Práctica 4: Métodos Lineales de Regresión y Clasificación

`Almacenes y Minería de Datos`

Grado en Ingeniería en Informática
Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
Escuela de Ingeniería y Arquitectura

Jordi Bernad, Sergio Ilarri y Daniel Pons

1. Objetivos

En esta práctica se pretende que el alumnado analice varios conjuntos de datos utilizando la regresión lineal y la regresión logística:

1. Aplicar la regresión lineal a un conjunto de datos interpretando los resultados obtenidos.
2. Utilizar la regresión lineal para la selección de predictores y el análisis de la varianza.
3. Clasificar un conjunto de datos utilizando la regresión logística.

Los conjuntos de datos que se utilizarán en la práctica se pueden obtener en Moodle como material adjunto a este enunciado.

2. Contexto de la práctica

A continuación se plantean dos problemas: uno relacionado con la calidad de la cerveza, cuyo objetivo es determinar los componentes que tienen más infuencia en su actividad antioxidante, y otro que representa un problema de clasificación.

2.1. Problema 1: calidad de la cerveza

La presencia de elementos oxidantes en la cerveza provoca que su sabor varíe a lo largo del tiempo de almacenaje, perdiendo propiedades organolépticas.

Con el conjunto de datos *lagerdata.csv*[1], adjunto al enunciado de esta práctica, se pretende estudiar la actividad antioxidante de tres compuestos químicos contenidos en la cerveza que evitan la oxidación: el contenido de compuestos fenólicos (tpc), el contenido de compuestos de melanoidina (ma), y el contenido de dióxido de sulfuro (tso2). La actividad antioxidante se ha medido en cinco ensayos diferentes que se corresponden a las columnas del conjunto de datos: "dsa", "asa", "orac", "rp" y "mca".

Comprobar, para cada uno de los cinco ensayos de la actividad antioxidante, qué componentes de tpc, ma y tso2 tienen más incidiencia o si alguno no presenta indicios de influir en alguno de los ensayos.

Material a entregar En primer lugar se adjuntará a la memoria la interpretación de todos los valores marcados en rojo o bajo las columnas marcadas en rojo de la Figura 1. En segundo lugar se explicará en la memoria qué proceso se ha utilizado para averiguar cuáles de los tres predictores influyen en la actividad antioxidante para cada uno de los cinco ensayos, y los resultados obtenidos. Se utilizarán dos páginas, como máximo, para el conjunto de esta problema.

```
> summary(lm(dsa~tpc+ma+tso2, data=beer.data))
 Coefficients :
             Estimate     Std. Error    t value   Pr(>|t|)
(Intercept) −0.0418869    0.0824789    −0.508    0.6147
tpc          0.0034765    0.0004144     8.390    5.42e−10 ***
ma           0.0032586    0.0065274     0.499    0.6207
tso2         0.0036574    0.0020140     1.816    0.0777 .

Residual standard error: 0.09466 on 36 degrees of freedom
Multiple R−squared: 0.7181,   Adjusted R−squared: 0.6946
F−statistic: 30.56 on 3 and 36 DF, p−value: 5.3e-10
```

Figura 1: Datos a interpretar de la regresión lineal: dsa~tpc+ma+tso2

2.2. Problema 2: un problema de clasificación

El inspector Augusto Puerrot está harto de llegar siempre tarde adonde están los "*malos*". Vive cerca de una zona costera en la que actúan varias bandas de narcotraficantes. Habitualmente, los narcos desembarcan la droga en una playa bajo la jurisdicción del inspector Augusto Puerrot. Si los narcos creen que han sido descubiertos mientras están descargando la droga, rápidamente se ponen a la fuga tirando al mar los fardos que les quedan en la embarcación utilizada para realizar las entregas. Las corrientes en esta zona del mar son caprichosas, y

[1] Obtenido del artículo: H. Zhao et al, *Assessment of endogenous antioxidative compounds and antioxidant activities of lager beers*, Journal of Food and Agriculture, vol. 93, n. 4, 2013

los fardos lanzados acaban en una de tres calas que hay a varias decenas de kilómetros de distancia. Los narcos tienen sicarios esperando en las tres calas para que en el momento que llega un fardo, recogerlo. Desgraciadamente, los medios del inspector Augusto Puerrot no son comparables a los de los narcotraficantes; cuando sucede un desembarcado fallido y se tiran al mar fardos de droga, el inspector Augusto Puerrot y su único subordinado se dirigen a una de las tres calas para requisar la droga que llegue. Pero tienen que elegir bien, puesto que si se dirigen a la cala equivocada, no tienen tiempo para reaccionar, y los narcos, con gente esperando los fardos en las tres calas, recogen inpunemente la droga.

