RESEARCH

Disease Prediction Using Machine Learning

Joel A Rodarte-Rivera

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract.

Background: the context and purpose of the study

Methods: how the study was performed and statistical tests used

Results: the main findings

Conclusions: brief summary and potential implications

Keywords: sample; article; author

Background

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

Correspondence: joel.rodarter@uanl.edu.mx
Facultad de Ciencias Fisicomatemática, UANL, Monterrey, Nuevo Léon, México
Full list of author information is available at the end of the article

Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

Discussion

Algoritmos no supervisados

IMPORTANTE: Por el momento, considero que la explicación del modelo No me queda muy claro dado existen conceptos de probabilidad con los que no estoy familiarizado. Sin emabargo, me es importante al menos saber de su existencia para dominarlo más adelante en el transcurso del curso.

Akaike Information CriterioBenroulli Mixture Model

¿Qué es?

- La distribución Bernoulli es una distribución discreta que predice la probablidad de éxito (1) o fracaso (0) de un evento. El término "mixture" se incluye dado que este método combina multiples distribuciónes de probabilidad para modelar una distribucón más compleja.

- Respecto al punto anterior, este algoritmo asumme que la información se genera de K mezclas de distribuciónes de Bernoulli, en donde cada componente de la mezcla representa a un cluster.
- Cada cluster es caracterizado por un vector de probabilidades que predice la probabilidad de encontrar un 1 en ese cluster. La distribución de la mezcla general es entonces una suma ponderada de las distribuciones de K Bernoulli, donde los pesos especifican la proporción de los datos que pertenecen a cada grupo. (?)
- Este último punto es el que no me queda muy claro.

Clustering para datos binarios

Akaike Information Criterion

El Akaike Information Criterion es un un solo numero que se utiliza para determinar cual dentro de múltiples opciónes de modelos mejor se ajusta para un set de datos. Este numero se puede utilizar en combinación con el Bernoulli Mixture Model para determinar la cantidad de cluster óptima.

Tarea 5

Se determina que para los 4920 pacientes los síntomas más comunes son los siguientes con su respectiva incidencia.

$fatigue$	0.393
$vomiting$	0.389
$high_fever$	0.277
$loss_of_appetite$	0.234
$nausea$	0.233

Kmeans

Se utilizó los 2 síntomas más comunes (fatiga y vómito) para ver la combinación de su aparición en todos los pacientes. Dado se sabe que sólo hay 4 combinaciones posibles de resultados para x enfermedad, se utilizó el algoritmo no supervisado de k medias con 4 clusters.

Los 4 resultados posibles son

- Que paciente no tenga ninguno de los dos síntomas
- Que paciente tenga fatiga pero no vómito
- Que paciente tenga vómito pero no fatiga
- Que paciente tenga ambos síntomas

Los centroides de los 4 clusters son los siguientes:

	x	y
0	0.000	0.000
1	1.000	0.000
2	−0.000	1.000
3	1.000	1.000

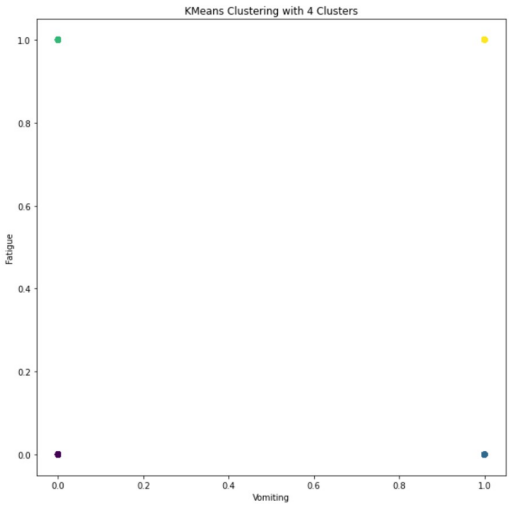
La cantidad de pacientes por cada centroide es la siguiente

	ck	nk
0	0	1836
1	1	1170
2	2	1152
3	3	762

Interpretación

- Se tienen 1826 pacientes que para alguna enfermedad NO tienen ni vomito ni fatiga
- Se tienen 1152 pacientes que para alguna enfermedad NO tienen fatiga pero SI vómito
- Se tienen 1170 pacientes que para alguna enfermedad SI tienen fatiga pero NO vómito
- Que QSe tienen 762 pacientes que para alguna enfermedad SI tienen fatiga y SI vómito

Dado se cuenta con información binaria se espera que el plot de k medias sea un cuadrado por que sólo hay 4 posibles resultados.

