Proyecto 2 Estadística Matemática

(MATE 3520)

Semestre 02-23

El objetivo de este proyecto es proponer un modelo de regresión logística que responda adecuadamente, dentro de los criterios especificados más adelante, a la pregunta siguiente: *¿cual es la probabilidad de pasar el curso de cálculo diferencial dados ciertos criterios tempranos de evaluación?*

Para la creación y evaluación del modelo se usarán los datos de calificaciones del curso de cálculo diferencial que he dictado durante los tres semestres anteriores al semestre en curso. Estos datos, debidamente anonimizados, han sido compartidos por canales internos con los estudiantes de este curso.

Como parte de la evaluación, los estudiantes deben leer (e implementar, aunque no necesariamente con el mismo software) la exposición sobre regresión logística disponible en este link.

El proyecto ha de resolverse en parejas, por lo que espero tener un total de cinco reportes.

Las tareas a realizar son las siguientes:

Todas las parejas deben

1. Eliminar como espurias las entradas que tienen nota cero en el examen final (se asume que estos estudiantes retiraron el curso).

2. Usar como variables independientes

   $$(x_1, x_2, x_3, x_4)$$

   las notas del primer parcial, segundo parcial, primer quiz, segundo quiz, respectivamente. **Nota:** hay versiones "con curvas", recomiendo ignorar estas y tomar las llamadas "versiones definitivas" o simplemente las notas antes de curva.

3. La variable dependiente es

   $$y = 0, \qquad y = 1$$

   de acuerdo a si la nota definitiva es inferior o mayor o igual a 3, respectivamente. El valor de esta variable debe computarse.

4. Producir tres modelos, cada uno de los cuales usa dos de las tres hojas de cálculo compartida para estimar los parámetros respectivos.
5. Presentar la matriz de confusión correspondiente a cada modelo, evaluada en los datos usados para producir las estimaciones (tres matrices), y en los datos dejados por fuera para cada una de ellas (tres matrices adicionales).
6. Discutir cuál de los tres modelos anteriores se prefiere y porqué.
7. Evaluar para el modelo escogido la hipótesis nula según la cual todos los coeficientes son cero (ver por ejemplo este link para comenzar).

Individualmente cada pareja debe:

Estimar **dos veces** los coeficientes del modelo logístico anterior partiendo **todos los datos** en dos subconjuntos de acuerdo al valor de:

- **Pareja 1:** Sexo (ya clasificado como 0 y 1).
- **Pareja 2:** Tiempo de permanencia en la universidad: 0 si está en el primer o segundo semestre, 1 de lo contrario. El valor de esta variable debe computarse usando los datos.
- **Pareja 3:** Programa: 0 si es ingeniería, 1 si es otro. El valor de esta variable debe computarse usando los datos.
- **Pareja 4:** Porcentaje de asistencias (contadas) a la magistral: 0 si es menor que 70%, 1 de lo contrario. **En este caso sólo dos de las tres hojas de cálculo serán usadas**.
- **Pareja 5:** Nota examen final: 0 si esta es menor a 2.5, 1 de lo contrario.
- **Todas las parejas**: evaluar las cuatro hipótesis nulas según las cuales cada uno de los coeficientes del modelo es la misma para los datos con la variable categórica 0 y la variable categórica 1 (puede comenzar leyendo acá).
- **Todas las parejas:** evaluar la hipótesis nula según la cual la probabilidad de ganar el curso para los datos con variable categórica 0 es la misma que la probabilidad de ganar el curso para los datos con variable categórica 1 (esto es fácil y está en los contenidos del curso).

Notas adicionales

Todas las especificaciones del proyecto anterior deben ser seguidas ene ste proyecto (compartir códigos, explicar métodos, etc). Debe adicionalmente crear "versiones estandarizadas" de los datos debidamente filtrados, y compartirlas en su entrega explicando de manera clara las entradas que seleccionó en las hojas de datos "crudos" ya compartidos. Estas deben incluir en particular los valores de las variables categóricas correspondientes al trabajo de la pareja, y deben ser las usadas para correr los códigos respectivos.