TEMARIO

Tratamiento de datos

LUMIA learning .com

Descripción

Sin duda el incremento y la disponibilidad de un mayor número de datos e información hace cada vez más fácil el poder efectuar análisis para responder los cuestionamientos que empresas, gobiernos e individuos se plantean. Sin embargo, el manejo adecuado de la información es lo que puede dar validez a los análisis y conclusiones que hagamos. De nada sirve la aplicación de un buen modelo si desde un comienzo la información no ha sido analizada y preparada adecuadamente.

En este curso se analiza el manejo adecuado de la información, empezando con los pasos necesarios que hay que seguir cuando queremos analizar los datos. Posteriormente, a lo largo de los siguientes temas se introducen técnicas de preprocesamiento de la información, así como de modelos que pueden utilizarse para resolver problemas relacionados con su manejo. Estos problemas incluyen la presencia de datos perdidos o muy diferentes al resto, técnicas para mejorar la calidad de la información al reducir su variabilidad, problemas de escala en los datos, selección de las variables adecuadas según el tipo de análisis deseado, etc. Todo esto, desde la perspectiva de aplicar el modelo o proceso adecuado según el tipo de datos. Cada uno de los Temas se ilustran a través de datos y código en Python.

El objetivo de este curso va más allá del conocimiento de las técnicas y su aplicación de forma sistemática. Más allá de lo anterior, el objetivo final de este curso es el de que el estudiante sea capaz preguntarse cuando y por qué debe usar cada uno de los métodos. Dando este paso será posible entonces aplicar con confianza los modelos modernos o clásicos que quieran efectuarse.

1. Análisis de datos
   a) Minerías vs análisis de datos y KDD
   b) Pasos en KDD
   c) Aprendizaje y sus tipos
   d) Preprocesamiento
   e) Modelos en minería de datos
2. Análisis descriptivo de datos
   a) Tipo de variables
   b) Análisis descriptivo básico según el tipo de variable
   c) Medidas de tendencia central y de dispersión
3. Análisis exploratorio de datos bivariado
   a) Medidas de asociación en variables de escala ordinal y superior

b) Matriz de covarianza y correlación
c) Otras medidas de asociación

4. Datos ausentes y duplicación
   a) Duplicación
   b) Datos ausentes
      (i) Tipos de pérdida
      (ii) Métodos de remplazo de la información.
5. Normalización y transformación de datos
   a) Normalización
      (i) Min-max
      (ii) Estandarización
   b) Transformaciones en variables cuantitativas: polinomiales, logarítmicas y a variables cualitativas.
   c) Transformaciones en variables cualitativas:
      (i) One-hot encoding o variables dummies
      (ii) Colapsar o recategorizar.
6. Detección de valores atípicos y ruido
   a) Ruido
      (i) Discrete binning
      (ii) Métodos basados en clustering (DBSCAN)
   b) Outliers: Análisis univariado
      (i) Z-score
      (ii) Desviación.
   c) Outliers: Análisis multivariado
      (i) Isolation Trees
      (ii) MCD
      (iii) Local Outlier Factor
      (iv) SVM de una clase.
7. Reducción de dimensiones
   a) Componentes principales
   b) Análisis factorial.
   c) Escalamiento multidimensional (MDS)
8. Selección de variables (características e instancias)
   a) Supervisado: Outputs cuantitativos
      (i) Modelos lineales
      (ii) Métodos univariados
      (iii) Métodos multivariados
   b) Supervisado: Outputs cuantitativos
      (i) Métodos univariados
      (ii) Métodos multivariados
   c) No supervisado (no outputs): Métodos basados en correlaciones
9. Discretización de datos
   a) Algoritmos no supervisados

(i) Binning
(ii) K-means

b) Algoritmos supervisados
(i) Partición a partir de árboles de clasificación
(ii) Fusión de intervalos con pruebas Ji cuadrado

c) Variables dummies o one-hot encoding

d) Distribución y transformaciones en variables cuantitativas
(i) Transformada de Box-Cox
(ii) Transformada de Yeo-Johnson
(iii) Transformación arco seno

10. Extracción, transformación y subida de datos
a) Bases de datos realcionales.
b) Unificar fuentes de datos
(i) Fuentes
(ii) Integración y consolidación
(iii) Data warehouse

Informes: lumialearning@gmail.com