Aprendizaje Automático y Análisis de Datos
Segundo Parcial
Abril 14 de 2020

Este parcial práctico debe enviarse resuelto a la profesora a más tardar el jueves 23 de abril. Las sustentaciones se realizarán el viernes 24 durante la mañana, incluyendo el horario de clase. Se deben realizar las siguientes actividades en forma individual:

Conceptos que vamos a aplicar:

- Métodos hold-out y cross validation para evaluación de sistemas de aprendizaje automático
- Conocimiento y preprocesamiento de datos
- Análisis de resultados
- Entrenamiento y prueba de modelos
- Mejora de resultados

En este parcial vamos a hacer tres iteraciones de trabajo sobre los datos de partida. En la primera iteración se deberán hacer las fases de:

- Conocimiento de los datos
- Preprocesamiento de los datos
- Entrenamiento de modelos: máquinas de vectores de soporte y perceptrón multicapa.
- Estimación de parámetros (pueden usar la función gridSearchCV que ahorra mucho trabajo)
- Evaluación del desempeño utilizando holdout de 10 iteraciones (por cada técnica).
- Visualización de los resultados del holdout (media y varianza por cada métrica a considerar. Mínimo accuracy, precision, recall y f1-score.)
- Análisis de los resultados y selección de la técnica que produce mejor desempeño.

En la segunda iteración se deberá retomar el mejor modelo y hacerle ajustes que logren mejorar aún más su desempeño. Algunas ideas sobre qué cambiar para lograr una mejora son:

- Con respecto a los datos: reconsiderar los atributos de entrada, el balanceo de los datos, la codificación de los datos, la normalización de valores, entre otros.
- Con respecto a los modelos: incluir otros parámetros adicionales en la estimación, cubrir rangos más amplios o más finos en el proceso de estimación.
- Otras variaciones que a ustedes se les ocurran.

En la segunda iteración entonces deberán realizarse las siguientes actividades:

- Realización de los ajustes que cada persona seleccione.
- Reentrenamiento del modelo manteniendo los valores de los parámetros estimados en la iteración uno y que no sean sujeto de modificación (es decir, no es necesario reestimar todos los parámetros sino sólo aquellos que ustedes decidan reconsiderar, si acaso seleccionan este aspecto como una mejora)

- Análisis de los resultados obtenidos y comparación con los de la primera iteración. Si no hay una mejora, deberán buscar otros cambios a realizar hasta lograr una mejora.

En la tercera iteración se debe hacer una nueva mejora al modelo obtenido en la segunda iteración, para lo cual se deben modificar variables diferentes a las modificadas en la segunda iteración. Por ejemplo, si para pasar de la primera a la segunda iteración se cambió el balanceo de los datos, para pasar de la segunda a la tercera no podrá usarse nuevamente esa opción y se deberá cambiar otra cosa diferente.

El conjunto de datos sobre el que van a trabajar está en el repositorio de UCI que ya conocen, el conjunto se llama Avila data set, y lo encuentran siguiendo este enlace: http://archive.ics.uci.edu/ml/datasets/Avila. Por favor tengan en cuenta que hay dos archivos, uno para entrenamiento y uno para prueba, deben usar todos los datos disponibles (o sea deben usar los dos archivos).

Criterios de evaluación:

- Conocimiento y preprocesamiento de los datos: 10%
- Primera iteración: 20%
    - Estimación de parámetros
    - Entrenamiento
    - Obtención de resultados
- Segunda iteración: 20%
    - Modificaciones realizadas (es importante explicar porqué se decidió explorar esas modificaciones en particular)
    - Análisis de los resultados obtenidos
    - Mejora lograda
- Tercera iteración: 20%
    - Modificaciones realizadas (es importante explicar porqué se decidió explorar esas modificaciones en particular)
    - Análisis de los resultados obtenidos
    - Mejora lograda
- Análisis comparativo de los resultados obtenidos en las tres iteraciones y conclusiones del proceso: 30%