Laboratorio 2 - Práctica de Aprendizaje Automático

Octubre 2016

1. Objetivo

El objetivo de este laboratorio es familiarizarse con herramientas de Aprendizaje Automático (AA), y utilizarlas para la resolución de problemas prácticos. En particular, en este laboratorio se trabajará en la implementación de varios clasificadores y técnicas.

2. Descripción del problema.

En este laboratorio se divide en dos partes. En la primera parte se trabajará con datos de un censo con diversos indicadores y en la segunda con imágenes de dígitos escritos a mano.

Para la primera parte deberán trabajar con un conjunto de datos obtenidos de un censo sobre los ingresos de personas junto con indicadores como edad, educación y ocupación. Se intentará entrenar clasificadores que aprendan a predecir el nivel de ingresos (mayor a 50K o menor o igual a 50K) en función del resto de los indicadores.

En segunda instancia se trabajará con un conjunto de datos compuesto por imágenes de dígitos escritos a mano. Cada instancia es un vector donde cada componente representa el valor en escala de grises de la imagen, y una etiqueta que indica el número que está dibujado. Se deberán entrenar dos clasificadores con algoritmos de aprendizaje diferentes que aprendan a predecir la etiqueta del número correcto en función de los valores de los pixeles de la imagen correspondiente.

3. Herramientas

La solución se implementará completamente en el lenguaje de programación Python en su versión 2.7. Python es un lenguaje multipropósito y multiparadigma que, entre otras cosas, es muy utilizado en el área de AA.

Para extender aun más la potencia de Python, utilizaremos las siguientes bibliotecas:

- **Pandas**
  Pandas [1] es una librería de código abierto implementada en Python, que permite manipular y analizar datos en tablas de una forma muy sencilla y rápida. Utilizaremos esta librería para importar, analizar y manipular los datos con los que vamos a trabajar.
- **Scikit-learn**
  Scikit-learn [2] es un conjunto de librerías de código abierto para AA, implementadas en Python. Es una de las librerías de AA más utilizadas en el área y cuenta con la implementación de varios de los algoritmos más conocidos.

- **Jupyter Notebook**
  Jupyter Notebook [3] es un ambiente de Python que se accede a través de un navegador. Permite trabajar con código Python de una forma muy amigable e interactiva. En él se pueden combinar ejecución de código, texto y gráficos en un solo documento. En este laboratorio se utilizará Jupyter Notebook como ambiente de desarrollo, las entregas se realizaran entregando los archivos generados por esta herramienta. Luego de instalado Python, Pandas, Scikit-learn y IPython, para abrir un notebook se debe de ejecutar el comando:

  >jupyter notebook

  en el directorio donde extrajeron los archivos del laboratorio. Esto abrirá una ventana de navegador donde podrá abrir el notebook que contiene las preguntas, guía y casillas donde ingresar las respuestas y el código que resuelva cada etapa de la tarea.

4. Se pide

Se deberá construir un sistema que:

- Importe los datos del censo.
- Importe los datos de los dígitos.
- Realice un preprocesamiento de los datos.
- Extraiga los atributos para realizar el aprendizaje.
- Entrene algoritmos de aprendizaje automático.
- Mida la performance de los modelos predictivos generados.

Los pasos a implementar serán solicitados en cada una de las celdas descritas en el Jupyter notebook adjunto a esta letra.

5. Formato y fecha de entrega

Cada grupo deberá entregar un archivo .zip de nombre L2GNN.zip (donde NN es el número de grupo) conteniendo:

- El Jupyter notebook con el código de las soluciones y las respuestas a las preguntas. Otros archivos que consideren pertinentes a la entrega.
- El trabajo puede entregarse hasta las 24 horas del día Lunes 31 de Octubre, a través de la plataforma EVA.

6. Evaluación

Para la evaluación del trabajo se tomará en cuenta:

- Resultados obtenidos por las soluciones.
- La calidad de las respuestas, en particular la explicación y justificación de las decisiones tomadas, así como el análisis de los resultados obtenidos.

Referencias

[1] Pandas sitio oficial. `http://pandas.pydata.org/pandas-docs/stable/`. Accedido: 14-10-2016.

[2] SciKit-Learn sitio oficial. `http://scikit-learn.org/stable/documentation.html`. Accedido: 14-10-2016.

[3] Jupyter Notebook sitio oficial. `http://jupyter.org/`. Accedido: 14-10-2016.