La falta de medios agudiza el ingenio, y al inspector Augusto Puerrot se le ocurrió un sencillo experimento para predecir lo mejor posible a qué cala llegarán los fardos. Junto con su subordinado, crearon 600 fardos con plásticos que recogieron de las calles, y los lanzaron al mar a lo largo de toda la playa. Cada fardo se identificaba con un número y se apuntaban las coordenadas GPS desde donde había sido lanzado y la cala a la que había llegado (cala 0, 1 ó 2). Tras realizar el experimento, el inspector Augusto Puerrot inmediatamente sacó una sencila conclusión: los fardos tirados al sur de la playa casi siempre llegan a la cala 0; los lanzados en el centro de la playa llegan a la cala 1; y al norte, a la cala 2 (ver Figura 2).

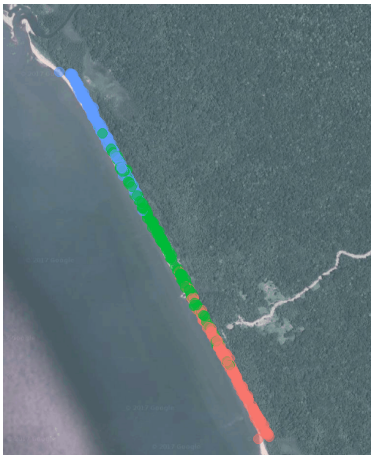

Figura 2: Puntos de lanzamientos de los fardos y cala a la que llegó: rojo, cala 0; verde, cala 1; azul, cala 2

Sin embargo, había dos zonas de la playa donde no estaba tan claro a qué cala llegaban los fardos. Tal y como nos deplazábamos desde la zona sur de la playa y llegamos a la zona media, algunos paquetes llegaban a la cala 0 y otros a la cala 1. De forma similar, al pasar de la zona media de la playa a la zona norte, los paquetes llegaban a las calas 1 ó 2.

2.3. Parte 1: ¿A dónde va el fardo?

El subordinado tenía una amiga aficionada al análisis de datos, Sada Tanita, y alguna vez le había oido hablar de que era capar de hacer predicciones con datos, o algo así. Augusto Puerrot y su subordinado fueron a visitar a Sada Tanita para preguntarle si podría ayudarles a predecir con más precisión a qué cala llegaría un fardo lanzado en las zonas que tenían un pronóstico más complicado. Cuando Sada Tanita vio el conjunto de datos que Augusto Puerrot y su subordinado habían conseguido hacer con tanto trabajo, comprendió enseguida que el problema se podría resolver fácilmente utilizando una regresión logística. Sada Tanita realizó los siguientes cálculos:

- La playa se modeló mediante un segmento que unía el punto más al norte de los almacenados en el conjunto de datos (coordenada y mayor) con el punto más al sur (coordena y más pequeña). Sada Tanita calculó dos puntos de esta recta que separaban la playa en tres tramos: zona norte, zona centro, zona sur.
- Con estos dos puntos, generó una función que, dada unas coordenadas GPS de la playa, devolvía un número, 0, 1 ó 2, con la cala a la que llegaría el fardo con mayor probabilidad
- También creó una función que, dadas unas coordenadas GPS, devolvía un vector con las probabilidades de que el fardo llegase a cada una de las tres calas.

El primer objetivo del lector de este enunciado es implementar código en R que analice el conjunto de datos *datacala.csv*, adjunto al enuncio de la práctica, y calcule los dos puntos que dividen la playa en tres tramos junto con las dos funciones mencionadas anterioremente. Una vez que tengamos las funciones, se hallará la matriz de confusión para los datos de entrenamiento, esto es, una matriz:

		Real		
		sur	centro	norte
Pred	sur	n_{11}	n_{12}	n_{13}
	centro	n_{21}	n_{22}	n_{23}
	norte	n_{31}	n_{32}	n_{33}

donde n_{ij} son el número de puntos que se han clasificado como i, pero que realmente son j.

