Tarea 2B. Selección de variables, bootstrap y aprendizaje no supervisado

Gonzalo Pérez, César Valle y Rodrígo Jiménez

Semestre 2024-2

La solución de la tarea se deberá subir al classroom antes de las 11:59 PM del 15 de mayo de 2024. La pregunta 1 vale 1.5 puntos, el resto de preguntas valen 3 puntos. Favor de argumentar con detalle las respuestas.

NOTA 1. En caso de que se identifiquen respuestas iguales en otras tareas, se procederá a la anulación de las tareas involucradas.

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. Bootstrap no paramétrico

Sea $X_1, ..., X_n$ una m.a de la distribución $Poisson(\theta)$. Supongamos que el parámetro de interés a estimar es $\tau(\theta) = e^{-\theta} = \mathrm{P}(X = 0)$. Se puede verificar que $\widehat{\tau}(\theta) = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}$ es el UMVUE de $\tau(\theta) = e^{-\theta}$, sin embargo, no es tan fácil encontrar la distribución de $\widehat{\tau}$ o la expresión de $V(\widehat{\tau})$.

a. Una forma para aproximar el valor de $V(\widehat{\tau})$ es el método Monte Carlo (MC). En general, con este método

$$E(g(Z)) \approx \frac{\sum_{b=1}^{B} g(Z_b)}{B},$$

donde $Z_1, ..., Z_B$ son números aleatorios de la distribución de la variable aleatoria Z. En este caso, para aplicar el método MC, deberíamos poder generar números aleatorios de la distribución de $\widehat{\tau}$, lo cual podemos realizar si se conoce θ y también el tamaño de la muestra n, de manera que considerando B posibles muestras aleatorias de tamaño n de una distribución $Poisson(\theta)$ se pueden estudiar algunas propiedades teóricas de $\widehat{\tau}(\theta)$ con los valores $\widehat{\tau}_1, ..., \widehat{\tau}_B$ que se consideran como los datos provenientes de la distribución de $\widehat{\tau}(\theta)$.

Suponga que $\theta = 1$, $n = 20$, $B = 10,000$ y que usará el método MC. Dé una aproximación de $E(\widehat{\tau})$ y $V(\widehat{\tau})$. También presente el histograma de $\widehat{\tau}_1, ..., \widehat{\tau}_B$.

b. A diferencia de a), en la práctica sólo observamos una m.a. $X_1, ..., X_n$ y quizás no sabemos ni suponemos una distribución específica. Aún así, dado un parámetro de interés, deseamos dar alguna estimación

y medidas de variabilidad, es decir, usamos un estimador específico, por ejemplo $\widehat{\tau}$, y estimamos la varianza de la estimación (estimamos $V(\widehat{\tau})$ sólo con la muestra). En estos casos podemos usar el método bootstrap no paramétrico.

i. Genere $n = 20$ números aleatorios de una distribución $Poisson(\theta)$, con $\theta = 1$.

ii. Suponga que la muestra generada en b.i) es la única que tiene en una base de datos y le indican que use $\widehat{\tau} = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}$ para estimar $\mathrm{P}(X = 0)$, dé la estimación de $\mathrm{P}(X = 0)$, así como una estimación de la varianza usando el método bootstrap no paramétrico con $B = 10,000$. También presente el histograma de $\widehat{\tau}^*_{(1)}, ..., \widehat{\tau}^*_{(B)}$. Comente los resultados, por ejemplo, comparando con lo obtenido en a).

Nota. Para la generación de número aleatorios use como semilla los tres últimos números del número de cuenta de uno de los integrantes del equipo.

2. 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 clínicas 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 o 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.

3. Componentes principales y análisis factorial exploratorio

Considere los datos en el archivo "Dat3Ex.csv". Estos datos corresponden a una encuesta que intenta analizar la personalidad de un grupo de 228 alumnos de licenciatura de una universidad de Estados Unidos. Las respuestas van del 1 al 5 de acuerdo con el cuestionario de la Figura 1.

Sólo considere las variables: V1, V2, V4, V6, V9, V12, V14, V16, V17, V26, V27, V29, V31, V34, V37.

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 transformada.

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.

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Disagree strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Agree Strongly 5

I see Myself as Someone Who...

___1. Is talkative	___23. Tends to be lazy
___2. Tends to find fault with others	___24. Is emotionally stable, not easily upset
___3. Does a thorough job	___25. Is inventive
___4. Is depressed, blue	___26. Has an assertive personality
___5. Is original, comes up with new ideas	___27. Can be cold and aloof
___6. Is reserved	___28. Perseveres until the task is finished
___7. Is helpful and unselfish with others	___29. Can be moody
___8. Can be somewhat careless	___30. Values artistic, aesthetic experiences
___9. Is relaxed, handles stress well	___31. Is sometimes shy, inhibited
___10. Is curious about many different things	___32. Is considerate and kind to almost everyone
___11. Is full of energy	___33. Does things efficiently
___12. Starts quarrels with others	___34. Remains calm in tense situations
___13. Is a reliable worker	___35. Prefers work that is routine
___14. Can be tense	___36. Is outgoing, sociable
___15. Is ingenious, a deep thinker	___37. Is sometimes rude to others
___16. Generates a lot of enthusiasm	___38. Makes plans and follows through with them
___17. Has a forgiving nature	___39. Gets nervous easily
___18. Tends to be disorganized	___40. Likes to reflect, play with ideas
___19. Worries a lot	___41. Has few artistic interests
___20. Has an active imagination	___42. Likes to cooperate with others
___21. Tends to be quiet	___43. Is easily distracted
___22. Is generally trusting	___44. Is sophisticated in art, music, or literature

Figura 1: Cuestionario sobre personalidad

4. Análisis de conglomerados

Considere los datos en el archivo *Dat4ExB.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 compañía. 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."

e2 "... the entire journey with Oddjob Airways will occur as booked."

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

e8 "... Oddjob Airways offers a comfortable on-board experience."

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

e10 "... the condition of Oddjob Airways's aircraft is immaculate."

e16 "... Oddjob Airways offers you a variety of foods and beverages that fits your personal needs."

e17 "... all of Oddjob Airways's personnel are always hospitable and welcoming."

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 transformada, 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.