**Prueba práctica – Profesional Unidad de Analítica en Salud**

Desarrolle la siguiente prueba[1] en un Jupyter notebook de Python. Se evaluará su capacidad para escribir código limpio, así como su habilidad para interpretar los resultados y comunicarlos de manera efectiva. Asegúrese de seguir las mejores prácticas de programación y de documentar su código adecuadamente. Incluya cualquier herramienta que haya usado y permita replicar los hallazgos.

Envíe el notebook con las pruebas completadas al correo nikhol.munoz@iets.org.co antes del lunes 05 de agosto de 2024 a la 1 pm.

- Entrene un SVM para detectar si una palabra pertenece a español o a inglés:

**(a)** Construya un conjunto de entrenamiento y otro de validación. Use el listado que se encuentra en el sitio web http:// en.wiktionary.org/wiki/Wiktionary:Frequency_lists. Considere solo las palabras de al menos 4 letras de longitud e ignore los acentos.

**(b)** Implemente los siguientes kernels:

i) Histogram cosine kernel: calcule una representación de bolsa de n-gramas (use CountVectorizer de scikit-learn) y aplique cosine_similarity de scikit-learn.
ii) Histogram intersection: calcule una representación de bolsa de n-gramas, normalícela (la suma de los bins debe ser igual a 1 $\forall i, \left|\left|x_i\right|\right|_1 = 1$) y calcule la suma del mínimo para cada contenedor del histograma.
iii) $\chi^2$ kernel: calcule una representación de bolsa de n-gramas y aplique el chi2_kernel de scikit-learn.
iv) SSK kernel: use el kernel del código que se encuentra en el siguiente repositorio https://github.com/helq/ python-ssk.

**(c)** Use scikit-learn para entrenar diferentes SVMs usando kernels precalculados. Utilice validación cruzada para encontrar los parámetros de regularización adecuados trazando el error de entrenamiento y validación frente al parámetro de regularización. Use una escala logarítmica para C, $\{2^{-15}, 2^{-14}, \ldots . . 2^{10}\}$. Pruebe diferentes configuraciones de los parámetros (en particular, diferentes valores de *n* para los n-gramas)

**(d)** Evalúe el desempeño de las SVMs en el conjunto de datos de prueba:

i) Informe los resultados en una tabla para las diferentes configuraciones evaluadas.
ii) Ilustre ejemplos de errores (palabras en inglés confundidas como español, palabras en español confundidas como inglés). Dé una posible explicación para estos errores.
iii) Discuta los resultados.

[1] Examen extraído y traducido del curso de Machine Learning del Profesor Fabio Gonzales: https://fagonzalezo.github.io/ml-2024-1/assign2.pdf