Syllabus

Desarrollo de modelos en Python

Programa IA&DL República Dominicana

Guillermo Sánchez González

(guillermo.gonzalez@strategybigdata.com)

V20170425

On Line

1. Presentación de la materia

El concepto de **minería de datos** hace referencia al descubrimiento de hechos y relaciones interesantes en grandes bases de datos. La minería de datos es un proceso con diferentes pasos. Entre estos pasos, un elemento crucial es tratar de extraer modelos de los datos que permitan bien comprenderlos mejor o bien realizar predicciones.

Por ejemplo, en una gran base de datos de clientes de una compañía de telefonía móvil, podemos tener datos de los clientes que han abandonado y se han ido a otra compañía. En este caso, nos interesaría tener un modelo matemático que nos dijese para un nuevo cliente la probabilidad de que nos abandone, para poder, por ejemplo, darle una oferta especial y retenerle. Para ello, tenemos que "aprender del pasado", es decir, obtener el modelo a partir de la base de datos existente, seleccionando las variables más determinantes. En nuestro ejemplo, podrían ser el número de llamadas al servicio técnico, el coste de ciertos servicios u otros.

Los algoritmos de **aprendizaje automático** esencialmente realizan esa tarea, aprender de los datos para poder entenderlos mejor y obtener modelos que nos permitan clasificar o predecir. Hay una gran variedad de algoritmos y el objetivo no es conocerlos todos, sino saber cómo buscar entre esa variedad y seleccionar los que podrían ser la solución a nuestra necesidad de negocio. Es importante resaltar que para utilizar estos algoritmos no es estrictamente necesario tener un conocimiento profundo de cómo funcionan, lo importante es entender cómo utilizarlos y evaluar sus resultados.

2. Objetivos de aprendizaje

El objetivo general del módulo es el de ser capaz de realizar modelos de aprendizaje automático básicos, y entender el papel de estos modelos en el contexto general de la minería de datos como proceso, y cómo se pueden utilizar estas y otras técnicas analíticas a datos obtenidos en la Web.

Los objetivos específicos son los siguientes:

1. Entender el proceso y objetivos de la minería de datos y el papel del aprendizaje automático y saber relacionarlo con otras tareas analíticas que realizan los *data scientists*.
2. Saber diferenciar entre aprendizaje supervisado y no supervisado, conocer algunos algoritmos en cada categoría y saber evaluarlos y utilizarlos.
3. Saber aplicar los conceptos de agrupamiento (*clustering*) y de regla de asociación e interpretarlos para casos concretos.
4. Comprender el concepto de recomendador y cómo se construyen los modelos de aprendizaje automático relacionados.

3. Programa de la materia: estructura y contenido

1. TEMA 1. Aprendizaje supervisado Regresión.
    - Regresión lineal y polinómica
    - Regularización
    - K-Vecinos
    - Validación cruzada
    - Selección de atributos
    - Actividades prácticas en clase.
2. TEMA 2. Aprendizaje supervisado clasificación.
    - Regresión logística.
    - Métricas de Evaluación de modelos (Matrices de confusión)
    - Regularización
    - Naybe Bayes
    - Random forest
    - Actividades en clase
3. TEMA 3. Aprendizaje no supervisado.
    - Clustering K-means.
    - Mixtura Gausiana

- PCA
- Teoría del aprendizaje
- Actividades en clase

4. TEMA 4. Reglas de asociación y sistemas de recomendación
   - Reglas de asociación (A-Priori)
   - Recomendadores basados en filtrado colaborativo.
   - Actividades en clase

4. Metodología y Actividades

El programa se desarrollará metodológicamente planteando los conceptos fundamentales de la actividad del data scientist, y fomentando la discusión de los principales aspectos de esa actividad.

El elemento central de la metodología docente es el trabajo práctico con un entorno de análisis de datos, para entender el trabajo del data scientist y adquirir habilidades básicas para seleccionar y aplicar algoritmos de aprendizaje automático. El entorno de análisis de datos es el IPython Notebook (ahora denominado Jupyter). Las clases presenciales se plantean con una primera parte teórica y una segunda parte de prácticas guiadas por el profesor.

Los alumnos que tengan un ordenador razonablemente potente (4 hilos y 8 GB de RAM o más), pueden instalar el *stack* de Anaconda que contiene todas las librerías incluído el Jupyter Notebook, en la versión Python 3.5+ entrando en https://www.anaconda.com/download y en la pestaña correspondiente a su sistema operativo. Se recomienda instalarlo si se puede antes de comenzar el módulo.

Itinerario	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Presentación del módulo	Videoconferencia apertura					
Introducción al Machine Learning	Aprendizaje supervisado regresión					
Python en Data Science **Tema 1**		Aprendizaje supervisado clasificación	Aprendizaje no supervisado	Reglas de asociación y sistemas de recomendación		
Actividad 2. **Caso de aprendizaje**				Realizar actividad		
Cierre y conclusiones semana online						Videoconferencia cierre

Descripción de las actividades:

Actividad 2. Caso de aprendizaje supervisado

Carácter: individual

Herramientas: Materiales Tema 2,3 Jupyter y scikit-learn

Desarrollo y plazo de ejecución:

Se proporcionará un notebook en el que se pedirá a los alumnos realizar un modelo de aprendizaje supervisado, se les dará un csv con el conjunto de datos.

¿Qué debes hacer?

Seguir el enunciado que se proporcionará y trabajar con el entorno de análisis para resolverlo.

Esta actividad es evaluable a través de un cuestionario en la plataforma.

5. Evaluación

El modelo de evaluación se basará en las actividades de curso desarrolladas (trabajos, casos, ejercicios, etc.). Las actividades se valorarán como sigue:

Actividad 2: Caso de aprendizaje supervisado:	50%
Participación en Foros y Videoconferencia:	50%

6. Bibliografía y materiales de consulta

Materiales obligatorios

Manuales de referencia de scikit-learn disponibles en:

http://scikit-learn.org/stable/documentation.html

Se proporcionan otros recursos on-line en los diferentes temas.

Referencias recomendadas

Una introducción muy básica al aprendizaje automático, muy recomendable para quien no tenga ningún conocimiento previo es el siguiente libro

- *Bootstrapping Machine Learning* (Louis Dorard), disponible en http://www.louisdorard.com/machine-learning-book/ Cuenta con edición electrónica y recursos adicionales para profundizar si se desea.

Hay muchos textos más avanzados que pueden utilizarse para profundizar en técnicas concretas más allá de los contenidos del módulo. Un ejemplo de uno de estos textos que los estudiantes pueden utilizar es el siguiente.

- Alpaydin, E. (2014). Introduction to machine learning. MIT press.
- Hastie, T. (2009). The Elements of Statistical Learning. Springer.

7. CV del Profesor

Guillermo Gonzalez, Licenciado en Matemáticas, científico de datos en Strategy Big Data donde desarrolla modelos avanzados de machine Learning usando redes neuronales e implementando publicaciones de MIT o Stanford para tener las mejores algoritmos.