Universidad Autónoma de Madrid
Escuela Politécnica Superior

Aprendizaje Automático: Teoría y Aplicaciones:

Práctica 1: Auditoría de Datos (Parte 1)

Luis Sánchez Calvo y Jaime Sánchez Fernández

27 de septiembre de 2019

Resumen

Una de las partes más costosa en aprendizaje automático es preparar los datos para ser procesados posteriormente. Lo primero que tenemos que hacer es entender el problema al que nos vamos a enfrentar, invirtiendo tiempo en estudiar/visualizar la base de datos que nos facilita el cliente/colaborador.

Para pre-procesar los datos se ha hecho un estudio de las variables, se ha realizado un proceso manual de selección de estas, y por último se ha utilizado la herramienta `One Hot Encoder` para reescribir sus valores de forma que se puedan utilizar en diversos algoritmos de ML.

Para realizar esta práctica hemos utilizado el conjunto de datos `Mushroom Data Set`[1].

Ejercicios

- Descripción de las variables y valores estadísticos:

  El conjunto de datos `Mushroom Data Set` consta de 8124 observaciones de diferentes setas. A todas estas observaciones se les ha medido 22 variables diferentes (todas ellas categóricas) y se les ha clasificado en 2 tipos diferentes: comestibles o venenosas (*Poisonous* o *Edible*). Hay algunos valores faltantes (*missing values*) que se encuentran todos en la variable número 11, raíz del tallo de la seta (**Stalk root**) y se representan con el símbolo ?. Las variables que se han estudiado son:

  - **Cap shape**: Forma de la copa de la seta. Toma los valores b si tiene forma de campana, c si es cónica, x si es convexa, f si es plana, b si tiene bultos o s si esta hundida.
  - **Cap surface**: Textura de la superficie de la copa de la seta. Toma los valores f si es fibrosa, g si tiene surcos, y si es escamosa o s en el caso de que sea lisa.
  - **Cap color**: Color de la copa de la seta. Se distinguen los colores marrón, béis, canela, gris, verde, rosa, púrpura, rojo, blanco y amarillo, con las letras n, b, c, g, r, p, u, e, w y y respectivamente.

[1] `https://archive.ics.uci.edu/ml/datasets/mushroom`

- **Bruises?**: Esta variable toma el valor t si la seta tiene magulladuras y f si no.
- **Odor**: Aroma de la seta. Toma los valores a, l, c, y, f, m, n, p o s si la seta huele a almendras, anís, creosota, pescado, asquerosa, mohoso, no tiene olor, fuerte o picante respectivamente.
- **Gill attachment**: Adherencia de las branquias (malla) de la seta. Puede ser adjunto, descendente, libre o con muescas, cuyos valores son a, d, f y n respectivamente.
- **Gill spacing**: Espacio entre las branquias de la seta. Esta variable toma los valores c si las branquias son cercanas, w si están abarrotadas o d si están distantes.
- **Gill color**: Color de las branquias de la seta. Estas pueden ser negras, marrones, béis, chocolateadas, grises, verdes, naranjas, rosas, púrpuras, rojas, blancas o amarillas. Estos colores se representan con las letras k, n, b, h, g, r, o, p, u, e, w e y respectivamente.
- **Stalk shape**: Forma del tallo. Puede ser agrandado (e) o estrecho (t).
- **Stalk root**: Forma de la raíz del tallo. Se distinguen las formas b (bulbosas), c (aporradas), u (acopadas), e (homogeneas), z (rizomorfas), r (arraigadas) o ? en caso faltante (*missing value*).
- **Stalk surface above ring**: Textura de la superficie del tallo por encima del anillo. Puede ser fibroso (f), escamoso (y), sedoso (k) o liso (s).
- **Stalk surface below ring**: Textura de la superficie del tallo por debajo del anillo. Puede ser fibroso (f), escamoso (y), sedoso (k) o liso (s).
- **Stalk color above ring**: Color del tallo por encima del anillo. Se distinguen los colores marrón, béis, canela, gris, naranja, rosa, rojo, blanco y amarillo, y se representan con las letras n, b, c, g, o, p, e, w y y respectivamente.
- **Stalk color below ring**: Color del tallo por debajo del anillo. Se distinguen los colores marrón, béis, canela, gris, naranja, rosa, rojo, blanco y amarillo, y se representan con las letras n, b, c, g, o, p, e, w y y respectivamente.
- **Veil type**: Tipo de velo de la seta. Puede ser parcial (p) o total (u).
- **Veil color**: Color del velo de la seta. Se distinguen los colores marrón, naranja, blanco y amarillo, con las letras n, o, w, y.
- **Ring number**: Número de anillos de la seta. Puede tener uno (o), dos (t) o ninguno (n).
- **Ring type**: Tipo de anillo de la seta. Puede ser en forma de telaraña (c), evanescente (e), ardiente (f), grande (l), colgante (p), revestido (r), zonal (z) o, en el caso de que no tenga, n.
- **Spore print color**: Color de las esporas. Se distinguen los mismos colores que en la variable **Stalk color above ring** con los mismos valores.
- **Population**: Población. Puede ser abundante (a), agrupada (c), numerosa (n), dispersa (s), variada (v) o solitaria (y).
- **Habitat**: Habitat de la seta. La seta puede vivir en un pasto (g), con hojas (l), en un prado (m), en un camino (p), en zona urbana (u), en la basura (w) o en un bosque (w).

