



Departamento de Matemáticas y
Física
Ingeniería de características

Presentación e Introducción

Dr. Gaddiel Desirena López

Primavera 2024

1. Descripción de la asignatura

Los datos representan un activo potencialmente valioso para la sociedad, empresas y gobiernos. Sin embargo, antes de poder aprovechar los datos mediante el uso de técnicas de áreas como aprendizaje automático y minería de datos, es importante poder seleccionar las características relevantes para un problema o pregunta con el objetivo de reducir la dimensionalidad del conjunto de datos. Por lo anterior, en este curso se analizará la importancia de la ingeniería de características, así como los diversos algoritmos utilizados para la selección de características, la construcción de características compuestas y la reducción de dimensionalidad de conjunto de datos numéricos y alfanuméricos.

2. Propósito general

Limpiar conjuntos de datos y sustraer sus características para crear nuevas variables que puedan ser usadas como entradas por un algoritmo de aprendizaje máquina.

3. Temario

1. Extracción de datos de diferentes fuentes
 - a) Archivos de texto
 - b) Archivos CSV
 - c) Archivos XLX o XLSX
 - d) Archivos JSON
 - e) Archivos XML
 - f) Archivos SHP
 - g) De imágenes
2. Identificación de datos
 - a) Variables numéricas y variables categóricas
 - b) Valores faltantes
 - c) Cardinalidad en variable categóricas
 - d) Relaciones lineales

- e)* Distribuciones de los datos
 - f)* Valores atípicos
- 3. Tratamiento de datos faltantes
 - a)* Eliminación de observaciones
 - b)* Sustitución por media y mediana
 - c)* Sustitución por moda y frecuencia
 - d)* Sustitución aleatoria
 - e)* Valores extremos
- 4. Codificación de variables categóricas
 - a)* Creación de variables binarias
 - b)* Variables categóricas ordinales
 - c)* Conteos o frecuencias de categorías
 - d)* Codificación en base la media
 - e)* Featuring hashing
- 5. Transformación de variables numéricas
 - a)* Transformación logaritmo y recíproco
 - b)* Transformación cuadrática y cúbica
 - c)* Transformación Box–Cox
 - d)* Transformación Yeo–Johnson
- 6. Escalamiento de variables
 - a)* Estandarización
 - b)* Normalización basada en la media
 - c)* Escalamiento de valores máximo y mínimo
 - d)* Escalamiento de máximo absoluto
 - e)* Escalamiento por cuantiles
- 7. Discretización de variables
 - a)* Intervalos de ancho constante
 - b)* Intervalos de frecuencia constante
 - c)* K–Means
 - d)* Árboles de decisión
- 8. Extracción de características en series de tiempo
 - a)* Intervalos
 - b)* Shapelets
 - c)* Diccionario de patrones

4. Evaluación

| Actividad | Ponderación |
|---|-------------|
| Exámenes | 30 % |
| Prácticas de laboratorio | 35 % |
| Tareas | 15 % |
| Proyecto de limpieza de conjunto de datos | 20 % |

| Evaluación parcial | Módulos |
|--------------------|----------|
| Primera | 1, 2 y 3 |
| Segunda | 4, 5 y 6 |
| Tercera | 7 y 8 |

5. Bibliografía

| Título | Autor | Editorial– Año | ISBN |
|--|---|----------------------------------|------------------------------|
| Feature Engineering for Machine Learning and Data Analytics | Guozhu Dong, Huan Liu | CRC Press, 2018 | 1351721267, 9781351721264 |
| Feature Engineering and Selection: A Practical Approach for Predictive Models | Max Kuhn, Kjell Johnson | CRC Press, 2019 | 1351609467, 9781351609463 |
| Python Feature Engineering Cookbook: Over 70 recipes for creating, engineering, and transforming features to build machine learning models | Soledad Galli | Packt Publishing Ltd, 2020 | 1789807824, 9781789807820 |
| Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists | Alice Zheng, Amanda Casari | O'Reilly Media, Inc., 2018 | 1491953195, 9781491953198 |
| Machine Learning Refined: Foundations, Algorithms, and Applications | Jeremy Watt, Reza Borhani, Aggelos Katsagelos | Cambridge University Press, 2020 | 1108480721, 9781108480727 |
| Texture Feature Extraction Techniques for Image Recognition | Jyotismita Chaki, Nilanjan Dey | Springer Nature, 2019 | 9811508534, 9789811508530 |

6. Introducción

6.1. ¿Qué es Ingeniería de características?

La ingeniería de características es el proceso de utilizar el conocimiento del dominio para extraer características de datos sin procesar mediante técnicas de minería de datos [1].

6.2. Motivación

- Es un paso crucial en la canalización del aprendizaje automático, porque las características adecuadas pueden aliviar la dificultad del modelado y, por lo tanto, permitir que la canalización produzca resultados de mayor calidad.

- Los profesionales están de acuerdo en que la gran mayoría del tiempo en la construcción de una canalización de aprendizaje automático se dedica a la ingeniería de características y la limpieza de datos.

6.3. Definiciones

Datos: son observaciones de fenómenos del mundo real.

Aprendizaje automático: Es un subconjunto de la Inteligencia Artificial y se refiere al estudio de algoritmos que mejoran a través de la experiencia [2].

Característica: Es una propiedad individual medible de un fenómeno que se observa. Éstas pueden ser numéricas o estructurales [3].

Conocimiento del dominio: Consiste en saber con precisión cómo se hace algo, tener una intuición de los principios subyacentes e integrarlos en la red de conocimiento existente.

Minería de datos: Es un conjunto de técnicas que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones que expliquen el comportamiento de estos datos.

6.4. Herramientas

<https://www.python.org/>
<https://colab.research.google.com>
<https://uniwebsidad.com/libros/python>

Referencias

- [1] https://en.wikipedia.org/wiki/Feature_engineering
- [2] https://en.wikipedia.org/wiki/Machine_learning
- [3] [https://en.wikipedia.org/wiki/Feature_\(machine_learning\)](https://en.wikipedia.org/wiki/Feature_(machine_learning))