API UR

http://apiur.es/apiweb/course/view.php?id=12

INGENIERÍA DEL CONOCIMIENTO

ALUMNO: Eduardo Dulce Chamorro

Práctica 1

1. **Desarrolle un modelo de regresión lineal del índice de mortalidad sin realizar ningún ajuste ni eliminando ninguna variable. Interprete el modelo y explique el significado de los errores de validación cruzada obtenidos.**

Comandos:

```
Weka > Explorer
OpenDB > pollution.arff
Classify > Crossvalidation > Linear regression > Start
```

Información obtenida:

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-
places 4
Relation:     pollution
Instances:    60
Attributes:   16
              PREC
              JANT
              JULT
              OVR65
              POPN
              EDUC
              HOUS
              DENS
              NONW
              WWDRK
              POOR
              HC
              NOX
              SO@
              HUMID
              MORT
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

MORT =

      1.8565 * PREC +
     -2.262  * JANT +
     -3.32   * JULT +
    -10.9205 * OVR65 +
   -137.3831 * POPN +
    -23.4211 * EDUC +
      4.6623 * NONW +
     -0.9221 * HC +
      1.871  * NOX +
   1934.0539

Time taken to build model: 0.1 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7423
Mean absolute error                     32.4407
Root mean squared error                 42.7162
Relative absolute error                 64.4106 %
Root relative squared error             68.4085 %
Total Number of Instances               60
```

ANOTACIONES:

MAE: Mean absolute error. Media de los errores absolutos.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

RMS: Root mean squared error. Media de la raíz de los cuadrados de los errores.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}.\ \mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

En este caso observamos que la diferencia entre los errores MAE y RMS no es muy alta 31%, por lo que la predicción del modelo no es muy errónea.

2. **Normalice los datos entre 0 y 1 (incluso la variable de salida). Vuelva a crear el modelo y muestre los errores de validación cruzada 10. Interprete los errores obtenidos.**

Comandos:

```
Weka > Explorer
Preprocess > Filter > Unsupervised > Attribute > Normalize
# así normaliza todos los parámetros excepto la salida (MORT)
# vamos a normalizar la salida
Pinchamos donde pone Choose Normalize > Ignore Class > Marcar True
```

Prep

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-
places 4
Relation:     pollution-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-unset-class-temporarily
Instances:    60
Attributes:   16
              PREC
              JANT
              JULT
              OVR65
              POPN
              EDUC
              HOUS
              DENS
              NONW
              WWDRK
              POOR
              HC
              NOX
              SO@
              HUMID
              MORT
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

MORT =

      0.2879 * PREC +
     -0.3859 * JANT +
     -0.2265 * JULT +
     -0.21   * OVR65 +
     -0.2599 * POPN +
     -0.2397 * EDUC +
      0.5451 * NONW +
     -1.8504 * HC +
      1.8454 * NOX +
      0.7976

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7423
Mean absolute error                      0.1006
Root mean squared error                  0.1325
Relative absolute error                 64.4106 %
Root relative squared error             68.4085 %
Total Number of Instances               60
```

Los errores ahora al estar normalizados los valores son porcentuales, por tanto el MAE sería 10,06% mientras que el RMS sería 13,25%.

3. **Identifique las variables más importantes que influyen en el modelo realizado en el punto 2 (con los datos normalizados). Calcule el porcentaje de influencia relativo de cada variable frente a la más influyente.**

Igual que lo que he indicado en el punto anterior. Al estar los valores normalizados tienen todos los mismos pesos en la fórmula de la regresión por lo que podemos comparar todas las variables directamente.

Las variables que más influyen son por orden: HC, NOX, NONW, POPN, EDUC, JANT, PREC, JULT.

4. **Compare los pesos que multiplican las variables del modelo lineal del punto 1 frente a las variables más importantes que usa el modelo con los datos normalizados entre 0 y 1 desarrollado en el punto 2. ¿Por qué no son proporcionales entre el modelo del punto 1 y el 2?**

La conclusión al comparar el modelo 1 y el 2 es que la función NO SE PUEDE COMPARAR, porque al no estar normalizados en el modelo 1 los pesos no son proporcionales.

5. **Realice modificaciones en los parámetros del modelo y/o seleccione/elimine manualmente las variables intentado reducir al máximo el error de validación cruzada RMSE. Repita el punto 3 identificando ahora las variables más importantes del mejor modelo posible. Explique la relación del RMSE frente al MAE del mejor modelo obtenido.**

Para hacer el modelo con las variables seleccionadas manualmente hay que hacer los siguientes ajustes:

```
        Classiffy linear regression > attribute selection method > No atribute selection
> Start
        # buscamos la variable que menos peso tiene y vamos a eliminarla
        Preprocess > Check en la variable y Remove
        # Volvemos a procesar
        Classify > Start
```

