Maestría en Electrónica

Reconocimiento de Patrones

Práctica 2: Métodos de Aprendizaje Automático

II Cuatrimestre, 2018

Esteban Martínez Valverde
estemarval@gmail.com

1. Estudio sobre los métodos supervisados y no-supervisados

Como parte de la *Práctica 2* y *Apuntes 2* asignadas, se provee el documento con la investigación de los métodos supervisados y no-supervisados vistos en el curso, a través del siguiente hipervínculo:

2. Selección de un conjunto de datos

Por lo general, la obtención o la selección del conjunto de datos (*Dataset*) siempre es el proceso que requiere de más tiempo en los proyectos de reconocimiento de patrones o *Machine Learning* (ML), ya que la respuesta a los diferentes métodos de ML depende de la naturaleza del *Dataset* escogido. Es por esto que se realizó una extensiva búsqueda en los diferentes repositorios de libre acceso. Los *Datasets* que se identificaron de interés se descargaron, analizaron y realizaron pequeñas pruebas para determinar el comportamiento de los datos.

Finalmente, se encontró un *Dataset* denomidado como *Forest type mapping Data Set*, en el repositorio de ML en linea de la Universidad de California Irvine. Este *Forest Dataset* contiene datos de teledetección (*remote sensing*) multitemporal de un área boscosa en Japón. El objetivo es mapear diferentes tipos de bosques utilizando datos espectrales.

Estos datos fueron obtenidos de un estudio de teledetección que mapeó diferentes tipos de bosque en función de sus características espectrales en longitudes de onda de infrarrojo visible a cercano, utilizando imágenes de satélite ASTER.El resultado (mapa de tipo de bosque) se puede usar para identificar y/o cuantificar los servicios del ecosistema (por ejemplo, almacenamiento de carbono, protección contra la erosión) proporcionados por el bosque.

Los datos se encuentran en dos archivos (training/testing) en el formato *Comma-separated values* (CSV), con las siguientes características:

- Nombre: **"training.csv"**
  - Atributos: 27
  - Instancias: 325 (62 %)
- Nombre: **"testing.csv"**
  - Atributos: 27
  - Instancias: 198 (62 %)

De lo cuál se crea un nuevo archivo (llamado **"forest_dataset.csv"**), conteniendo todos los datos de los dos archivos anteriores, con un total de 523 instancias y 27 atributos Esto, con el fin de poder obtener un mejor control en cuanto a porcentajes entre los *training/testings Datasets*.

La descripción de los atributos se detalla a continuación:

- **Class:**
  - 's'('Sugi'forest)
  - 'h'('Hinoki'forest)
  - 'd'('Mixed deciduous'forest)
  - ó'(Ótherñon-forest land)

- **b1 - b9:**
  ASTER image bands containing spectral information in the green, red, and near infrared wavelengths for three dates (Sept. 26, 2010; March 19, 2011; May 08, 2011.
- **pred_minus_obs_S_b1 - pred_minus_obs_S_b9:**
  Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 's'class (b1-b9).
- **pred_minus_obs_H_b1 - pred_minus_obs_H_b9:**
  Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 'h'class (b1-b9).

sdasd

3. Aplicación de métodos de aprendizaje

- **kNN para clasificación:**

  Accuracy with k=5: 0.8215384615384616
  Accuracy with k=7: 0.8184615384615385
  Accuracy with k=9: 0.8174358974358974

- **Regresión Lineal:**

- **SVM**

- **SVM**

Repositorio del *notebook*
El *Jupyter notebook* del ejercicio de PCA, se puede encontrar con el nombre: *aplicacion_pca_poverty.ipynb*. Mientras que el *Dataset* se puede encontrar en la dirección: */data/MPI_national_labeled.csv*.

Referencias