No tiene sentido calcular los estadísticos de estas variables (salvo, a lo mejor, la moda) ya que son variables categóricas y carecen de un orden.

🗁 Describe y realiza modificaciones en la base datos si lo consideras necesario.

Se ha modificado el documento `agaricus-lepiota.data` y se ha añadido un encabezado para favorecer la comprensión de los datos. En este encabezado se ha escrito el nombre de las variables.

Además, tras analizar cuales son las mejores variables para clasificar las setas en *Poisonous* o *Edible* (esto se verá con más detenimiento en los siguientes puntos), se ha reducido el número de variables a una (**Odor**). Tras esto se ha utilizado la herramienta `One Hot Encoder` para transformar los valores categóricos de las variables a valores binarios.

🗁 Estudia si es necesario normalizar los datos y cómo lo harías.

No se han normalizado los datos ya que se está trabajando con variables categóricas que carecen de un orden.

🗁 Detección de valores extremos (outliers).

Al igual que antes, cómo estamos trabajando con variables categóricas, no tiene sentido hablar de valores extremos.

🗁 Detección de valores perdidos (*missing values*) y descripción de cómo actuarías para solventar el problema.

La única variable que contiene valores faltantes es la variable **Stalk root**. Como podemos ver en la Figura 1, la gran mayoría de las setas que tienen esta característica faltante son, en su mayor parte, aquellas en las que la variable **Gill color** toma el color marrón (*b*). Por lo tanto, para estimar el valor faltante ? de la variable **Stalk root** calcularía cual es la forma del tallo más popular (moda) de entre todas las setas marrones y lo extrapolaría.

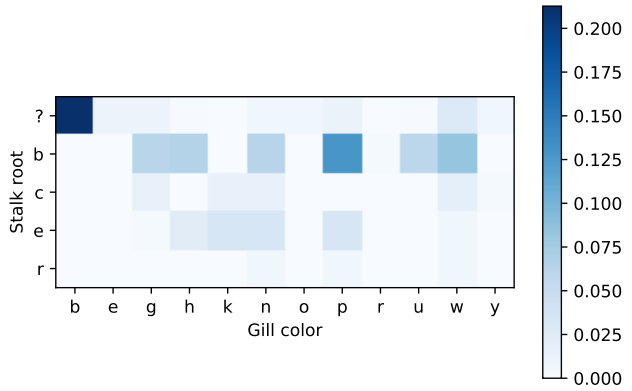

Figura 1: Tabla de contingencia de las variables **Stalk root** y **Gill color**.

