Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO III: Minería de datos

PARTE 03: Regresión Lineal

Introducción:

En este trabajo se abordará una técnica estadística de predicción, utilizada para estudiar la relación entre diferentes variables de tipo numéricas, la Regresión Lineal.

En primera instancia, se introducen técnicas de para explorar los datos e intentar verificar los supuestos que permitan realizar en el análisis de regresión. Luego, se avanzará sobre la estimación de los parámetros y el estudio de la bondad del ajuste de la recta resultante.

Se utilizarán las herramientas de software **R** y **Weka** con el objetivo de resolver problemas de la disciplina, los cuales son una combinación ejercicios clásicos de minería de datos complementados con ejercicios propuestos por el equipo docente.

Consignas:

1. **Regresión lineal simple.** A partir del dataset presentado a continuación, el cual consiste en un pequeño relevamiento del tiempo que demandan las llamadas a servicio técnico de una empresa y la cantidad de unidades de hardware reparadas a partir de la misma, realice las siguientes operaciones:

#	MINUTOS LLAMADA	UNIDADES REPARADAS
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5

7	96	6
8	97	6
9	109	7
10	119	8
11	149	9
12	145	9
13	154	10
14	166	10

a. Realice un *scatter plot* y comente brevemente que observa sobre la relación entre las variables.
b. Ahora, calcule el coeficiente de correlación, es consistente con su hipótesis del punto anterior?
c. Ahora verifique los supuestos de la regresión lineal simple a efectos de determinar si es posible avanzar sobre el análisis de regresión. Qué conclusiones obtiene? ¿Cuál es la importancia de validar estos supuestos?
d. En caso de ser posible, estime los parámetros de la regresión manualmente, utilizando el *método de mínimos cuadrados*.
e. Grafique la recta obtenida sobre el *scatter plot* del primer punto. Que observa a simple vista sobre la bondad del ajuste?
f. Calcule el R^2 y compare los resultados contra lo observado en el punto anterior.
g. Realice las predicciones para los valores [160, 25, 119]. Que observa en este problema en particular?

2. Se cuenta con el dataset *animals.csv* con el registro del peso medio del cerebro y el cuerpo de un grupo de especies de mamíferos. Incorpore tal dataset en **R** y repita el procedimiento de la consigna anterior. Documente las conclusiones.

3. **Regresión lineal múltiple.** Cargue el dataset *machines.csv* en **R**:
    a. Explore las características de los datos, instancias, cantidad de atributos, su significado y el tipo de datos de cada uno.
    b. Realice un gráfico *scatter plot* en busca de la relación entre los diferentes atributos y la performance del CPU y concluya al respecto.
    c. Ahora verifique los supuestos de la regresión lineal múltiple a efectos de determinar si es posible avanzar sobre el análisis de regresión. Qué conclusiones obtiene? ¿Cuál es la diferencia con relación a los supuestos de la regresión lineal múltiple?
    d. Utilizando la herramienta, estime los parámetros de la regresión y grafique la recta ajustada sobre el *scatter plot*.
    e. Concluya acerca de la bondad del error de estimación de la recta ajustada. ¿Cuál es la diferencia con respecto a la regresión lineal simple?
    f. Utilizando la función de regresión resultante, realice la predicción para las siguientes instancias del nuevo modelo bdm-2016 y concluya respecto de esas predicciones:

Adviser	bdm-2016	80	2048	16000	256	4	32
Amdahl	bdm-2016	95	4096	32000	32	8	128
Bti	bdm-2016	188	1024	6000	0	16	32

4. Ahora, explore el dataset *crimen.arff* en **Weka** y realice el procedimiento para el análisis de regresión.
    a. En caso de ser posible, estime los parámetros utilizando una de las herramientas y concluya al respecto.
    b. Realice las estimaciones para las instancias que aparecen en crimen-prediccion.csv. Documente los resultados.

5. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada tp0303-<legajo> que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Referencias sugeridas:

Regression Analysis by Example, Chapter 1, 2 & 3. Samprit Chatterjee and Ali S. Hadi, Ed. Wiley, 2012.

Data Mining: Practical Machine Learning Tools and Techniques
http://www.cs.waikato.ac.nz/ml/weka/

An Introduction to R
https://cran.r-project.org/doc/manuals/r-release/R-intro.html