Proyecto de MLLib

Juan M. Alberola, Victor Sánchez

Grado en Tecnologías Interactivas

Recursos

- Documentación de MLlib

Introducción

En este proyecto vas a poner en práctica todo lo que has aprendido sobre Apache Spark para afrontar un proyecto de aprendizaje automático empleando la API de MLlib. Como bien se ha visto, esta API se emplea principalmente para la construcción y evaluación de proyectos de aprendizaje automático. Pese a no tener tantos algoritmos implementados como otras variantes no Big Data como ahora Scikit-Learn o Weka, la ventaja de emplear los modelos incorporados en MLlib es su habilidad para lidiar con volumenes muy grandes de datos (siempre que la infraestructura acompañe, claro está). Para este proyecto, evaluaréis el rendimiento de diversos métodos de aprendizaje automático, y sus respectivas configuraciones en una tarea de clasificación. Para el desarrollo de este proyecto se empleará exclusivamente Apache Spark y su ecosistema de librerías asociadas y compatibles. A continuación detallaremos el trabajo a realizar.

Introducción

En PoliformaT encontraréis, junto a este enunciado, un fichero llamado *train.csv* que contiene el conjunto de datos a emplear para la tarea de clasificación. Concretamente, el conjunto de datos cuenta con 55.532 registros de reservas hoteleras y una etiqueta que marca si la reserva fue finalmente cancelada. El objetivo de la tarea es ser capaz de predecir si una reserva será cancelada o no en base a una serie de características de la reserva. Concretamente, el conjunto de datos cuenta con la información que aparece en el cuadro 1. Algunas filas pueden presentar valores nulos.

El objetivo de este proyecto es encontrar el modelo de aprendizaje automático implementado en MLLib, y su correspondiente configuración, que mejores resultados obtiene atendiendo al *F1 score*. Para ello, tendréis que llevar a cabo experimentación con diferentes modelos y configuraciones. Para encontrar el mejor modelo, se deben seguir una serie de recomendaciones:

- Recomendamos que empleéis validación cruzada para evaluar la calidad conseguida por los hiperparámetros de un modelo.
- Recomendamos que empléeis la búsqueda en rejilla para encontrar la mejor combinación de parámetros de vuestro modelo. Puede que sea interesante hacer diferentes búsquedas en rejilla

Dominio	**Descripción**	**Dominio**
lead_time	Número de días de antelación de la reserva	Entero
arriva_date_week_number	Semana del año en la que comienza la reserva	Entero
stays_in_weekend_nights	Número de noches de fin de semana de la reserva	Entero
stays_in_week_nights	Número de noches entre semana de la reserva	Entero
adults	Número de adultos de la reserva	Entero
children	Número de niños de la reserva	Entero
babies	Número de bebés de la reserva	Entero
meal	Tipo de dieta incluida BB - Desayuno HB - Media pensión FB - Pensión completa Sin categorizar	Categórico
country	País de origen de la reserva	Categórico
market_segment	Origen de la reserva TA - Agente de viaje TO - Operador de tours ...	Categórico
distribution_channel	Canal de distribución de la reserva	Categórico
is_repeated_guest	Si ya ha estado en el hotel	Booleano
previous_cancellations	Número de cancelaciones previas	Entero
previous_bookings_not_canceled	Número de reservas previas no canceladas	Entero
reserved_room_type	Tipo de habitación reservada	Categórico
booking_changes	Número cambios realizados en la reserva	Entero
deposit_type	Si se ha hecho un depósito en la reserva Sin depósito Con depósito Con depósito con devolución	Categórico
days_in_waiting_list	Número de días hasta que se le confirmó la reserva al cliente	Entero
customer_type	Tipo de cliente	Categórico
adr	Coste diario promedio	Real
required_car_parking_spaces	Número de plazas de parking solicitadas	Entero
total_of_special_requests	Número total de peticiones especiales	Entero
is_canceled	Si se canceló la reserva	Booleano

Cuadro 1: Descripción de los datos de entrenamiento

para un mismo modelo si encontramos que el modelo se comporta particularmente bien en un subespacio de las posibles configuraciones probadas.

