TEC | Tecnológico de Costa Rica

Instituto Tecnológico de Costa Rica
Área Académica de Ingeniería en Computadores
CE 5506 — Introducción al Reconocimiento de Patrones

Proyecto 1
Explorando los Algoritmos de Aprendizaje Supervisado

Fecha de asignación:	12 de Abril, 2023	Fecha de entrega:	3 de Mayo, 2023
Grupos:	3 personas	Profesor:	Jason Leitón Jiménez

1. Objetivo

Evaluar los diferentes algoritmos de aprendizaje supervisado junto con las implicaciones de los set de datos, con el fin de comprender el efecto que genera un buen tratamiento de datos, utilizando sets de uso público y real.

2. Motivación

Un buen proyecto de Machine Learning (ML) suele comenzar con definir un problema específico que amerita el uso de ML. Una vez que se detecta que amerita su uso, se elije un set de datos útil y representativo. Seguidamente se analiza las variables independientes (features) y dependientes (labels/target), con el fin de normalizar. estandarizar, modificar o, en algunos casos, generar nuevas features (con base en las existentes). Seguidamente se debe elegir un algoritmo de ML acorde al problema, ya sea si es regresión, clasificación, tomando en cuenta las limitaciones de cantidades de datos, limitaciones computacionales, así como explicabilidad deseada. Adicionalmente, para escoger el modelo, muchas veces no basta con conocer con anticipación los algoritmos, sino que requiere probar y comparar.

3. Descripción General

El proyecto consiste en evaluar de manera objetiva 3 algoritmos de aprendizaje supervisado, los cuales son:

1. Regresión Logística.
2. Árboles de decisión.
3. KNN

Se puede utilizar la implementación vista en clase. Además, se debe de realizar una comparación del algoritmo creado por cada grupo contra el algoritmo correspondiente de la biblioteca **sklearn**. Se pretende que cada grupo explore su propio algoritmo y desarrolle mejoras con el fin de obtener los mejores resultados.

También se debe de establecer la interpretación de los resultados, con el fin de realizar un análisis profundo y basado en métricas para indicar cuál algoritmo es el mejor de forma objetiva para cada uno de los sets de datos.

3.1. Métricas

Cómo mínimo, se espera que para cada algoritmo y set datos se puedan obtener las siguientes métricas:

1. Accuracy
2. Precision
3. Recall
4. AUC (gráfica)
5. ROC (gráfica)

   Cualquier otra métrica adicional será bienvenida y aportará información en el análisis de los resultados. Cabe resaltar que los resultados deben ser en términos de los test sets, sin embargo, se debe de analizar si hubo overfitting o underfitting en cada caso, por lo que se debe tener en cuenta las métricas de los training sets.

   Cada grupo se encargará de ejecutar las implementación con las mismas distribuciones de set de datos, con el fin de que la comparación sea lo más equitativa posible. Se aconseja ejecutar más de una vez para asegurarse que los experimentos sean **reproducibles**.

   Nota: Los resultados finales deben ser los que arrojaron los modelos con los parámetros ya ajustados. Para lograr esto se recomienda utilizar GridSearch de sklearn.

3.2. Set de datos

Para cada uno de los set de datos se espera que se realice *feature engineering* con el fin de tener mejores resultados en el entrenamiento. Además, de que realicen un análisis de los problemas que presentan los sets, por ejemplo, si se presenta algún tipo de sesgo. Cada grupo deberá elegir el porcentaje del set de testeo de tal manera que arroje los mejores resultados (recordar utilizar el *stratify*).

Se utilizarán 3 set de datos:

3.2.1. Red Wine Quality

Este set de datos contiene 11 propiedades fisicoquímicas del vino (features) con más de 1000 muestras y una variable target llamada calidad. La calidad del vino, en números del 3 al 8, puede ser agrupada en MAL VINO y BUEN VINO (por ejemplo, valores de 3,4,5 pueden ser MAL VINO, y 6,7,8 BUEN VINO, o similar). De esta manera, este set de datos representa el problema de poder predecir si el vino es bueno o no, con base en propiedades fisicoquímicas (clasificación binaria) Link.

3.2.2. Notas del curso de Arquitectura de Computadores 1

Este set de datos es proporcionado por el profesor Luis Alberto Chavarría, consiste en las notas de los estudiantes del curso de Arquitectura I durante varios semestres. El objetivo con este set de datos es crear un modelo que se le pase como datos la nota del proyecto 1, proyecto 2, examen 1 y tarea 1, con el fin de que el modelo prediga si la nota será mayor o menor a 67,5 (clasificación si pasará o no pasará el estudiante).

Este set de datos lo podrá encontrar en tecdigital y uno de los retos más importantes es que se debe de uniformar la información, así como eliminar datos que no se necesitan, verificar si existen outliers, es decir, aplicar feauture engineering.

3.2.3. Set de datos a elegir

Se debe de elegir un set de datos que les interese, que permita hacer clasificación binaria y que sea factible correr los 4 algoritmos en él. Deben tener en cuenta que el set de datos no sea grande (posiblemente en el orden de miles de muestras esté bien), para que sea manejable en memoria y en CPU.

3.2.4. Puntos extras (15 puntos)

Se debe realizar un modelo que prediga la nota del estudiante (utilizando el set de datos de Arquitectura de Computadores I). La predicción se debe de hacer a partir de las siguiente notas: proyecto1, proyecto 2, examen 1, taller 1. El objetivo del modelo es predecir la "futura"nota (problema de regresión).

4. Documentación- Estilo IEEE-Trans

- Introducción: Teoría necesaria, breve descripción del proyecto y qué es lo que se espera en el escrito.

- Detalles del diseño del programa desarrollo. En esta sección se espera observar un diagrama de flujo de cada algoritmo, así como la elección de los parámetros a utilizar para cada problema.
- Resultados y análisis. Se espera ver todas las métricas así como la generalización de las mismas, aquí se debe mencionar los problemas de los set, si hubo overfitting, underfitting, cuál fue el mejor, las razones de los mejores y peores resultados, mencionar comparaciones de resultados con parámetros y cualquier otro análisis que sea relevante
- Conclusiones
- Referencias

5. Entregables

1. Notebook: Un solo notebook(.ipynb) con el código y con los mejores resultados de cada modelo.
2. Paper: PDF con el contenido de la documentación

6. Fecha de entrega y revisión

3 de mayo a las 15:00. Revisiones en horario de clase.

7. Evaluación

- Red Wine 15 %
- Notas del curso 15 %
- Set de datos arbitrario 15 %
- Feature Engineering 15 %
- Ajuste de parámetros 10 %
- Paper 30 %