2.4. Parte 2: ¿Podría hacerse con menos fardos?

Sada Tanita, además de tener conocimientos de Minería de Datos, era una afamada activista ecológica. Cuando le contaron que habían lanzado 600 fardos de plástico al mar para hacer el experimento, se imaginó toda esa basura flotando y comentó: "*Bastante basura hay en el mar para que vosotros andéis echando*

una poca más". Les hizo ver al inspector y al subordinado que no eran necesario tirar al mar tanto fardo del siguiente modo:

- Seguimos modelizando la playa mediante la recta explicada anteriormente: la recta que pasa por los puntos de mayor y menor latitud del conjunto de datos con la información de todos los fardos.
- Dado un porcentaje $0 < p < 1$, se eligen al azar $p * 600$ fardos del total.
- Con esos $p * 600$ fardos se calculan los dos puntos que dividen la playa en tres tramos
- Si repetimos este proceso 1000 veces, obtendremos 2000 puntos distintos, 1000 puntos que separan la zona norte, y otros 1000 para la zona sur.
- De los 1000 puntos calculados para la zona norte (sur), elegimos los 100 que están más lejos del punto calculado para separar la zona norte (sur) usando todos los datos, esto es, los peores comparados con el experimento con 600 fardos. Si estas 100 peores distancias están en media por debajo de 100 metros, se considerará que el experimento se podría haber hecho con $p * 600$ fardos[2].
- Del mismo modo, una vez calculado los puntos que separan la playa para los 1000 distintos conjuntos de datos con $p * 600$ fardos, podríamos hallar el porcentaje de aciertos con respecto al total del conjunto de datos. En este caso, habríamos calculado 1000 porcentajes. Si en media los peores 100 porcentajes no difieren en más de un punto porcentual del porcentaje de acierto obtenido usando todos los fardos, también podremos considerar que el experimento se podría haber llevado a cabo con $p * 600$ fardos.

El segundo objetivo del lector es probar con valores de p =0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, y 0,9, para calcular qué porcentaje de fardos se podrían haber lanzado para obtener un modelo similar al obtenido lanzando 600 fardos.

Implementación con R Tenéis en moodle un enlace a varias chuletas con instrucciones básicas de R. Para resolver esta práctica son muy útiles instrucciones como: `rbind` y `cbind` para combinar por filas o columnas vectores y/o matrices; `install.packages` para instalar nuevos paquetes; `library` para cargar paquetes; `sample` para sacar muestras; *plot* para hacer gráficos con puntos y líneas; `points` para añadir puntos a un gráfico; `abline`, añadir una recta a un gráfico. Todas estás instrucciones las tenéis descritas en las chuletas o tecleando `help("instruccion")` en la línea de comando.

Se pueden definir funciones mediante `function` y `return(...)`. Un ejemplo de función que os será de utilidad para resolver los problemas es:

[2]Para calcular la distancia entre dos coordenadas GPS en metros, se puede usar la función `distHaversine` del paquete `geoshpere`.

```
sampledataset = function(dataset, perc=1) {
  loc.data.2 = dataset
  if (perc != 1) {
    ind.sample = sample(1:nrow(loc.data.2), perc*nrow(loc.data.2))
    loc.data.2 = loc.data.2[ind.sample,]
  }
  return(loc.data.2)
}
```

Con este código, definimos una función que se llama `sampledataset`, que tiene dos paráme-tros, el segundo con un valor por defecto de 1. Observad que en R no se declaran los tipos de los parámetros: `dataset` es el nombre del primer parámetro; y `perc`, el del segundo. Esta función, dado un data frame o matriz, `dataset`, con n filas, devuelve otro data frame o matriz con `perc`$*n$ filas elegidas al azar de `dataset`.