Finalmente he conseguido este modelo con ridge 0.1:

```
=== Classifier model (full training set) ===

Linear Regression Model

MORT =

      0.2486 * PREC +
     -0.2345 * JANT +
     -0.1648 * JULT +
     -0.14   * EDUC +
      0.5299 * NONW +
     -0.7873 * HC +
      0.7673 * NOX +
      0.1601 * SO@ +
      0.3987

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.815
Mean absolute error                      0.0884
Root mean squared error                  0.1112
Relative absolute error                 56.5678 % -> va a coincidir con el modelo sin
normalizar
Root relative squared error             57.4047 % -> va a coincidir con el modelo sin
normalizar
Total Number of Instances               60
```

6. Explique las otras métricas que aparecen, además del RMSE y MAE.

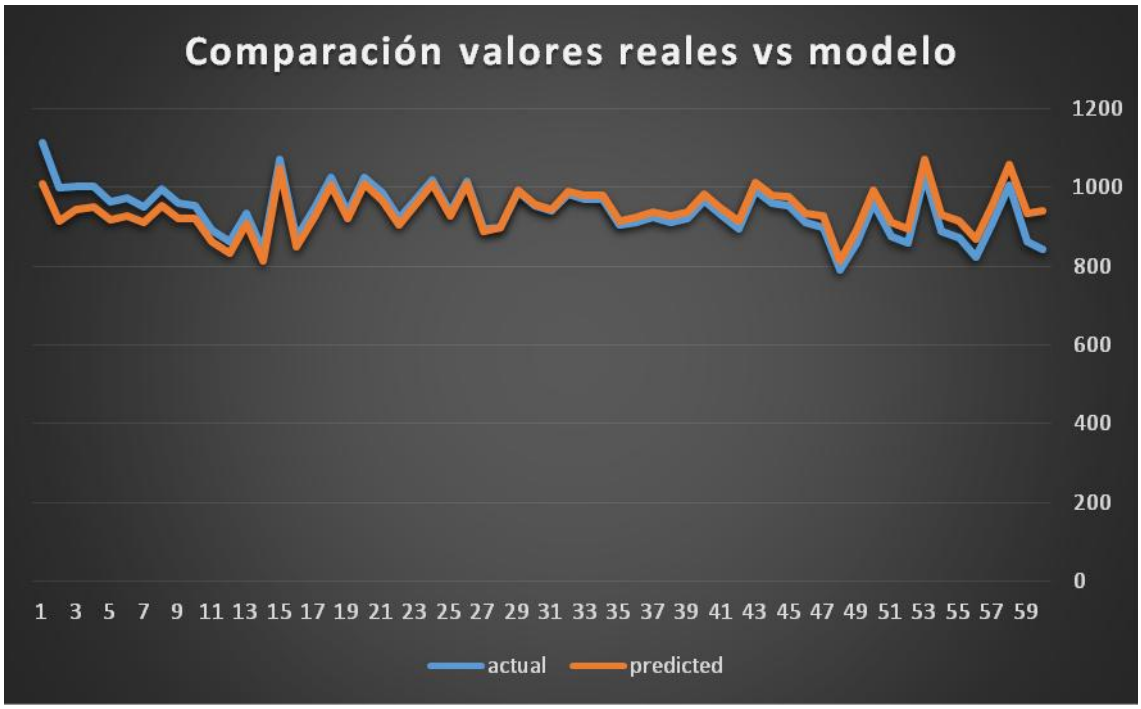
RMSE: Es el error MAE relativizado en tanto por ciento. Por tanto, en el error normalizado y sin normalizar coincide esta métrica.

MAE: Es el error RMS relativizado en tanto por ciento. Por tanto, en el error normalizado y sin normalizar coincide esta métrica.

7. **Extraiga los resultados de predicción creando un modelo con los parámetros del paso 5 pero usando la base de datos de entrenamiento SIN NORMALIZAR, páselos a EXCEL y presente una gráfica con los valores reales y los valores predichos (usando validación cruzada 10) para el mejor modelo lineal.**

Para exportar el modelo a CSV hay que seguir estos pasos:

```
Classify > More options > Output model > Choose > CSV >  casilla blanca > output
file > save > dar un nombre .csv > guardar > ok > VOLVER A CORRER EL MODELO CON
START!
```



8. **Identifique cuales son los países con mayor error de predicción.**

país	inst#	actual	predicted	error
34	4	1113.156	1007.869	-105.287
7	1	997.875	913.63	-84.245
1	1	1003.502	942.91	-60.592
3	3	1001.902	951.74	-50.162
14	2	962.354	917.799	-44.555
32	2	972.464	928.95	-43.514
47	5	950.672	910.139	-40.533
36	6	994.648	954.966	-39.682
39	3	958.839	920.339	-38.5

NOTAS:

Correlación cuanto más alto mejor.
Correlación va entre 0 y 1.
Cuando la correlación es 1 es porque los valores se aproximan a la diagonal de la regresión lineal.