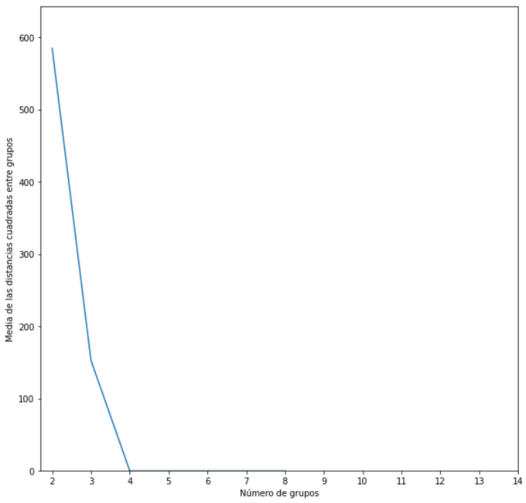


Selección número de clusters adecuado

Por el método del codo se calcula la inercia para entender que sucedería con la dispersión de los datos considerando 2 a 8 clusters.

	nclusters	inertia
0	2	584.658
1	3	152.878
2	4	0.000
3	5	0.000
4	6	0.000
5	7	0.000
6	8	0.000

Dada la información es binaria se sabe que debería ser suficiente con 4, pero se quiere entender como cambiaría añadiendo hasta 4 más, se espera que la diferencia sea pequeña o nula despues de cuatro. Además, se espera que sea muy grande menos de 4.



Tarea 6

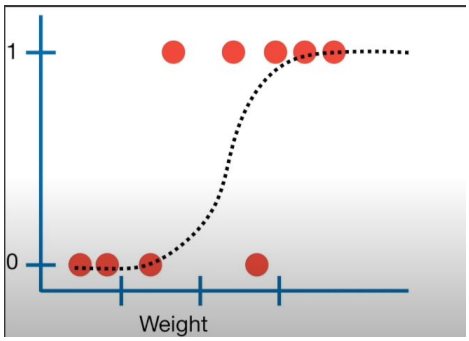
¿Qué es el la regresión logistica y porqué funciona con data frames de datos binarios?

La regresión logística es una algoritmos supervisado común para problemas que involucracon clasificiación binaria. Se basa en modelar la probabilidad de que una variable de respuesta tome el valor de 1 en función a variables predictoras. Se dice que tome el valor de 1 dado que la variable de respuesta solo puede ser 0 o 1. Aplicado al set de datos trabajado este tetramestre 1 representaría la presencia de una enfermedad mientaras que 0 sería la ausencia de la misma.

El algoritmo calcula la suma ponderada de las variables predictoras para producir un calificación. Esta calificación después se evalua con una función logistica para obtener la probabilidad estimada para que la varibales de salida tome el valor de 1.

La función logística tiene una forma de S, como la siguiente imágen lo cual quiere decir que la respuesta de la función es 0 cuando el valor de entrada es negativo y es 1 cuando el valor de entrada es positivo. La función logisitica se asegura que las probabilidades siempre esten entre 0 y 1 por lo que se presta a utilizarla en problemas de clasificación binaria.

Una vez que se entrena el modelo, se puede utilizar para predecir la probabilidad que la variable de salida tome el valor de 1. Una vez que se obtiene dicha probabilidad cualquier valor por encima de .5 se le clasifica como 1, presencia de enfermedad y 0 cuando es menor de .5, o ausencia de enfermedad.



Métricas para analizar performance Logistic Regression

Para evaluar el desempeño de este algoritmo se utiliza técnas como accuracy y ROC (Receiver Operating Characteristic). Accuracy no es más que la division de los casos correctamente predecidos entre los casos totales. ROC es una manera gráfica de interpretar una matriz de confusión para diferenciar de manera gráfica cuantos faltos positivos y verdaderos positivos hay. Más adelante en esta tarea se muestra como realizarlo en python.

Estrategia para tarea 05

Se cuenta con un set de datos de 4920 pacientes. Voy a dividir esos datos en 70

El objetivo es preddecir dados 133 síntomas posibles, predecir la presencia o no de diabetes.

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations should be provided.

Declarations

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
JRR was a major contributor in writing the manuscript.

Acknowledgements
Universidad Autónoma de Nuevo León

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

References
Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.