Examen 2A. Selección de variables y aprendizaje no supervisado

Gonzalo Pérez, Francisco Zárate y Dioney Rosas

Semestre 2023-1

La solución del examen se deberá subir al classroom antes de las 11:59 PM del 30 de noviembre de 2022. Todas las preguntas valen 3.5 puntos. Favor de argumentar con detalle las respuestas.

NOTA 1. En caso de que se identifiquen respuestas iguales en otros examenes, se procederá a la anulación de los examenes involucrados.

NOTA 2. Usar una confianza de 95 % o una significancia de .05 en los casos en donde no se requiera otro nivel de forma explícita.

NOTA 3, sobre el formato de entrega. La solución de cada pregunta deberá incluir un reporte ejecutivo (pdf) y por separado un archivo donde se pueda replicar TODO resultado que se presente en el reporte ejecutivo, así como lo correspondiente al preprocesamiento y diferentes opciones exploradas (R, Rmd, etc.). El reporte ejecutivo **no debe pasar de 4 páginas por pregunta** y deberá incluir la descripción del modelo y/o los resultados, tablas, figuras y de las pruebas de hipótesis relevantes. Toda figura, tabla o resultado que se incluya DEBE estar descrito (explicado) y referido en el texto, de otra forma aunque se presente en el documento o se pueda generar con los scripts no se tomará en cuenta como parte de la solución. Por otra parte, los scripts deben estar comentados, al menos de grosso modo, es decir, indicando el objetivo de conjuntos de líneas de código.

1. Selección de variables.

Considere la base de datos *fat* del paquete *faraway*, considere todas las variables, excepto siri, density y free. También eliminé del análisis los casos con valores extraños en weight y height, así como valores cero en brozek. Suponga que el objetivo del estudio es usar las variables clinicas observadas en los pacientes para estudiar cuáles de éstas son los factores que ayudan a modelar mejor el promedio del porcentaje de grasa corporal en los hombres (var brozek).

i. Considere un modelo para datos continuos con liga identidad y distribución Gaussiana. Realice una selección de variables considerando sólo los efectos principales de las variables y usando: a) mejor subconjunto, b) un método stepwise y c) método lasso. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.

ii. Considere un modelo para datos continuos con liga identidad y distribución Gaussiana. Realice una selección de variables considerando en el modelo los efectos principales de las variables, así como su interacción, sólo considerando: a) un método stepwise y b) método lasso. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.

iii. Considere posibles modificaciones a los incisos i) y ii) realizando lo siguiente. A) usar distribución Gamma (ligas identity y log); B) usar en los modelos de forma adicional la versión al cuadrado de las variables. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.

iv. Presente en una sóla tabla los diferentes modelos obtenidos, así como el BIC de cada uno. Comente sobre los resultados, por ejemplo, qué variables aparecen en la mayoría de modelos, si parece necesario incluir interacciones o realizar un preprocesamiento a los datos y, considerando el mejor de todos, qué interpretación se puede dar a algunos de los coeficientes del modelo.

2. Componentes principales y análisis exploratorio factorial

Considere los datos en el archivo "Dat2Ex.csv", sólo los casos donde se tiene respuesta en todas las variables. Estos datos corresponden a una encuesta realizada a los alumnos de una universidad sobre el desempeño de un curso. Las respuestas van de 1 a 5, donde 5 es que consideran que el aspecto analizado en la pregunta es muy bueno, mientras que 1 es deficiente. La descripción de las preguntas es:

ITEM13 ÏNSTRUC WELL PREPARED"

ITEM14 ÏNSTRUC SCHOLARLY GRASP"

ITEM15 ÏNSTRUCTOR CONFIDENCE"

ITEM16 ÏNSTRUCTOR FOCUS LECTURES"

ITEM17 ÏNSTRUCTOR USES CLEAR RELEVANT EXAMPLES"

ITEM18 ÏNSTRUCTOR SENSITIVE TO STUDENTS"

ITEM19 ÏNSTRUCTOR ALLOWS ME TO ASK QUESTIONS"

ITEM20 ÏNSTRUCTOR IS ACCESSIBLE TO STUDENTS OUTSIDE CLASS"

ITEM21 ÏNSTRUCTOR AWARE OF STUDENTS UNDERSTANDING"

ITEM22 Ï AM SATISFIED WITH STUDENT PERFORMANCE EVALUATION"

ITEM23 ÇOMPARED TO OTHER INSTRUCTORS, THIS INSTRUCTOR IS"

ITEM24 ÇOMPARED TO OTHER COURSES THIS COURSE WAS"

i. Asumiendo que las variables son continuas, obtenga los componentes principales e indique si se pueden identificar dimensiones interesantes de estos datos. Explore el uso de los datos en la escala original y con alguna escala tranformada.

ii. Asumiendo que las variables son continuas, aplique la técnica de análisis exploratorio factorial e indique si se pueden identificar dimensiones interesantes de estos datos. Explore el uso de los datos en la escala original y con alguna escala tranformada.

iii. Realice modificaciones en i) y ii) considerando a) que los datos son categóricos ordinales y b) rotaciones a los resultados. Considerando todos los resultados, sólo seleccione un conjunto de componentes o factores, los que le parecen más adecuados, e interprete.

Nota. la interpretación de los resultados es lo más importante, así que trate de argumentar ésta, puede incluir algunas gráficas.

3. Análisis de conglomerados

Considere los datos en el archivo *Dat3ExA.csv*, sólo los casos con respuesta en todas las variables. Estos datos corresponden a una encuesta realizada por la compañía Oddjob Airways con la intención de conocer las expectativas de sus clientes sobre ciertos aspectos del servicio de la compañia. El objetivo es analizar si se pueden identificar grupos de clientes que en un futuro se puedan usar para focalizar la publicidad de la empresa. Las respuestas van de 1 a 100, donde 100 es que la persona considera que ese aspecto es crucial en el servicio, mientras que 1 corresponde a que no lo es. La descripción de los aspectos que se consideran es:

e1 "... with Oddjob Airways you will arrive on time."

e3 "... in case something does not work out as planned, Oddjob Airways will find a good solution."

e4 "... the flight schedules of Oddjob Airways are reliable."

e5 "... Oddjob Airways provides you with a very pleasant travel experience."

e6 "... Oddjob Airways's on board facilities are of high quality."

e9 "... Oddjob Airways gives you a sense of safety."

e11 "... the interior of Oddjob Airways's aircraft is well maintained."

e15 "... the food and beverage items served by Oddjob Airways are of a high quality."

e18 "... Oddjob Airways's entire personnel are customer and service-oriented."

e21 "... Oddjob Airways makes traveling uncomplicated."

e22 "... Oddjob Airways provides you with interesting on-board entertainment, service, and information sources."

i. Asumiendo que las variables son continuas, obtenga algunos grupos considerando el método k-means. Explore el uso de los datos en la escala original y con alguna escala tranformada.

ii. Asumiendo que las variables son continuas, obtenga algunos grupos considerando el método de conglomerados jerárquico aglomerativo. Explore el uso de los datos en la escala original y con alguna escala tranformada, así como varias disimilaridades (entre clientes y clusters).

iii. Realice modificaciones en i) y ii) considerando que se usan algunos componentes principales en lugar de las variables originales. Considerando todos los resultados, sólo seleccione un conjunto de conglomerados, los que le parecen más adecuados, e interprete.

Nota. la interpretación de los resultados es lo más importante, así que trate de argumentar ésta, puede incluir algunas gráficas y estadísticas por grupo.