Adquisición de datos de aprendizaje

Como sucede en todos los aprendizajes, tenemos que disponer de datos. La máquina tendrá que ser capaz de distinguir entre una cereza y un albaricoque a partir del peso y el tamaño de cada fruta. La primera opción que se nos ocurre es ir a una frutería, tomar las medidas necesarias para cada fruta y listarlas en un documento. Este trabajo es bastante repetitivo.

Sin embargo, hay una segunda opción. Como seguro que ya sabe, las características de las frutas están «normalizadas». Hay normas para definir el tamaño y el peso mínimo y máximo de una cereza, y lo mismo para los albaricoques. Por lo tanto, vamos a formar nuestro conjunto de datos a partir de estas normas y generando frutas de manera aleatoria.

Para conseguir estas normas, solo hay que hacer una sencilla búsqueda en Internet. Así hemos podido completar las tablas que aparecen a continuación:

Tras consultar algunas páginas web, hemos creado las siguientes tablas:

**Las cerezas**:

| **Diámetro mínimo (mm)** | **Diámetro máximo (mm)** | **Peso mínimo (g)** | **Peso máximo (g)** |
| --- | --- | --- | --- |
| 17 | 19 | 1 | 5 |
| 20 | 21 | 5 | 6 |
| 22 | 23 | 6 | 7 |
| 24 | 25 | 7 | 8,5 |
| 26 | 27 | 8,5 | 10 |
| 28 | 29 | 10 | 11,5 |

**Los albaricoques**:

| **Diámetro mínimo (mm)** | **Diámetro máximo (mm)** | **Peso medio (g)** |
| --- | --- | --- |
| 35 | 39 | 27 |
| 40 | 44 | 41 |
| 45 | 49 | 54 |
| 50 | 54 | 74 |
| 55 | 59 | 100 |

Un pequeño script de Python para ayudarnos

Vamos a crear un script de Python que nos permitirá generar una cantidad de cerezas y albaricoques en función de sus características propias.

Después de crear un proyecto nuevo e instalar los módulos Pandas, Matplotlib y Scikit-Learn, le invitamos a introducir estas líneas de programa en un archivo de script nuevo al que llamaremos generacionFrutas:

#---- IMPORTAR MÓDULOS --

import random

import pandas as pnd

#---- CARACTERÍSTICAS------

#CEREZAS

caracteristicasCerezas =

[[17,19,1,5],[20,21,5,6],[22,23,6,7],[24,25,7,8.5],[26,27,8.5,10],

[28,29,10,11.5]]

#ALBARICOQUES

caracteristicasAlbaricoques =

[[40,44,41],[45,49,54],[50,54,74],[55,59,100]]

Como puede comprobar, hemos creado dos tablas. Una contiene las características de las cerezas y, la otra, las características de los albaricoques. Cada tabla contiene subtablas que corresponden a cada categoría de fruta.

Así, la tabla [40,44,41] presente en la tabla caracteristicaAlbaricoque corresponde a una característica (calibre) que contempla un diámetro mínimo de 35 mm, un diámetro máximo de 39 mm y un peso medio de 27 g.

A continuación podemos definir la cantidad de frutas que hay que generar:

cantidadObservaciones = 2000

Generamos 2000 cerezas y 2000 albaricoques.

#Generación de las cerezas

cerezas = []

random.seed()

for iteration in range(cantidadObservaciones):

   #elección al azar de una característica

   cereza = random.choice(caracteristicasCerezas)

   #Generación de un diámetro

   diametro = round(random.uniform(cereza[0], cereza[1]),2)

   #Generación de un peso

   peso = round(random.uniform(cereza[2], cereza[3]),2)

   print ("Cereza "+str(iteration)+" "+str(cereza)+" :

"+str(diametro)+" - "+str(peso))

   cerezas.append([diametro,peso])

#Generación de los albaricoques

albaricoques = []

random.seed()

for iteration in range(cantidadObservaciones):

   #elección al azar de una característica

   albaricoque = random.choice(caracteristicasAlbaricoques)

   #Generación de un diámetro

   diametro = round(random.uniform(albaricoque[0], albaricoque[1]),2)

   #Generación de un peso

   limiteMinPeso = albaricoque[2] / 1.10

   limiteMaxPeso = albaricoque[2] \* 1.10

   peso = round(random.uniform(limiteMinPeso, limiteMaxPeso),2)

   print ("Albaricoque "+str(iteration)+" "+str(albaricoque)+" :

"+str(diametro)+" - "+str(peso))

   albaricoques.append([diametro,peso])

Se elige una característica al azar para cada cereza y cada albaricoque. A continuación se recuperan los diámetros mínimo y máximo de esta característica, y el diámetro se genera de manera aleatoria dentro de este rango.

Procedemos de la misma manera para el peso de la cereza. Como solo tenemos un peso medio para el albaricoque como característica, elegimos crear aleatoriamente un peso comprendido entre el peso medio - 10 % y el peso medio + 10 %. Luego, el diámetro y el peso se guardan en una tabla.

Finalmente, concatenamos las tablas de cerezas y de albaricoques y mezclamos el resultado.

Ahora ya hemos formado nuestro conjunto de observaciones. Podemos guardarlo en formato CSV con ayuda del módulo Pandas y de la función to\_csv del Dataframe.

#Constitución de las observaciones

frutas = cerezas+albaricoques

print(frutas)

#Mezcla de las observaciones

random.shuffle(frutas)

dataFrame = pnd.DataFrame(frutas)

dataFrame.to\_csv("datas/frutas.csv", index=False,header=False)

Para evitar errores de ejecución, considere la posibilidad de crear el directorio datas en su proyecto antes de ejecutar el código.