Guión cutre práctica 2 ML

Entrega: 20 de octubre

1. Reducción de dimensionalidad

Hay que cambiar la clase de 1-2 a 0 - 1 restando la mediana (o restando 1 y ya)
Se puede quitar el target de la matriz para simplificar trabajo.

1.1. Mutual information

Se usa la función de sklearn (v 0.19) de sk.feature_selection.mutual_info_classif(var, target). Se puede hacer con regresion y clasificación pero en general la de regresion da mejores resultados. Aquí se usa clasificacion.

Se ordenan la mutual information entre la variable y la clase, y nos quedamos con las variables con una MI más alta.

La función de MI te discretiza las variables él solito y sin ayuda.

Si se hace con regresión existe una pequeña diferencia al final. Si se pintar las clases con respecto a la variable (scatterplot)

1.2. χ^2 test

Solo se puede aplicar cuando los valores de las variables son positivas, que lo son.

Chi2 tambien esta en sklearn.

Mirando los valores de los estadísitcos, se pueden seleccionar aquellos valores más altos que serán los que tengan una probabilidad más alta de caer en la zona crítica de la distribución.

Se comparan todos los metodos y se escogen aquellos con mejores resultados en ambos métodos.

1.3. PCA

Se usa la librería de sklearn. Es un método no supervisado (OJO CUIDAO).

Se puede pintar la varianza explicada (plot(PCA.explained_variance_)) y el ratio.

Si la primera PCA explica toda la variaza(99 %), eso es que o hay un problema de escala (no están normalizados) o todas están claramente correlacionadas.

Cuando se escalan es cuando las cosas empiezan a tener sentido ya que el método no es invariante a la escala.

El porcentaje de varianza explicada que se quiere llegar es el de 90 %.

Se calcula en X_{train} y se transforma el X_{test} para que esten en el mismo espacio. Esto ocurre de forma similar al normalizar.

PCA al ser np-supervisado es invariante a la clase. **Hacerla directamente normalizada**

1.4. LDA

Para LDA se usa otra vez (SORPRESA) la librería de sklearn.

LDA es invariante a escala por lo que no haría falta normalizar.

En nuestros datos, no deberíamos tener demasiada esperanza con LDA debido a que nuestras variables de gaussianas tienen bien poco por no decir nada.

El scatterplot otra vez nos da una indicación de si se ha separado de forma correcta.

1.5. Seleccion del numero de variables en el caso anterior

1.5.1. Metodo forward

Se empieza siempre cogiendo la variable o componente más significativas, aumentando uno hasta alcanzar el mínimo en el error o una convergencia (se producirá un mínimo, después llegaría al overfitting (Eso dice eso)).

1.5.2. Backward

Al reves, ir quitando de una en una las variables menos significativas. **Esta no hay que hacerla**

2. Regresion logistica

Se hace con todas las reducciones y se da el error en X_{test} para comprobar cual ha sido mejor.

Hay que entregar una tabla dando la Accurancy, precision, sensitivity y Specifity para cada uno de los metodos de selección de variables.