Material a entregar Se entregarán los siguientes ficheros de código escrito en R:

- Un fichero con código para mostrar por pantalla: las coordenadas del punto más al norte y al sur de la playa; las coordenadas de los puntos que separan la zona norte y sur; la tabla de confusión que se explica anteriormente; la cala a la que se predice que llegan los fardos lanzados en las coordenadas (9,558359, 1,1), (9,564329, 1,089) y (9,568671, 1,081); y las probabilidades de que un fardo lanzado desde uno de los tres puntos anteriores alcance cada una de las tres calas. En este fichero se incluirá todo el código necesario para calcular estos datos.
- Un fichero con código que muestre por pantalla la medias de las peores distancias y las medias de los peores porcentajes de acierto para los distintos valores de p descritos en el enunciado.

Se explicará razonadamente en la memoria todas las decisiones tomadas a la hora de calcular los puntos de separación de la zona norte y sur. Se valorará que se incluya un gráfico donde aparezca una línea simulando la playa, y los dos puntos de separación de la playa. ¿En qué país es muy probable que vivan los protagonistas de esta historia? (Dos páginas máximo para este ejercicio.)

3. Entrega de la práctica

La práctica se realizará en equipos de tres personas (salvo que existan problemas logísticos, que deberán comentarse y resolverse previamente con el profesor). Cuando se finalice la práctica se debe entregar un fichero .zip denominado *pN_NIP1_NIP2_NIP3.zip* (donde NIP1, NIP2 y NIP3 son los NIP de los autores de la práctica, con NIP1 $<$ NIP2 $<$ NIP3, y en pN la N representa el número de práctica) con el siguiente contenido:

1. Un fichero de texto denominado *autores.txt* que contendrá el NIP, los apellidos y el nombre de los autores de la práctica en las primeras líneas del fichero.

2. Un fichero de texto o PDF denominado informe.txt o informe.pdf, que contendrá la memoria de prácticas (respuestas a las cuestiones planteadas, esquemas y código desarollados, etc., según el caso). En dicho fichero se identificará claramente al comienzo los componentes del grupo de prácticas (nombre y apellidos y NIP de cada uno) y el número de práctica. El informe de prácticas debe contener al final un apartado de conclusiones personales que incluirá, entre otras cosas, información sobre el tiempo invertido por cada miembro del grupo de prácticas en la realización de la misma.

3. Todos los fuentes y programas de prueba desarrollados, si es el caso.

Al descomprimir el fichero .zip se deben extraer los ficheros y directorios necesarios en el directorio pN_NIP1_NIP2_NIP3. Es importante seguir las convenciones de nombrado y la estructura de ficheros y directorios descrita.

Para la entrega del fichero .zip, se utilizará Moodle 2 del Anillo Digital Docente de la Universidad de Zaragoza. La fecha límite de entrega es el día anterior al de la siguiente sesión de prácticas en el laboratorio a las 23:59.

4. Recomendaciones

Una vez realizadas las prácticas y entregadas estas, cada grupo debe presentárselas al profesorado de prácticas en la siguiente sesión de prácticas. Al realizar la presentación el profesorado le formulará cuestiones sobre las decisiones que ha tomado. La práctica debe entregarse en los términos indicados anteriormente y debe ser original (no debe haber sido copiada). Se considerarán los siguientes aspectos:

- La consecución de los objetivos planteados y la adecuada realización de las tareas correspondientes.

- La justificación de las afirmaciones incluidas en el informe.

- La inclusión de referencias a fuentes de información, si procede.

- La metodología para llevar a cabo la búsqueda de información necesaria para la realización de la práctica.

- La estructura y presentación del informe elaborado (completitud del trabajo, precisión, explicaciones adecuadas y completas, justificación de las decisiones tomadas, referencias adecuadas, coherencia de discurso, sin errores tipográficos y ortográficos, etc.).

Es importante extraer conclusiones: resultados obtenidos, indagaciones realizadas, diferencias observadas entre las distintas alternativas y/o herramientas, dificultades encontradas, valoración de los aprendizajes o mejoras de habilidades conseguidas, opinión personal, etc. Se recomienda destacar especialmente aquellos aspectos que muestren la capacidad de indagación, búsqueda de información, autonomía, y curiosidad personal que se hayan desarrollado.