Fundamentos de Aprendizaje Automático 2016/2017

PRÁCTICA Nº 0

Objetivo

El objetivo de esta práctica es familiarizarse con el lenguaje de programación Python, ya que será el lenguaje de programación a utilizar a lo largo de las prácticas. Además, se deberá implementar parte de la clase *Datos* que se utilizará para gestionar los diferentes conjuntos de datos con los que probar los algoritmos de reconocimiento de patrones a implementar a lo largo de las prácticas.

Preliminares

El lenguaje de programación Python es uno de los lenguajes que está adquiriendo mayor popularidad en el campo del reconocimiento de patrones[1]. Dada su versatilidad y sencillez, es el lenguaje que vamos a utilizar a lo largo de las prácticas.

Durante las prácticas emplearemos el lenguaje Python y los Jupyter Notebooks (anteriormente denominados IPython Notebooks). Utilizaremos Anaconda, una distribución totalmente gratuita de Python que incluye más de 300 paquetes como NumPy, para computación científica básica y análisis de datos, y Scikit-learn, paquete con diversos algoritmos de aprendizaje automático. Además, Anaconda incluye también los Jupyter Notebooks, que permiten combinar celdas de texto y código, y serán útiles para combinar las implementaciones de código realizadas y la discusión de los resultados obtenidos.

Anaconda permite instalar Python 2.7 o Python 3.5. Se puede descargar en https://www.continuum.io/. La versión instalada en los laboratorios y a utilizar durante las prácticas es **Python 2.7**.

IMPORTANTE. Para poder usar Anaconda en los laboratorios de prácticas, debe establecerse la variable de entorno PATH como:

```
# export PATH=${PATH}:/opt/anaconda-2.3.0/bin
```

Para arrancar el intérprete de IPython de Anaconda, se debe ejecutar el binario que se encuentra en el directorio de instalación de Anaconda (`/opt/anaconda-2.3.0/`):

```
# /opt/anaconda-2.3.0/bin/ipython
```

Y para arrancar el servidor de Notebooks, éste se debe ejecutar desde el directorio de instalación de la siguiente forma:

```
# cd /opt/anaconda-2.3.0
# ./bin/ipython notebook $HOME
```

[1] http://machinelearningmastery.com/best-programming-language-for-machine-learning/

Las diapositivas disponibles en *Moodle* junto al enunciado de prácticas proporcionan una introducción a Python (variables y tipos de datos, funciones, secuencias de control, ficheros, clases, módulos, paquete NumPy, …).

Estilo de programación

En conveniente (y se valorará) utilizar el estilo de programación recomendado que se puede encontrar en: https://www.python.org/dev/peps/pep-0008/ . El utilizar un estilo común facilita la lectura de código a otros programadores. Los puntos más importantes de esta guía de estilo son (extraído del tutorial de python):

- Utiliza indentación con 4 espacios en vez de tabuladores.
- No escribir líneas de más de 79 caracteres.
- Usar líneas en vacías para separar funciones clases y bloques de código largos.
- Cuando sea posible, poner comentarios en una línea aparte.
- Usar *docstrings*, esto es, documentación de las funciones.
- Usar espacios alrededor de los operadores y después de comas: a = f(1, 2) + g(3, 4).
- Nombra tus clases y funciones de forma consistente. Una posibilidad es utilizar CamelCase para clases y lower_case_with_underscores para funciones y métodos.
- Programar usando ASCII básico.

Ficheros de datos

Los métodos de aprendizaje automático permiten construir modelos a partir de un conjunto de datos. En las prácticas, utilizaremos algunos conjuntos de datos del repositorio de la Universidad de California Irvine (UCI)[2], así como el conjunto de datos de dígitos manuscritos recogido durante las clases de teoría. Los conjuntos de datos que utilizaremos vendrán dados por un fichero de datos (extensión `.data`) y un fichero con una descripción de los mismos (extensión `.names`). En caso de ser necesario, cada fichero de datos debe tratarse para que siga la siguiente estructura:

- 1ª Fila: Número de datos del conjunto
- 2ª Fila: Nombres de los atributos. El último atributo corresponderá a la clase en el caso de problemas de aprendizaje supervisado.
- 3ª Fila: Tipos de los atributos: Nominal o Continuo
- Resto de filas: Conjunto de datos, uno por fila y campos separados por comas.

Como ejemplo, en esta práctica se proporcionan en *Moodle* los ficheros correspondientes a los conjuntos de datos ***tic-tac-toe*** (http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame) y ***credit approval*** (http://archive.ics.uci.edu/ml/datasets/Credit+Approval). Además de los conjuntos de datos originales *tic-tac-toe* y *credit aproval*, se proporciona una modificación del conjunto de datos *credit approval* en que se han eliminado aquellos patrones con valores desconocidos (*missing values*) en algún atributo.

