Data science para geociencias

Berenice Martínez Téllez berenice@cicese.edu.mx

Michelle Sainos Vizuett sainos@cicese.edu.mx

Enero, 2021

Descripción del curso

Introducción a la ciencia de datos como herramienta para resolver desafíos de las áreas de ciencias de la tierra, oceanografía, meteorología, ciencias ambientales, etc. Será un curso tanto práctico como teórico.

Materiales de clase

- Todo el material de apoyo se proporcionará vía Slack y también estará a la disposición en github.

Prerequisitos

Nociones básicas de programación y matemáticas básicas, a continuación se enlistan algunos temas básicos

- Álgebra Lineal: manejo de vectores, matrices, concepto de plano, espacio.
- Conceptos básicos de cálculo.
- Probabilidad: teorema de Bayes, probabilidad condicional.
- Estadística descriptiva: media, varianza, distribución de probabilidad.
- Programación estructurada: conceptos básicos como manejo de ciclos, estructuras condicionales, funciones, tipos de datos en python.
- Programación orientada a objetos: conceptos básicos de clases, atributos y métodos.
- Python: manejo básico de Numpy, Pandas y Matplotlib.

Con que tengas la idea es suficiente, no necesitas ser una experta. Lo que si se requiere es tener una computadora con acceso a internet y muchas ganas de aprender.

Objetivos

En este curso aprenderás:

- Conceptos fundamentales de la ciencia de datos.
- Programación de modelos de machine learning y deep learning.
- Aplicaciones de estos modelos a problemas reales.
- Manejo básico de algunos frameworks para ciencia de datos.

Estructura del curso

El curso tendrá una parte teórica donde se verán conceptos y fundamentos matemáticos de como funcionan los algoritmos y modelos. Y tendrá una parte práctica donde programaremos en Python algunos algoritmos y aplicaciones selectos.

Al final del curso se desarrollará un proyecto de su interés con las herramientas propuestas.

Estructura de la parte teórica

1. Introducción a la ciencia de datos
    1.1. Conceptos de ML, DL, Big Data y data science.
    1.2. Aprendizaje supervisado y no supervisado.
    1.3. Problemas de clasificación y regresión.
    1.4. Flujo de trabajo en proyectos de ciencia de datos.
    1.5. Frameworks para ML y DL.
    1.6. Propuestas de proyectos
2. Preparación de los datos
    2.1. Manejo de outliers, normalización y transformación.
    2.2. Visualización y análisis exploratorio de los datos.
3. Técnicas de Validación
    3.1. Hold Out Validation
    3.2. K-Fold Cross Validation
4. Ingeniería de características
    4.1. Eliminación de variables dummy.
    4.2. Reducción de dimensionalidad mediante análisis de componentes principales.
    4.3. Técnica de filtros para la selección de características
5. Rendimiento de los modelos

    5.1. Métricas de rendimiento.

6. Métodos de Machine Learning
    6.1. Regresión lineal.
    6.2. Regresión logística.
    6.3. Árboles de decisión.
    6.4. Naive-Bayes
    6.5. Clusterización
    6.6. K-vecinos cercanos
    6.7. Máquinas de soporte vectorial
7. Modelos de Ensamble
    7.1. Bagging
    7.2. Boosting

Actividades a realizar

- Implementación de reducción de dimensiones mediante PCA.
- Implementación de feature engineering.
- Clasificación de cultivos con imágenes satelitales.
- Predicción de temperatura y suspended particle matter a partir de registros de calidad de aire.
- Predicción si lloverá al siguiente día con datos de clima.
- Predicción de esfuerzo en concreto.
- Clasificación de cobertura de suelos mediante imágenes satelitales.
- Clustering de sismos históricos.
- Detección de fracturas en imágenes de concreto.
- Detección de sargazo.
- Clasificación de asteroides.

Proyecto Final

El proyecto a realizar será decisión de cada participante. Encuentra un problema de las geociencias que te parezca interesante y relevante y te ayudaremos a llevarlo a cabo paso a paso.