- Recomendamos que probéis diferentes transformaciones de los atributos de entrada al modelo. Como sugerencias, podéis probar:
  - Escalar los atributos numéricos a una misma escala para una mejor comparabilidad. Existen diferentes alternativas como la normalización entre -1 y 1, la normalización entre 0 y 1, o la normalización a normál estándar (media=0 y desviación típica=1). Podéis explorar los transformadores y estimadores implementados en MLLib para ayudaros en esta tarea[1].
  - Visualizar el histograma de los atributos numéricos para ver su similitud a una normal. En caso de distribuciones asimétricas, a veces ayuda aplicar una transformación como raíz cuadrada ($\sqrt{x}$), inversa ($\frac{1}{x}$), cuadrado, cubo o logaritmo del atributo para hacer que el atributo se asemeje más a una normal. A veces esto ayuda a algunos modelos de aprendizaje automático a encontrar mejores resultados.
  - Algunos modelos no trabajan bien con atributos categóricos, por lo que estos pueden ser transformados a atributos numéricos binarios empleando transformaciones como *OneHotEncoding*.
  - Derivando nuevos atributos como combinación o subconjunto de los existentes (e.g., productos, ratios, reducción de categorías).
  - Reducir el número de características empleando algún selector de características que elimine información redundante o ruido[2].
  - Imputar valores en el caso de valores nulos a otros valores como medias, medianas, o valor del registro más similar.
  - Discretizar atributos continuos.

Los requisitos del proyecto son los siguientes:

- El trabajo se debe realizar **preferiblemente en grupos de tres**.
- Las transformaciones sobre el conjunto de datos, así como el entrenamiento de los modelos debe realizarse integramente en Spark. Podéis apoyaros en otras librerías como Matplotlib o Seaborn para realizar visualizaciones.
- Se debe entregar el **código desarrollado**,en formato *ipynb*, para:
  - El código de todos los experimentos realizados usando validación y búsqueda en rejilla para los diferentes modelos, así como un fichero con los resultados de cada experimento o su salida en el notebook.
  - Las transformaciones necesarias para preparar el conjunto de datos para el entrenamiento (e.g., transformaciones de atributos, selectores de características, filtrados, etc.)

[1] `https://spark.apache.org/docs/latest/ml-features.html`
[2] `https://spark.apache.org/docs/latest/ml-features.html#feature-selectors`

  - Un script final configurado con vuestro mejor modelo que prepare el conjunto de datos de entrada y haga el entrenamiento del modelo con la mejor configuración. Esto es importante, puesto que se realizará un torneo con vuestros mejores modelos para ver quién ha obtenido mejores resultados. La última parte de la nota del proyecto depende directamente de este script y estos resultados. Si no se proporciona, se obtendrá un cero en la rúbrica de este apartado. Para la evaluación del torneo se empleará un conjunto de test diferente al que contáis para el entrenamiento y se usará el F1-score como métrica de calidad. Se puede encontrar un ejemplo en `https://colab.research.google.com/drive/1MmmvyWmgyp6Cb95TLhnudCuILs5X8P-A?usp=sharing`

- Se deberá entregar una **presentación en Powerpoint de 15 minutos** que contenga:
  - Portada con el nombre de la asignatura, título del trabajo e integrantes del equipo.
  - Las transformaciones del conjunto de datos que mejor os han funcionado, describiendo en forma de tubería las transformaciones que se han llevado a cabo y en qué consisten.
  - Una descripción del proceso de entrenamiento que habéis seguido, detallando los modelos que habéis probado así como las configuraciones de hiperparámetros que habéis probado.
  - Una descripción del mejor modelo que habéis encontrado. Se deberá describir la lógica, a alto nivel, de cómo funciona dicho modelo y las bases en las que se fundamenta. Para ello, tendréis que investigar sobre vuestro mejor modelo. También deberéis comentar cuáles han sido los mejores valores de hiperparámetros que habéis encontrado y, aproximadamente, cómo influye cada uno sobre el modelo.
  - Una tabla resumen de los resultados que habéis obtenido y el porqué os habéis decantado por un modelo y no por otros.
  - Bibliografía empleada durante el proyecto y la presentación.
- La entrega tanto de la presentación como del código se realizará en zip a través de PoliformaT antes del **11/06/2025 a las 23:55**.
- La presentación del trabajo se llevará a cabo el **12/06/2025** en horarios de clase (15:00-18:15).

Criterios de evaluación

A continuación se adjunta la rúbrica de evaluación. Cabe destacar que la rúbrica es **orientativa** y es **IMPOSIBLE** que contenga todas las casuísticas que se pueden dar en un proyecto abierto de esta índole.

