Tarea 4: Sistemas de recomendación y evaluación de modelos

F.C | GD M.D - II-2015

Sistema de recomendación

Un importante periódico le entrega a usted un dataset limpio con información acerca del acceso a su portal web. El mismo contiene 131300 posibles transacciones en un tiempo determinado. Se sabe que el portal ofrece 9 tipos de contenidos y nos ofrecen solo información de 9 artículos. Los contenidos son:

```
- Deportes
- Politica
- Variedades
- Internacional
- Nacionales
- Sucesos
- Comunidad
- Negocios
- Opinión
```

El periódico tiene sospechas de que existen bots que están ganando dinero al hacer clicks en artículos con promociones. En consecuencia, le piden a usted que realice un análisis exploratorio sobre las transacciones para determinar el número de posibles **transacciones bot** que tienen en su dataset (ellos aceptan que si una persona ve un artículo más de 20 segundos entonces no es un bot).

Aunado a esto, tienen una lista de demandas que debe suplir usted, el experto.

1. Modificar su dataset de tal manera que no se lean los identificadores de los artículos como **itemN** sino por su tipo de contenido **contenido/articuloN**. Ejemplo: {item1, item10, item81} es la transacción {deportes/articulo1, politica/articulo1, opinion/articulo9}.
2. Conocer los **tipos de usuarios** que ingresan a su página (ellos creen que son 8 tipos de usuarios) y tratar de determinar la proporción de cada tipo de usuario.
3. Dado un usuario nuevo que haya ingresado a n artículos (n variable), poder recomendar un artículo n+1 y así aumentar el compromiso del cliente con su portal web. Como usted sabe, para poder calcular las reglas necesita como entrada **MinSupport** y **MinConfianza**. Sin embargo, el cliente desconoce cuáles son estos valores en consecuencia es tarea de usted determinar y **justificar** los mismos de acuerdo a su criterio.
4. Conocer las 10 visitas con mayor tiempo de estadía en la página y las 10 visitas con menor tiempo de estadía en la página.
5. Conocer las 10 transacciones con mayor número de apariciones en el dataset.

Todos estos requerimientos deben ser analizados en un informe que deberá entregar al encargado del proyecto con los códigos necesarios para realizar las actividades mencionadas. De esta manera, el mismo debería tener al menos 2 secciones una dedicada al análisis exploratorio de las **transacciones bot** y una sección dedicada al número de **tipos de usuarios** que ingresan a la página.

Aplicación Shiny (OPCIONAL)

Puede escoger sustituir el informe por una aplicación Shiny que contenga los requerimientos del mismo en el caso de ser de su agrado.

Curvas ROC

Las curvas ROC(**Receiver Operating Characteristics**) son gráficos usados como técnica de visualización, organización y selección de clasificadores basados en su rendimiento.

En la segunda sección de su tarea, se le pide que implemente un graficador de curvas ROC para clasificadores cuya salida sea un score. De esta manera, los parámetros de graficador serán:

1. Los scores por instancia (no necesariamente ordenados).
2. La verdadera clase de las instancias.
3. La clase target. En el caso de que $nclass > 2$ entonces haga un enfoque 1 vs all.

```
generate_ROC = function(scores, real, target){
  # Generar curva
}
```

Consideraciones generales

Para el sistema de recomendación

1. Use el lenguaje estadístico R y las librerías **arules** y **arulesviz**.
2. Adjunto el archivo "periodico.csv" con las transacciones crudas que provee el usuario.
3. Adjunto un archivo "ejemplo.csv" con las primeras 20 transacciones ya transformadas según la primera actividad que solicita el periódico.

Para generación de Curvas ROC

1. Puede usar el lenguaje Python o R según sea su agrado.
2. Tomar en cuenta el paper ***An introduction to ROC analysis*** de Tom Fawcett como referencia hasta la sección 5. *Efficient generation of ROC curves.*
3. Tome en cuenta el caso explicado en la figura 6 (casos donde el score de dos instancias es el mismo).

Consideraciones de forma:

Ingrese a la dirección RecomendacionModelos y haga *fork* del repositorio donde encontrará los archivos csv y un README.

Este repositorio será propiedad de usted. En consecuencia, solo podrá realizar cambios en el mismo. El repositorio debe poseer lo siguiente:

1. Todo Script .py intradocumentado o .Rmd reproducible y documentado.
2. README.md explicando la configuración del ambiente en el cual trabajó:
   - Ejemplo: README.md de Bootstrap
   - En el caso de usar python, hacer pip freeze de su versión para conocer librerías instaladas.
3. Un **informe.pdf** relativo a la sección del sistema de recomendación.

Consideraciones de contenido:

- La tarea es **estrictamente** individual. Se promueve la participación y discusión de la misma en un ambiente responsable. Sin embargo, cualquier evidencia de copia será severamente sancionada colocando una nota mínima de cero (0) puntos según lo establecido en la Ley de Universidades. **Cualquier tarea entregado** debe ser fruto de su propio trabajo.
- Fecha de Entrega: **Domingo 14 de mayo de 2016**.
  - Hasta este día se aceptarán push's en los repositorios.
  - No se recibirá ninguna tarea por correo electrónico.
  - La regla de extensión de entregas se aplicará hasta el martes siguiente.