- Buscar relaciones entre las variables predictoras y las variables predictoras y la clase.

  Para buscar estas posibles relaciones entre las variables predictoras, creamos las tablas de contingencia de los predictores 2 a 2. Sin embargo, para estudiar las relaciones entre las variables predictoras y la clase (*Poisonous* o *Edible*) generamos los gráficos de barras de las variables, mostrando en cada caso la proporción de setas que son comestibles y venenosas.

  Como puede verse en la Figura 2[2], las variables **Bruises** y **Cap surface** están altamente correladas, por lo que alguna de ellas podría omitirse si una gran pérdida de información. Por otro lado, puede observarse que la variable **Odor** parece ser altamente significativa para predecir si la seta es comestible o venenosa.

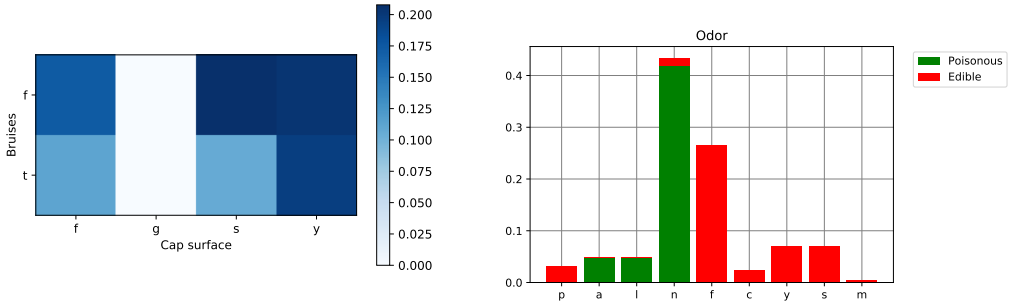

Figura 2: A la izquierda la tabla de contingencia de las variables **Bruises** y **Cap surface**. A la derecha el gráfico de barras de la variable **Odor**, donde el verde representa la proporción de setas comestibles y el rojo lo propio con las venenosas.

- Detecta, si hubiera, falsos predictores.

  No parece que halla falsos predictores (ya que ninguna de las variables predictoras predice bien la clase de la seta, salvo **Odor** que no parece ser un falso predictor), y si los hubiera seguramente necesitaríamos algunos conocimientos sobre micología que, sin lugar a dudas, desconocemos.

- Estudia si fuera conveniente segmentar alguna de las variables.

  Parece bastante conveniente segmentar la variable **Veil type**, ya que, al tener todas las setas observadas el mismo tipo de velo, no aporta ninguna información. Ver Figura 3.

- Estudia si fuera conveniente crear nuevas variables sintéticas basada en las variables originales.

  Podríamos valorar la creación de alguna variable sintética a partir de las 22 anteriores, aunque no parece que sea de gran ayuda ya que la variable **Odor** ya clasifica bien los datos por si sola (como puede verse en la Figura 2).

[2]Aquí solo mostramos algunos de los gráficos de barras y tablas de contingencia más releventes. En el notebook de Jupyter pueden verse el resto.

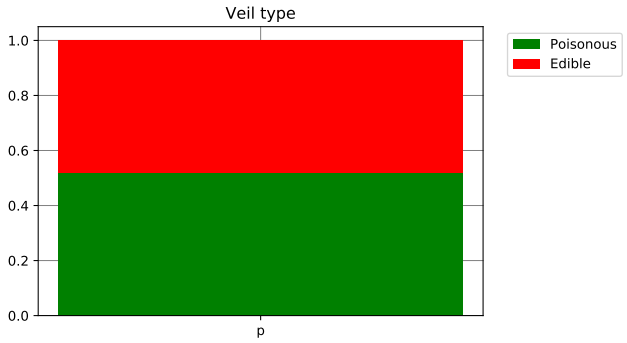

Figura 3: Gráfico de barras de la variable **Veil type**. Al tener todas las setas observadas el mismo tipo de velo, esta gráfica no aporta información.