Laboratorio 2015

1da. Entrega

Permisos de Construcción en Montevideo

Objetivos

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 el estudio de un conjunto de datos y en la implementación de clasificadores.

Descripción del Problema

En este laboratorio se utilizarán datos producidos por la Intendencia Municipal de Montevideo y extraídos del repositorio de datos abiertos [1], sobre permisos de construcción en Montevideo. Dichos datos contienen información sobre el tipo de obra (obra nueva, reforma, etc.), los m^2 involucrados, el destino de la obra (vivienda, comercio, etc.), la ubicación en la ciudad (dirección y Centro Comunal Zonal) y la fecha del permiso. Se propondrán varios casos de clasificadores que predigan una de estas variables en función de otras.

Herramientas

La solución se implementará completamente en el lenguaje de programación *Python* [2,3,4] en su versión 2.7[1]. Python es un lenguaje multipropósito y multiparadigma que, entre otras cosas, es muy utilizado en el área de AA.

Se utilizarán bibliotecas de Python especializadas en análisis de datos y AA.

- **Pandas**
  Pandas [5] es una biblioteca de código abierto implementada en Python, que permite manipular y analizar datos en tablas de una forma muy sencilla y rápida.
  Se utilizará para importar, analizar y manipular los datos con los que se va a trabajar.

- **Scikit-learn**
  Scikit-learn [6,7] es un conjunto de bibliotecas de código abierto para AA, implementadas en Python. Es uno de los ambientes de AA más utilizados en el área y cuenta con la implementación de varios de los algoritmos más conocidos.

- **IPython Notebook**
  IPython Notebook [8] 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 IPython 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:

```
> ipython notebook
```

[1] Es obligatorio en este laboratorio utilizar la versión 2.7 y no instalar la versión 3.0 en donde se han hecho grandes cambios sin compatibilidad hacia atrás.

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.

Se pide:

Se deberá :

1. Importar los datos de los permisos en un ambiente python.

2. Realizar un estudio de las características de los datos
   1. ¿Son todos correctos o hay errores?
   2. Para los atributos numéricos, calcular la media, desviación típica, mínimo, máximo, etc.
   3. ¿Hay *outliers* en los datos numéricos?
   4. ¿Hay valores faltantes?

3. Realizar transformaciones de los datos de modo de que puedan ser estudiados y utilizados como entrada para clasificadores. En particular, se tratará de predecir el destino de la obra en función de otros atributos.

4. Entrenar algoritmos de aprendizaje automático.

5. Medir la performance de los modelos predictivos generados.

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

Formato y fecha de entrega
Cada grupo deberá entregar un archivo *.zip* de nombre *LXGNN.zip* (donde *X* es el número de entrega de laboratorio y NN es el número de grupo) conteniendo:

- El IPython 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 19 de Octubre, a través de la plataforma EVA.

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] Catálogo de datos abiertos - Sitio Oficial - https://catalogodatos.gub.uy/
[2] Python - Documentación Oficial - http://docs.python.org/2/
[3] Dive Into Python - Sitio Oficial - http://www.diveintopython.net/
[4] Python Essential Reference - Addison-Wesley Professional (July 19, 2009) – ISBN 0672329786
[5] Python Data Analysis Library – Sitio Oficial - http://pandas.pydata.org/

[6] Scikit-learn – Sitio Oficial - http://scikit-learn.org/
[7] Learning scikit-learn: Machine Learning in Python – Packt Publishing (November 25, 2013) – ISBN 1783281936
[8] IPython Notebook – Sitio Oficial - http://ipython.org/notebook.html