Guía para el desarrollo del proyecto del curso

Julián D. Arias-Londoño
Dpto. of Systems Engineering and Computer Science
Universidad de Antioquia

Resumen
Las actividades propuestas para el desarrollo del proyecto buscan que cada uno de los grupos de estudiantes aborden todo el análisis, el diseño, y la simulación de un sistema de predicción basado en técnicas de aprendizaje de máquina; describiendo el problema, evaluando sus similitudes con otros proyectos y soluciones reportados en la literatura especializada, especificando cada una de las etapas del desarrollo del trabajo, los modelos usados con sus respectivas restricciones, el ajuste de hiperparámetros, la metodología de validación, los resultados de las simulaciones y las conclusiones obtenidas.

Índice

1. Introducción

Cada equipo de trabajo debe seleccionar un problema que pueda ser abordado a partir de las técnicas y aproximaciones vistas en el curso de Modelos y Simulación II. Los problemas abordados pueden ser propuestos por el grupo de estudiantes, o seleccionado a partir del listado que será compartido por el profesor del curso. **Sólo se permite que un proyecto/dataset sea desarrollado por un equipo de trabajo**. El listado de proyectos sugeridos puede ser consultado en el siguiente enlace.

Cada equipo de trabajo debe crear un repositorio GitHub en el cual se debe alojar un reporte del proyecto (en formato pdf) y, al menos, un notebook en el que se puedan reproducir los resultados incluidos en el reporte. El readme del repositorio debe dar indicaciones claras de cómo interactuar correctamente con el repositorio. Para el reporte se debe usar la plantilla IEEE para los artículos publicados en Transactions. Se recomienda usar un editor de texto LaTeX, en particular Overleaf. Tanto el reporte como el/los notebooks deben contener las secciones descritas a continuación.

2. Descripción del problema (20 %)

La primera parte del informe está destinada a realizar una descripción del problema, hacer un análisis exploratorio de los datos que permita comprender cómo está compuesta la base de

datos y a definir la aproximación que desde el punto de vista de *Machine Learning - ML* será seguida por parte del equipo para la solución del problema. Todos los informes deben contener:

1. Una descripción clara del contexto del problema, donde se refleje la utilidad de desarrollar una solución basada en ML para dicho problema.
2. Una descripción de la composición de la base de datos incluyendo (pero no limitándose) el número de muestras, el número de variables, su significado, la existencia de datos faltantes y si es del caso, la estrategia de imputación de datos seguida. Además, se debe describir claramente el tipo de codificación usado para las diferentes variables del problema.
3. El tipo de configuración o paradigma de aprendizaje que el equipo de trabajo decidió como apropiado para el abordaje del problema y su justificación.

3. Estado del arte (10 %)

El propósito fundamental de esta fase del proyecto es que los grupos de trabajo revisen literatura especializada del campo de ML en la que se presenten soluciones a problemas similares al descrito en la sección 2.

4. Realice una búsqueda de al menos 4 artículos que hayan abordado el mismo problema dque Uds están trabajando. Incluya, en la medida de lo posible, trabajos que hayan empleado la misma base de datos. Describa brevemente:
   - ¿Qué configuración o paradigma de aprendizaje se usa en el trabajo?
   - ¿Qué técnica(s) de aprendizaje usan los autores para solucionar el problema planteado?
   - ¿Qué metodología de validación usaron?
   - ¿Cuáles fueron las métricas empleadas para evaluar el desempeño del sistema? En caso de que alguna métrica no haya sido vista en clase, se debe describir en qué consiste y presentar su definición matemática.
   - ¿Cuáles fueron los resultados obtenidos en cada uno de los trabajos citados?

   Se recomienda buscar trabajos en las bases de datos: Elsevier y IEEE. También se pueden buscar trabajos en la base de datos http://link.springer.com, pero se debe tener en cuenta que el acceso que tiene la Universidad es mucho más limitado para dicha base de datos. Incluir preferiblemente artículos publicados en revista, no en congresos o conferencias. Se recomienda utilizar el buscador Google Scholar para encontrar artículos que hayan citado la base de datos seleccionada. **No utilice más de una página del informe para esta descripción**.

