Introducción al Aprendizaje Estadístico - Actividades de Clase - Unidad 5
Lucía Babino - Paula Spano

Variedades de hongos:

En un bosque de Bariloche hay dos variedades de hongos, que identificaremos como la variedad I y variedad II. En el archivo `hongos_clasificados.txt`, disponible para su descarga en el campus, se encuentran $n = 500$ registros correspondientes a la altura y variedad de cada uno de los hongos examinados. El objetivo es poder clasificar un nuevo hongo de este bosque con solo medir su altura.

1. Realizar un gráfico que pueda dar información sobre la relación entre la altura de los hongos y su variedad.

2. Crear una nueva variable dentro del data frame donde guardó los datos que valga 0 si el hongo es de la variedad 1 y que valga 1 si es de la variedad 2. Inspeccione el nuevo data frame para chequear que esté bien generada.

3. Como primera propuesta, clasificar a la variedad de un nuevo hongo a partir de su altura, utilizando la información que brindan los `k = 8` vecinos más cercanos:

   a) Si se toma un hongo del bosque y se obtiene una altura de 5,2 cm, ¿en qué variedad debe clasificarse?
   (*Hint: para responder a esta pregunta, calcular las proporciones de hongos de las variedades I y II entre los $k = 8$ vecinos más cercanos.*)

   b) Si ahora se considera un hongo cuya altura es 6 cm, ¿en qué variedad debe clasificarse?

4. Ahora clasificaremos utilizando el método de proporciones locales con ancho de ventana $h = 0{,}1$.

   a) Clasificar a un hongo de altura 5,2 cm y a otro de altura 6 cm. Recordar que para ello, se deben calcular las proporciones de hongos de las variedades I y II entre aquellos cuya altura diste de 5,2 ó 6, según el caso, en menos de 0,1.

   b) En cada caso, ¿se utilizan más, menos o igual cantidad de hongos que cuando implementó vecinos más cercanos? En base a eso, ¿alrededor de qué valor de altura (5.2 ó 6) cree que habrá mayor concentración de hongos? ¿qué gráfico podría realizar para corroborarlo?

5. Implementar dos funciones, `clas_knn(x, y, x_nuevo, k)` y `clas_prop_loc(x, y, x_nuevo, h)` que en base a un conjunto de valores `x` y sus correspondientes clases `y`, clasifique a una nueva observación con valor `x_nuevo` mediante el método de `k` vecinos más cercanos en el primer caso y de proporciones locales con ancho de ventana `h` en el segundo. Utilizar dichas funciones para realizar las mismas clasificaciones de los ítems anteriores y así corroborar que están bien programadas.

6. Clasificar a un hongo cuya altura es 5.2 cm utilizando el método generativo. Para ello:

   a) Estimar la función de densidad de las alturas de los hongos de la variedad I, utilizando el núcleo normal con la ventana óptima de validación cruzada y evaluarla en 5,2.

   b) Repetir el ítem anterior para la variedad II.

   c) Obtener la proporción muestral de hongos en cada variedad.

   d) Clasificar según la regla que determina el método.

7. Definir una función `clas_gen(x, y, x_nuevo, h0, h1)`, que aplique el método generativo para clasificar a un valor `x_nuevo`, donde `(x,y)` son los datos del conjunto de entrenamiento y `h0` y `h1` son las ventanas utilizadas para las estimaciones de las funciones de densidad con núcleo normal.
   Utilizar dicha función para realizar la misma clasificación del ítem anterior usando los mismos anchos de ventana h_0 y h_1 para corroborar que esté bien programada.

8. Calcular el valor de la función objetivo de Validación Cruzada Leave One Out (LOOCV), $CV(h_0, h_1)$, para las ventanas usadas en el ítem 6. Para ello, primero implementar una función `loocv(h0, h1)` que calcule el error de validación cruzada para un (h_0, h_1) dado. Hint.: adaptar la función `loocv(h)` que definió en el ítem 19.d de la guía de Actividades 2.

9. Hallar las ventanas óptimas $h_{0,opt}$ y $h_{1,opt}$ mediante el método de LOOCV realizando una búsqueda en una grilla de valores de 0,01 a 1 con paso 0,05 para h_0 y otra de valores de 0,03 a 1,73 con paso 0,15 para h_1. Recordar que el 0 corresponde a la variedad 1 y el 1 a la variedad 2.
   Calcular el valor de $CV(h_{0,opt}, h_{1,opt})$ y compararlo con el de $CV(h_0, h_1)$ con (h_0, h_1) las ventanas del ítem 6. ¿Cuál es menor?