Sesión Laboratorio
Parte I – Manejo de Datos Faltantes y Normalización

CI-2600

Objetivos

Familiarizarse con los fundamentos de las estrategias de manejo de:

- datos faltantes
- normalización de datos
- técnicas de balanceo
- selección y extracción de variables

Manejo de datos Faltantes

El manejo de datos faltantes es una estrategia de pre-procesamiento utilizada en casos de entradas faltantes

En casos de datos faltantes se puede recurrir a varias maneras

1- Solicitar que se rellenen -> el cliente provee

2- Rellenar a cero -> quizás la herramienta para medir no es lo suficientemente sensible?

3- Imputar con valores de media -> asume que los datos faltantes se comportan como la media de los datos

4- Interpolar con valores de vecinos cercanos -> asume comportamiento de datos basado en patrones

5- Eliminación

Efectos del Tratamiento

Cada método tiene sus ventajas y desventajas

Eliminar datos implica que se pierde posible información pero no se añade información potencialmente ruidosa

El uso de ceros se adhiere a una realidad de la sensibilidad de los aparatos analíticos con umbrales experimentales, pero extrapola valores que pueden no ser reales (no todo cero es realmente un cero)

La media genera información nueva y se reduce posible pérdida de información por eliminación pero se le dan valores que no son enteramente reales a los datos

La interpolación asume que los datos tienen un comportamiento similar, lo cual no es necesariamente cierto

Entonces… Que debo Usar?

No hay meetodo perfecto

Sin embargo dependiendo del problema o la preginta científica yl el contexto se debe implemewntar cada meetodo!

El el contexto de datos biológicos casi nunca se pretende eliminar datos, por qué? Los aparatos analíticos tienen un umbral de sensibilidad, y un NA puede ser o muy baja expresión o un error de detección

Además si hay pocas muestras, se corre el riego de perder informacieon

Imputar con RFs

En R se puede imputar de distintas maneras, por ejemplo en este caso, para este laboratorio vamos a imputar usando RF en un set de datos de mamografía

Al ser un RF es un método no-paramétrico, por lo que puede imputar variables continuas y aquellas de tipo categóricas

Pero que hace? Pues mediante la construcción de RF de las variables predice los valores de las variables faltantes

- Recuerden que sies continua es mediante regresión y por lo tanto es la media
- Si es carteg´∑orica es una clasificación y es la que tiene más votos

Comprobación

Como determino yo que método usar para cada caso?

En primera instancia comparo el desempeño de los modelos que voy a contruir

Es decir, que debo ver la métricas y su representación o lo que implican

Que me dicen la sensibilidad, la especificidad

Normalización de Datos

El proceso de normalizar los datos cumple una función importante

Me permite compensar el efecto de valores extremos

Me ayuda a convertir los datos en equicomparables

Es decir que convierte a los tratamientos en “iguales”

Esto es importante porque la normalización reduce las diferencias entre tratamientos

Que significa esto?

OK! Resulta que yo tengo dos efectos, el efecto dentro de mi tratamiento pero también el efecto entre los tratamientos

Los Efectos de la Normalización

Siguiendo

Entonces… si yo no normalizo, yo no puedo comparar!

Por qué? Sencillamente porue cada muestra ha sido tratada de forma diferente, es decir podemos verla como un ente distinto que debe ser convertida a un esteandar comparable con las otras… como lo logro?

Aplicando un meetodo de equivalencia o equiparación… la normalización

Existen varios métodos… por ejemplo:

Log2

Min/max cc transformación 0/1

Z-score

MAD

Cuando y por qué?

Esto es una excelente pregunta!

Las técnicas de normalización, al igual que las de manejo de datos faltantes tienen mucho que ver con el tipo de problema!

Por ejemplo… un set de datos donde tengo el efecto de valores extremos, como debe ser tratado? Dependiendo de la kurtosis y de un análisis preliminar o descriptivo debo determinar:

el tipo de normalización

cómo me va a afectar en análisis posteriores

Cuidados de la Normalización

En el ejemplo de mamografía de hoy podemos ver que la normalización tiene un efecto inesperado

Por qué puede suceder esto?

Como lo podríamos corregir?

Es esto el producto de la metodología de normalización aplicada?

Probablemente no! Por qué??

Sesión de Laboratorio Parte II – Balanceo y Reducción de la Dimensionalidad

CI-2600