilice los datos procesados disponibles en el enlace presentado a continuación para el desarrollo del ejercicio,

http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

Con el conjunto de datos completo, construya un modelo de regresión logístico con función de enlace logit **tomando como respuesta la presencia de la enfermedad cardiaca**, use las demás variables como explicativas en el modelo de regresión. Revise las URL dadas para la definición de cada una de las variables y note que debe obtener la variable respuesta categorizando una de las variables del conjunto de datos. Siga los siguientes pasos en la realización del ejercicio:

1. Imputar datos: El conjunto de datos tiene datos perdidos en algunas variables. Estos están notados con un ?. Impute los valores perdidos como la mediana de los datos para las variables corresponientes.

2. Revisar las distribuciones bivariadas: Revise la distribución de la variable respuesta para cada una de las covariables categoricas de manera bivariada. ¿observa algún inconveniente con alguna de las variables al hacer el análisis?.

3. Modelo bivariado: Calcule manualmente (como lo vimos en clase, a partir de la tabla de contingencia), los parámetros estimados de regresión logística considerando únicamente la variable fbs (glucemia en ayunas) y la variable respuesta. Verifique el resultado ajustando el `glm` correspondiente.

4. Modelo multivariado: Ajuste un nuevo modelo con todas las variables. ¿cuáles variables son significativas mediante el test de Wald? ¿cuáles no lo son?.

5. Visualización de probabilidades predichas bajo modelo multivariado: Usando el modelo del punto anterior, encuentre las probabilidades de presentar enfermedad cardiaca y visualicelas junto a la variable respuesta. ¿Describe el modelo la presencia de enfermedad cardiaca?.

Problema 3 - 30 pts (práctico) El conjunto de datos `AAD-taller03.xlsx` contiene la predicción de incumplimiento de pago de tarjeta de crédito bajo dos modelos logísticos diferentes para un total de 9080 clientes. Se cuenta además con la variable de incumplimiento observada al finalizar el periodo. ¿Cuál de los dos modelos logísticos tiene mayor poder de predicción? Explique con fundamento estadístico su resultado.

Problema 4 - 10 pts (práctico) Este punto es **opcional** para la nota, pero puede mejorar su nota en 10 puntos adicionales. **De obtener la nota máxima en el presente taller, los puntos podrán subir la nota del taller 1 o el taller 2**.

Repita el problema 2, pero en lugar de imputar los datos mediante la mediana en el punto 1, utilice el algoritmo EM.