4. Entrenamiento y Evaluación de los Modelos (30 %)

En esta sección se debe describir el diseño experimental y los resultados obtenidos durante las simulaciones.

4.1. Configuración experimental

5. Describa la metodología de validación escogida para entrenar y evaluar los modelos de ML. En aso de requerir etapas adicionales, como la utilización de técnicas de sub o sobre muestreo, la forma en la fueron usadas debe estar descrita en este apartado.

6. Incluya una tabla en la que especifique el conjunto de hiperparámetros que fueron analizados durante el proceso de evaluación de cada modelo, junto con la descripción de la malla de valores usada para cada uno. Cada proyecto debe comparar al menos 5 modelos de aprendizaje, entre los que se deben incluir:

   - Un modelo parámetrico basado en funciones discriminantes gausianas o en regresión (lineal o logística según se requiera).
   - Un modelo no parámetrico
   - Un modelo basado en el ensamble de árboles de decisión
   - Una red neuronal artificial
   - Una máquina de vectores de soporte

7. Describa cuáles serán las métricas de desempeño que se usaràn para evaluar el sistema y justifique su utilización.

4.2. Resultados del entrenamiento de Modelos

8. Presente los resultados de la experimentación para cada uno de los modelos evaluados, en los que se evidencie el efecto de los hiperparámetros en el desempeño del modelo. Haga uso de tablas y gráficas que ayuden al lector a comprender de manera clara los resultados. **TODAS** las figuras y tablas deben ser introducidas en el texto del informe y contener un *caption* que la describa. Todas las figuras deben incluir el nombre de los ejes para que el lector pueda comprender la información que se le presenta. Las medidas de desempeño deben incluir los correspondientes intervalos de confianza estimados durante la fase de validación. **Recuerde incluir resultados de entrenamiento, validación y test.**

---

5. Reducción de dimensión (20 %)

En esta fase del proyecto se deben evaluar diferentes técnicas de reducción de dimensión y establecer si es posible reducir la complejidad del modelo final.

5.1. Selección de características

9. Realice un análisis individual de cada una de las características, a partir de medidas de correlación y o índice que refleje la capacidad discriminativa de las variables (según sea el caso). Identifique de acuerdo con este análisis, ¿si existen y cuáles son las características candidatas a ser eliminadas?

10. Realice selección de características por el método de búsqueda secuencial ascendente o descendente y evalúe el subconjunto de características encontrado, en los 2 mejores modelos predictivos encontrados en la sección 4. Para realizar este punto cada grupo debe decidir la función criterio a usar en el algoritmo de selección, justificando su decisión. Incluya una tabla con los resultados, indi-cando el porcentaje de reducción alcanzado y las demás medidas de desempeño. Recuerde incluir en el informe cuál fue el criterio de selección usado y porqué.

5.2. Extracción de características

11. Use el método PCA para encontrar un conjunto de componentes principales inferior al número de variables original y evalúe nuevamente en los 2 mejores modelos de predicción encontrados en la sección 4. Cada grupo debe decidir el criterio para seleccionar el número de componentes principales justificando su decisión. Incluya una tabla con los resultados, indi- cando el porcentaje de reducción alcanzado y las demás medidas de desempeño.
12. Incluya una sección de discusión y conclusiones sobre la evaluación completa de la solución desarrollada, e incluya una comparación de sus resultados con los descritos en la sección 3.

---

6. Evaluación final (20 %)

Para la evaluación final del proyecto se tendrán en cuenta los siguientes criterios:

- Alcance de los obsjetivos de cada una de las secciones anteriores.
- No superar las 10 páginas en el reporte.
- Orden y claridad del repositorio GitHub para la reproducción de los resultados.
- Sustentación: cada grupo debe preparar un video de 10 minutos en el que se presenten un resumen del proyecto. El video será proyectado en una sesión de clase y los estudiantes deberán responder a las preguntas del publico. La capacidad de cada integrante del grupo de responder a las preguntas será evaluada de manera individual.