CIFF

Araceli Macia Barrado <aracelimacia@campusciff.net>

Dudas del Ejercicio Final

6 mensajes

Araceli Macia Barrado <aracelimacia@campusciff.net> 29 de junio de 2016, 0:05
Para: Alberto Torres Barrán <albertotb@gmail.com>

Buenas!
Te he escrito por el classroom, pero igual es mas cómodo por mail..
Es que revisando el ejercicio final, tengo dudas de esto:

> Escalar los datos para que tengan media 0 y varianza 1, es decir, restar a cada variable num´erica su media y dividir por la desviaci´on t´ıpica. Calcular la media y desviaci´on en el conjunto de train, y utilizar esa misma media y desviaci´on para escalar el conjunto de test.
>
> Lo que yo entiendo que hay que hacer es esto:
>
> Tengo dos DataSet, el de Entrenamiento(train) y el de test.
>
> 1) En el de train, de cada columna numérica calculo la media y la desviación tipica. Es decir, tendré una media y una desviación por cada columna numerica.
>
> 2) En el Dataset de Test, a cada valor le resto la media y divido por la desviacion tipica (de la columna correspondiente calculada en el Data Set de Train)
>
> De hecho, ya lo he hecho, y efectivamente en el DataSet de Test, redondeando ya me sale que la media es 0 y la varianza es 1.

¿ me confirmas que lo he entendido bien?
Es que avanzado en el ejercicio, veo este otro punto :

> (Opcional) Realizar un modelo de regresi´on lineal de la variable de respuesta sobre el resto y ajustarlo por m´ınimos cuadrados usando u´nicamente los datos del conjunto de entrenamiento.

Y la variable de respuesta, ya si que no entiendo cual es???
Queria intentar hacer los ejercicios opcionales, aunque la verdad que lo de estadistica de momento me cuesta mucho...

Gracias!

--
Araceli Macía Barrado

Alberto Torres <albertotb@gmail.com> 29 de junio de 2016, 12:00
Para: Araceli Macia Barrado <aracelimacia@campusciff.net>

Hola Araceli,

La primera parte la has entendido bien. Después de restar la media y dividir por la desviación en train, obviamente te tiene que salir 0 y 1 si la calculas. En test, como comentas, puesto que es un submuestreo aleatorio de train debería salirte algo muy pequeño como media después de normalizar y algo muy parecido a 1 al dividir por la desviación.

En cuanto a la segunda parte, la variable respuesta es la Y.

Un saludo,
Alberto
[El texto citado está oculto]
--
Un saludo,

Alberto

Araceli Macia Barrado <aracelimacia@campusciff.net> 30 de junio de 2016, 10:56
Para: Alberto Torres <albertotb@gmail.com>

OK, gracias!

otra duda.. en este punto :

>Calcular la media para las filas que tienen SEX=M y la media para las filas que tienen SEX=F, utilizando la funcíon tapply.

te refieres a hacer algo como esto :
tapply(Datos$AGE,Datos$SEX,mean)

pero de todas las columnas numericas?? o lo que hay que hacer es mostrar solo dos datos, uno de la media de todos los valores en conjunto del dataSet de las filas que SEX sea M y otro con valor F?

Es que lo que yo he interpretado, es ejecutar taply con todas las columnas numericas,y para ello he hecho una funcion para pasarle tapply todas las columnas.. y me va dando los resultados de F y M de cada uno de los valores, algo asi:

$S5 (con la columna S5)
F M
4.731533 4.569392

$S6
F M
93.86207 88.94348

$Y
F M
155.8079 149.0391

pero releyendo el ejercicio tengo dudas por si he interpretado mal lo que hay que hacer.

gracias!

[El texto citado está oculto]
--
Araceli Macía Barrado

Alberto Torres <albertotb@gmail.com> 30 de junio de 2016, 16:43
Para: Araceli Macia Barrado <aracelimacia@campusciff.net>

Hola,

Efectivamente, me refería a lo que has hecho. Básicamente se trata de calcular las medias de todas las columnas menos SEX agrupadas por sexo.

Un saludo,
Alberto
[El texto citado está oculto]
--
Un saludo,

Alberto

Araceli Macia Barrado <aracelimacia@campusciff.net> 1 de julio de 2016, 8:27
Para: Alberto Torres <albertotb@gmail.com>

Genial, gracias!
y la ultimisima pregunta.. el modelo de regresión lineal de la variable Y sobre el resto, es hacer n modelos de regresión lineal, es decir de la variable Y con cada una de las columnas? Es que he estado mirando en internet que es un modelo de regresión multiple.. y no se si seria mas bien esto que los n modelos?

[El texto citado está oculto]
--
Araceli Macía Barrado

Alberto Torres <albertotb@gmail.com> 1 de julio de 2016, 10:53
Para: Araceli Macia Barrado <aracelimacia@campusciff.net>

Hola Araceli,

No, es hacer un único modelo de regresión lineal. Únicamente tienes que hacer varios si tienes múltiples variables de respuesta (o uno de regresión múltiple, como mencionas). Si solo tienes una, es un único modelo y que no tiene límite en el numero de variables de entrada que puede tener (con un pequeño matiz, pero en este caso no es importante).

Un saludo,
Alberto
[El texto citado está oculto]
--
Un saludo,

Alberto