Aprendizaje Automático
TP4: Regresión Logística y Métodos de Aprendizaje no supervisado

Profesores: J. Gambini - J. Santos

1. El archivo acath.xls contiene datos de pacientes que recurrieron al centro médico con dolor en el pecho. Estos datos fueron extraídos de http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/acath.html.

   Se trata de una muestra de 3504 pacientes para los que se recogieron diversas variables cuyos nombres en la base de datos y descripción es la siguiente:

   - sigdz : variable binaria que toma valores 1 y 0, indicando si el paciente presenta estrechamiento de al menos un 75 % de alguna de las arterias coronarias importante (enf = 1) o no (enf = 0).
   - tvdlm : lo mismo que la anterior pero corresponde a tres arterias con estrechamiento.
   - sex: variable categórica que indica el sexo del paciente, 0 masculino, 1 femenino.
   - age: variable continua que representa la edad en años del individuo.
   - choleste: variable continua que expresa los Mg/dl de colesterol.
   - duracion: variable continua que recoge la duración, en días, de los síntomas de la enfermedad coronaria.

   a) Dividir aleatoriamente el conjunto de datos en dos conjuntos, uno de entrenamiento y uno de prueba.

   b) Clasificar la variable categórica sigdz que indica si el paciente posee o no una enfermedad coronaria, utilizando el modelo de regresión logística y las variables numéricas. Calcular la matriz de confusión.

   c) Utilizando el modelo anterior calcular la probabilidad de que una persona tenga estrechamiento arterial si el colesterol es de 199, la edad es de 60 años y la duración es de 2 días.

   d) Realizar el ejercicio 1*b* diferenciando mujeres y varones, o sea tomando la variable sex como factor. Interpretar los resultados.

e) Realizar el ejercicio $1b$, pero utilizando el método K-NN con $k = 5$. Calcular la matriz de confusión y comparar con el ejercicio $1b$.

f) Utilizando las variables numéricas del conjunto de datos y el método de k-medias, dividir el conjunto de datos en dos grupos, los registros correspondientes a personas que poseen la enfermedad y las que no, ignorando la verdadera clasificación. Evaluar el resultado.

2. Considere 50 textos periodísticos de aproximadamente 500 palabras:

   - 10 pertenecientes al periodista Jorge Fonteveccia, del diario Perfil.
   - 10 pertenecientes al periodista Horacio Verbitsky, del periódico El cohete a la Luna.
   - 10 pertenecientes al periodista Carlos Pagni, del diario La Nación
   - 10 pertenecientes a la periodista Romina Calderaro, del diario Página 12.
   - 10 pertenecientes al periodista Eduardo van der Kooy, del diario Clarín.

   Calcular al menos 6 atributos que caractericen texto. Utilizar k-medias, Agrupación Jerárquica y redes de Kohonen para agrupar los textos del mismo autor. Considerar la posibilidad de agregar o eliminar atributos y agregar o eliminar textos.

   Atributos sugeridos:

   - **Cantidad promedio de palabras por oración** relativa a la cantidad de palabras de todo el texto.
   - **Suma de las frecuencias relativas de las cinco palabras más repetidas**.
   - **Cantidad de palabras diferentes en el texto**, relativa a la cantidad total de palabras.
   - **Cantidad de Conjunciones subordinantes**: En las oraciones con conjunciones subordinantes, existe una oración principal y una oración secundaria que es introducida por la conjunción subordinante y que depende de la principal.
     - Conjunciones subordinantes causales: porque, pues, ya que, puesto que, a causa de, debido a.
     - Conjunciones subordinantes consecutivas o ilativas: luego, conque, así que.
     - Conjunciones subordinantes condicionales: si.
     - Conjunciones subordinantes finales: para que, a fin de que.
     - Conjunciones subordinantes comparativas: como, que.
     - Conjunciones subordinantes concesivas: aunque, aun cuando, si bien.
     - Conjunciones subordinantes completivas: que, si.
   - **Cantidad de conjunciones coordinantes**: Unen palabras u oraciones que tengan la misma jerarquía.
     - ni, y, o, o bien, pero aunque, no obstante, sin embargo, sino, por el contrario.

- **Frecuencia relativa de artículos determinados**: La, el, los, las.
- **Frecuencia relativa de artículos indeterminados**: un, una unos, unas.
- **Cantidad de adverbios que terminen en mente**.