Taller 1: Minería de datos 2020-II

Profesores: Andrés Marino Álvarez Meza, Ph.D.
amalvarezme@unal.edu.co
Diego Fabian Collazos Huertas, Ph.D.(c)
dfcollazosh@unal.edu.co
Monitor: Jhon Bryan Bermeo Ulloa, Msc.(c)
jbbermeou@unal.edu.co

Especialización en estadística
Universidad Nacional de Colombia - sede Manizales

1. Instrucciones

El taller debe ser enviado en formato .ipynb al correo electrónico amalvarezme@unal.edu.co desde su correo institucional (no se aceptarán envios desde correos diferentes a @unal.edu.co) incluyendo desarrollos matemáticos, conceptuales, códigos en Python, resultados y discusiones. Se puede resolver de manera individual o en parejas. Fecha máxima de entrega: noviembre 15 de 2020.

2. Preparación y depuración de datos

- Revise y discuta las etapas de preproceso y análisis explortorio de datos (modelos matemáticos y códigos implementados) del cuaderno plantilla FIFA-19
- Con base en el cuaderno del punto anterior, realice una codificación OneHotEncoder de al menos un atributo sobre el paquete de códigos en la clase dummyFIFA. Grafique la proyección de los datos codificados en 2D mediante PCA.

3. Modelos de regresión

- Implemente un cuaderno de Python que permita realizar una comparación en términos del error absoluto medio para la predicción de nuevos contagiados y contagios acumulados de Covid-19 bajo los siguientes horizontes: 1 día, 7 días y 15 días. Se deberá implementar una validación cruzada anidada para la sintoinización de hiper-parámetros, comparando los siguientes modelos: i) Elastic-Net, ii) Kernel Ridge Regression (utilizando kernel rbf), iii) Bayesian Ridge Regression, iv) SVR (utilizando un kernel rbf), v) RandomForestRegression, y vi) KNN. Ver cuaderno guía Covid-19. Se debe escoger al menos un país para mostrar los resultados.

4. Modelos de clasificación

- Extienda el análisis de regresión implementado en el punto 3 para detectar el aumento o no de nuevos contagiados bajo los horizontes expuestos. Si se detecta un aumento la salida deberá ser +1, de lo contrario la salida se fijará en -1. Realice un análisis comparativo de los siguientes modelos de clasificación (bajo el esquema de validación cruzada anidada), utilizando acierto, precisión, exhaustividad, y F1-score: i) Clasificador linear SGD, ii) KNN, iii) SVM (kernel rbf), y iv) RandomForest.

Referencias

https://github.com/amalvarezme/MineriaDatos

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). Massachusetts, USA:: MIT press.

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.