[2] http://archive.ics.uci.edu/ml/

Actividades

La planificación temporal sugerida y las actividades a llevar a cabo son las siguientes:

- *1º semana*: Leer y comprender la presentación de introducción a Python y realizar los ejemplos y ejercicios que hay a lo largo de la misma. Implementar la clase *Datos* para leer los datos del fichero de entrada y almacenar la información necesaria para su posterior uso por los algoritmos de aprendizaje automático y métodos de particionado. Una posible estructura de implementación se comenta a continuación.

Diseño

Con el objetivo de desarrollar una aplicación lo más flexible y general posible se plantea la siguiente estructura para la clase *Datos*:

```
import numpy as np

class Datos(object):

  supervisado=True
  TiposDeAtributos=('Continuo','Nominal')
  tipoAtributos=[]
  nombreAtributos=[]
  nominalAtributos=[]
  datos=np.array(())
  # Lista de diccionarios. Uno por cada atributo.
  diccionarios=[]

  # TODO: procesar el fichero para asignar correctamente las variables
supervisado, tipoAtributos, nombreAtributos,nominalAtributos, datos y
diccionarios
  def __init__(self, nombreFichero,sup):

  # TODO: hacer en las proximas practicas
  def extraeDatosTrain(idx):
    pass

  def extraeDatosTest(idx):
    pass
```

Se deberá implementar el constructor de la clase (`__init__`) que recibe como parámetros el nombre del fichero de datos y un flag indicando si se trata de un problema supervisado y, en caso de ser así, el último de los atributos corresponde a la clase de cada patrón. Las funciones extraeDatosTrain y extraeDatosTest se implementarán en próximas prácticas.

Los atributos de la clase deberán guardar la siguiente información:

- **`supervisado`**: flag para indicar si se trata de un conjunto de datos con etiquetas.

- **`TiposDeAtributos`**: esta variable no debe tocarse. Guarda las dos posibles descripciones de los tipos de atributos.
- **`tipoDeAtributos`**: Lista con la misma longitud que el número de atributos del problema (incluyendo la clase) y que contendrá el tipo de atributo de cada variable (Continuo o Nominal). En caso de que el fichero de datos contenga algún tipo de datos que no se corresponda a uno de estos dos tipos, se deberá informar del error. Por ejemplo, se puede lanzar una excepción del tipo `ValueError`.
- **`nombreDeAtributos`**: Lista con la misma longitud que el número de atributos del problema (incluyendo la clase) y que contendrá el nombre de cada variable.
- **`nominalAtributos`**: Lista de valores booleanos con la misma longitud que el número de atributos del problema (incluyendo la clase) que contendrá True en caso de que el atributo sea nominal y False en caso contrario. Esta estructura será útil posteriormente para el uso del paquete scikit-learn.
- **`datos`**: Array bidimensional de NumPy (matriz) que se utilizará para almacenar los datos. El uso de NumPy facilitará el indexado de los datos, así como el uso del paquete scikit-learn.
- **`diccionarios`**: lista de diccionarios con la misma longitud que el número de atributos. La posición i-ésima de la lista contendrá el diccionario asociado al atributo i-ésimo. Los diccionarios solo deben generarse para las variables nominales, para variables continuas, se guardará un diccionario vacío. Para las variables nominales, el diccionario establecerá la relación entre los valores categóricos (claves) y un entero (valor). Para garantizar que este mapeo es consistente para diferentes ficheros y permutaciones de los datos, los enteros deben asignarse en orden lexicográfico de las claves. Así, por ejemplo, en el caso del conjunto de datos tic-tac-toe, donde el primer atributo puede tomar los valores 'x', 'b', 'o', el diccionario correspondiente deberá ser: `{'b': 0, 'o': 1, 'x': 2}`. Una vez obtenidos los diccionarios de cada atributo, las variables categóricas deberán ser codificadas a los valores enteros asignados en el diccionario para su posterior almacenamiento en el array NumPy `datos`.

Esta clase se definirá en el módulo faa (fichero faa.py). La plantilla de la clase Datos se puede encontrar en *Moodle.* De esta forma, se pueden instanciar elementos de la clase Datos como sigue:

```
from faa import Datos
dataset=Datos('./ConjuntosDatos/tic-tac-toe.data',True)
```

Fecha de entrega y entregables

Esta práctica no tiene entrega asociada y su duración es de una semana.