
 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años Ministerio de Educación Presidencia del Estado 7 de febrero de 2015</p>	<p>Ingeniería de Sistemas y Computación</p> <p><b>Maestría</b></p> <p>MINE4206 – <i>Machine Learning</i> en el modelaje y análisis de información</p> <p>Semestre: 2019-10</p>	
--	--	---

### **TALLER No. 1**

#### **Proceso de aprendizaje a partir de datos**

#### **Preparación de los datos**

##### **Objetivos.**

Comprender los aspectos relacionados con la preparación de datos, tomando en cuenta las características y complejidad de su representación, así como los requerimientos de los algoritmos de aprendizaje que serán empleados para resolver el problema.

##### **Descripción.**



**Para los problemas de clasificación que se exponen a continuación, realice la fase de comprensión de los datos, y proponga y aplique una estrategia para la limpieza y preparación de estos, tomando en cuenta las características del algoritmo de aprendizaje que será utilizado en la etapa de modelado.**

1.- Actualmente, las instituciones bancarias deben competir entre ellas con servicios que brinden confianza, comodidad y accesibilidad, como por ejemplo, el depósito de cheques sin tener que asistir a una oficina comercial. Este servicio crea eficiencias dentro de las instituciones financieras, ayuda a atraer a nuevos clientes y abre nuevos canales para la prestación de servicios.

En una de tales instituciones existe la necesidad de que los ejecutivos de cuentas del área de negocios identifiquen los clientes potenciales para ofrecerles el servicio de depósito remoto de cheques. La institución cuenta con datos relacionados con campañas de marketing directo basadas en llamadas telefónicas. El conjunto de datos disponible está relacionado con 17 campañas; para cada contacto, se almacenó una gran cantidad de atributos y si hubo un éxito. Para todo el conjunto de datos considerado ("bank-marketing"), hubo una tasa de éxito del 11.69%.

***Para este problema el algoritmo que será utilizado puede tratar con atributos numéricos y nominales, pero es sensible al desbalance de las clases.***

2.- Uno de los peligros inherentes a la actividad minera es la amenaza sísmica que ocurre con frecuencia en muchas minas subterráneas. Los factores que influyen en la naturaleza de estos eventos son muy diversos, y las relaciones entre estos factores son muy complejas y muy poco conocidas. Los métodos utilizados hasta ahora para anticipar la actividad sísmica peligrosa no cubren las necesidades en este sector, ya que resultan insuficientes para lograr una buena sensibilidad y especificidad en las predicciones. Es por esto que se ha planteado verificar si los métodos de *machine learning* pueden ser capaces de predecir eventos sísmicos peligrosos. En un primer intento, se quiere construir un modelo de clasificación a partir del conjunto de datos "seismic-bumps", el cual está relacionado con la predicción de riesgos sísmicos. (Importante: utilizar el conocimiento del dominio reflejado en el artículo de los autores "*Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines*")

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años Ministerio de Educación Resolución 3815 de 2015</p>	<p>Ingeniería de Sistemas y Computación</p> <p><b>Maestría</b></p> <p>MINE4206 – <i>Machine Learning</i> en el modelaje y análisis de información</p> <p>Semestre: 2019-10</p>	
--	--	---

***Para este problema el algoritmo que será utilizado requiere que los datos sean sólo numéricos y es sensible a las diferencias entre los rangos de las variables. Además, es capaz de tratar con el desbalance de las clases a través de un parámetro de regularización.***

3.- Uno de los problemas al que se enfrentan los negocios del mercado de consumidores y de sectores empresariales es el abandono de los clientes. La solución tradicional pasa por predecir cuáles clientes exhiben una alta propensión a abandonar y abordar sus necesidades a través de un servicio de asistencia personalizada, campañas de marketing o mediante la aplicación de exenciones especiales. Un interesante aporte en este dominio es el análisis con técnicas de *machine learning*. Las empresas recopilan datos exhaustivos e históricos sobre sus clientes, y es posible utilizarlos para determinar modelos predictivos que permitan identificar clientes con intención de renuncia a los servicios contratados con la empresa. El conjunto de datos a utilizar es: “WA\_Fn-UseC\_-Telco-Customer-Churn”.

***Para este problema el algoritmo que será utilizado puede tratar con datos numéricos y nominales, pero los expertos sugieren hacer un estudio previo para determinar las variables que puedan resultar más informativas. Esto a su vez facilitaría la interpretación del modelo generado.***

4. Con el fin de construir un buscador de noticias que se adapte a los perfiles y gustos de los diferentes usuarios de un servicio de noticias on-line (*newswire*) se requiere construir un categorizador de textos. Para ello utilice el conjunto de datos disponible en <http://mlg.ucd.ie/datasets/bbc.html>, de la BBC (utilice el conjunto “raw text file”).

***Para este problema aplique análisis de componentes principales como mecanismo para la reducción de la dimensionalidad.***

**Importante: justifique cada decisión tomada en cada paso.**

### **Consideraciones**

- Utilice sólo los conjuntos de datos que se suministran.
- La limpieza de los datos debe incluir, según sea el caso, el tratamiento de ausencias, datos fuera de rango, atributos inútiles o redundantes, entre otros aspectos que considere conveniente.
- El taller podrá ser realizado en grupos máximo tres personas.
- Dos problemas (a su elección) deben ser resueltos con el software RapidMiner.
- Dos problemas (a su elección) deben ser resueltos utilizando la librería scikit-learn en el ambiente Jupyter Notebook.
- Entregables: un documento (máximo 5 páginas) que explique brevemente cómo se realizó la preparación de los datos y la justificación de las decisiones tomadas en cada paso, los procesos RMP (comentados) y los archivos ipynb (comentados).

**Fecha de entrega: 06/03/19**