Aspecto	<50 %	50-69 %	70-89 %	90-100 %
Calidad visual y oral presentación (1 punto)	-Mal uso de elementos gráficos, tablas y figuras -Presentación sin orden y estructura -Faltan índice, portada o bibliografía -Estilo visual no apropiado para la presentación -El estudiante no es claro en su mensaje	Fallan entre 3 y 4 de los siguientes: -Uso adecuado de elementos gráficos, tablas y figuras -Presentación bien definida en secciones -Índice, portada y bibliografía -Estilo visual apropiado para la presentación -El estudiante es claro en su mensaje	Fallan entre 1 y 2 de los siguientes: -Uso adecuado de elementos gráficos, tablas y figuras -Presentación bien definida en secciones -Índice, portada y bibliografía -Estilo visual apropiado para la presentación -El estudiante es claro en su mensaje	-Uso adecuado de elementos gráficos, tablas y figuras -Presentación bien definida en secciones -Índice, portada y bibliografía -Estilo visual apropiado para la presentación -El estudiante es claro en su mensaje
Experimentación (5 puntos)	Fallan más de 3 de los siguientes: - Se han probado diferentes tipos de transformaciones y selectores, y se describen claramente. Estas transformaciones tienen sentido. - Se describen adecuadamente los modelos probados, sus hiperparámetros, y búsquedas en rejilla realizadas - Se describe adecuadamente el análisis de los experimentos realizados - Se describe correctamente y adecuadamente la tubería final de transformaciones llevada a cabo - Se ha realizado un estudio profundo de modelos e hiperparámetros	Fallan 3 de los siguientes: - Se han probado diferentes tipos de transformaciones y selectores, y se describen claramente. Estas transformaciones tienen sentido. - Se describen adecuadamente los modelos probados, sus hiperparámetros, y búsquedas en rejilla realizadas - Se describe adecuadamente el análisis de los experimentos realizados - Se describe correctamente y adecuadamente la tubería final de transformaciones llevada a cabo - Se ha realizado un estudio profundo de modelos e hiperparámetros	Fallan entre 1 y 2 de los siguientes: - Se han probado diferentes tipos de transformaciones y selectores, y se describen claramente. Estas transformaciones tienen sentido. - Se describen adecuadamente los modelos probados, sus hiperparámetros, y búsquedas en rejilla realizadas - Se describe adecuadamente el análisis de los experimentos realizados - Se describe correctamente y adecuadamente la tubería final de transformaciones llevada a cabo - Se ha realizado un estudio profundo de modelos e hiperparámetros	- Se han probado diferentes tipos de transformaciones y selectores, y se describen claramente. Estas transformaciones tienen sentido. - Se describen adecuadamente los modelos probados, sus hiperparámetros, y búsquedas en rejilla realizadas - Se describe adecuadamente el análisis de los experimentos realizados - Se describe correctamente y adecuadamente la tubería final de transformaciones llevada a cabo - Se ha realizado un estudio profundo de modelos e hiperparámetros
Investigación modelo (2 puntos)	Fallan al menos 3 de los siguientes: - Se describe en detalle el funcionamiento e ideas detrás del mejor modelo encontrado. - Se ilustra gráficamente y con ejemplos su funcionamiento - Se describe detalladamente el impacto de los hiperparámetros configurados - La investigación se ha basado en artículos y/o libros académicos	Fallan dos de los siguientes: - Se describe en detalle el funcionamiento e ideas detrás del mejor modelo encontrado. - Se ilustra gráficamente y con ejemplos su funcionamiento - Se describe detalladamente el impacto de los hiperparámetros configurados - La investigación se ha basado en artículos y/o libros académicos	Falla uno de los siguientes: - Se describe en detalle el funcionamiento e ideas detrás del mejor modelo encontrado. - Se ilustra gráficamente y con ejemplos su funcionamiento - Se describe detalladamente el impacto de los hiperparámetros configurados - La investigación se ha basado en artículos y/o libros académicos	- Se describe en detalle el funcionamiento e ideas detrás del mejor modelo encontrado. - Se ilustra gráficamente y con ejemplos su funcionamiento - Se describe detalladamente el impacto de los hiperparámetros configurados - La investigación se ha basado en artículos y/o libros académicos
Resultados torneo (2 puntos)	No se ha entregado código funcional para el torneo	Fuera del top 50 % del torneo	Top 50 % del torneo	Ganador del torneo