Instituto Tecnológico de Costa Rica Semestre 1, 2021
Escuela de Ingeniería en Computación
Maestría en Ciencias de la Computación
Curso: Aprendizaje Automático
Profesor: Dr. José Carranza-Rojas
Valor: 15%
Proyecto en parejas

Proyecto 1 – Explorando Aprendizaje Supervisado

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, escogemos un set de datos útil y representativo. Seguidamente analizamos las variables independientes (features) y dependientes (labels/target), y normalizamos. estandarizamos, modificamos o, en algunos casos, generamos nuevas features (con base en las existentes). Seguidamente se debe escoger 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.

Tanto para comparar, así como para tener una idea objetiva de qué tan bueno es el modelo resultante, requerimos usar métricas que nos den esta información. Una vez que tenemos los resultados de las métricas, podemos proceder a interpretar los resultados y entender si nuestro algoritmo ha aprendido a resolver el problema planteado.

Este proyecto nos lleva a desarrollar estos pasos con dos set de datos, y a explorar 4 algoritmos de ML, compararlos de forma objetiva, y por supuesto, a interpretar los resultados. Finalmente, en forma de paper científico, nos hace describir toda la experimentación realizada, reportar los resultados, y concluir sobre ellos.

Respecto a los algoritmos de ML, es importante recalcar que el objetivo de este proyecto NO es programarlos desde cero. El objetivo es saberlos usar. Esto quiere decir que se recomienda tomar los algoritmos de sklearn o de alguna otra librería similar.

Sets de datos

Los set de datos a usar en este proyecto deben evaluar clasificación binaria, es decir, deben servir para clasificar en una de dos clases nada más.

1. **Red Wine Quality** [1]

Instituto Tecnológico de Costa Rica Semestre 1, 2021
Escuela de Ingeniería en Computación
Maestría en Ciencias de la Computación
Curso: Aprendizaje Automático
Profesor: Dr. José Carranza-Rojas
Valor: 15%
Proyecto en parejas

Link: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Este set de datos contiene 11 propiedades fisicoquímicas del vino (features) con más de 4 mil samples 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).

2. Set de datos a escoger (arbitrario)

Deben escoger 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 samples esté bien), y que no tenga demasiados features, para que sea manejable en memoria y en CPU (por el momento no estamos lidiando con GPUs). En particular, para este proyecto no se recomienda hacer uso de set de datos para computer vision o con texto abierto (nlp), sino más bien hacer uso de set de datos con observaciones.

Algoritmos

No se recomienda programar ninguno de los 4 algoritmos a usar, sino usarlos directamente de alguna librería (en Python). Lo que sí deben hacer es buscar los “mejores” hiperparametros para cada uno (por lo que deben saber a groso modo cómo funcionan). Los algoritmos a usar son:

1. Regresión Logística
2. Árboles de Decisión
3. kNN
4. Redes Neuronales

Herramientas

Se recomienda el uso de sklearn o alguna otra librería en Python. Se deben entregar los fuentes con los resultados en un Jupyter Notebook. Se recomienda hacer uso de matplotlib para generar los plots. El Jupyter Notebook debe traer todos los resultados sin necesidad de correrlo.

Métricas

Como mínimo, deben reportar Accuracy, Precision, Recall, AUC y ROC. Métricas adicionales son bienvenidas. Recuerden que los resultados deben ser en términos de test sets (aunque pueden mencionar los de training set, especialmente para hablar sobre Overfitting o Underfitting). En general, el test set debería ser el mismo para las 4 implementaciones, sin embargo, se recomienda correr los experimentos varias veces con test sets distintos para

asegurarse que son reproducibles (por ejemplo, escoger 5 cortes del set de datos de forma 80/20%, y cada corte correrlo con cada algoritmo).

Paper en Latex / PDF

Deben crear en latex (y entregar los fuentes de latex) un paper donde describen los experimentos realizados, su metodología respecto a la búsqueda de hiper parámetros (GridSearch?), así como de feature engineering (qué cambios y por qué se le hicieron a los features). Deben incluir los resultados y por supuesto sus propias conclusiones. Adicionalmente deben incluir el por qué se escogió el set de datos arbitrario, qué condiciones y propiedades tiene. El documento debe seguir el template de la IEEE para publicaciones en ingeniería el cual pueden encontrar aquí:

https://www.ieee.org/conferences/publishing/templates.html

Evaluación

Tarea	Puntaje Máximo
Red Wine Dataset	
Regresión Logística - Análisis de resultados	7.5
kNN - Análisis de resultados	7.5
Árbol de decisión - Análisis de resultados	7.5
Red neuronal - Análisis de resultados	7.5
Análisis general Completitud Sobre el set de datos, sobre los features Cúal es el mejor resultado, por qué? Overfitting? Underfitting?	20
Set de datos arbitrario	
Regresión Logística - Análisis de resultados	7.5
kNN - Análisis de resultados	7.5
Árbol de decisión - Análisis de resultados	7.5
Red neuronal - Análisis de resultados	7.5
Análisis general Completitud Sobre el set de datos, sobre los features, por qué se escogió el set de datos? Cúal es el mejor resultado, por qué? Overfitting? Underfitting?	20
TOTAL	100

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. “Modeling wine preferences by data mining from physicochemical properties”. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Instituto Tecnológico de Costa Rica Semestre 1, 2021
Escuela de Ingeniería en Computación
Maestría en Ciencias de la Computación
Curso: Aprendizaje Automático
Profesor: Dr. José Carranza-Rojas
Valor: 15%
Proyecto en parejas