MACHINE LEARNING ESCALABLE

PEC – CLOUD

MIGUEL PÉREZ

1. ANÁLISIS

Uno de los grandes proveedores Cloud que existen en el mercado es Google Cloud Platform, y el producto que se pretende analizar en este ejercicio de la PEC es BigQuery Machine Learning, que es una funcionalidad de BigQuery.

BigQuery ML permite crear y ejecutar modelos de aprendizaje automático con consultas de SQL estándar, por lo que ofrece a los profesionales de SQL la posibilidad de compilar modelos de aprendizaje automático con las herramientas y habilidades de SQL existentes. BigQuery es en sí un motor de consulta SQL rápido y un almacén de datos, que permite al usuario centrarse en escribir sentencias SQL en conjuntos de datos, ya sean grandes o pequeños, y permite transformar y alimentar el conjunto de datos directamente en modelos de aprendiza automático, por lo que BigQuery ML aumenta la velocidad de desarrollo gracias a que se elimina la necesidad de mover los datos.

El producto está disponible en diversos elementos como la API REST de BigQuery o herramientas externas tales como Jupyter. Para el uso de esta herramienta con grandes conjuntos de datos se requiere de una programación extensa y conocimiento de los marcos de trabajo del aprendizaje automático, por lo que se restringe el desarrollo de soluciones solo para aquellos usuarios que cumplan con dichos requisitos.

Los modelos generados con BigQuery ML representan lo que un sistema de aprendizaje automático aprendió de los datos de entrenamiento, pudiendo usar un modelo con datos de varios conjuntos de datos de BigQuery para el entrenamiento y la predicción, y dichos modelos se pueden exportar posteriormente a Cloud AI Platform o a una capa de servivio propia para poder predecir datos de forma online. Es compatible con muchos tipos de modelos como la regresión lineal, la regresión logística (binaria o multiclase), K-Means, series temporales, XGBoost, redes neuronales, AutoML Tables o con modelos importados de TensorFlow.

Las principales ventajas que ofrece el uso de esta herramienta en comparación con el uso de otras herramientas de aprendizaje automático que hacen uso de almacenes de datos en la nube son:

- Democratizar el uso del aprendizaje automático al permitir compilar y ejecutar modelos mediante hojas de cálculo y herramientas de inteligencia comercial existentes.

- No es necesario programar con Python o Java, sino con SQL.

- Aumenta la velocidad de desarrollo al eliminar la necesidad de exportar los datos.

Finalmente, un proceso end-to-end usando esta herramienta podría ser el mostrado en la siguiente imagen:

2. EJEMPLO DE USO

Como se ha comentado en el apartado anterior, BigQuery ML es compatible con multitud de modelos y uno de ellos son las series temporales, por lo que permite hacer predicciones sobre todo tipo de series temporales. Se puede encontrar un ejemplo ilustrativo en la documentación de BigQuery ML que se resume a continuación pero que puede ser aplicable a cualquier otro conjunto de datos temporales.

En el tutorial se hace uso de una tabla que contiene información sobre un parte de los datos de sesión que se recopiló con Google Analytics. Una vez creado el conjunto de datos de BigQuery donde se almacenará el modelo de aprendizaje automático, un primer paso que se suele dar es la visualización de la serie temporal, realizando posteriormente las modificaciones necesarias a los datos para adaptarlos al modelo.

Cuando los datos estén preparados se puede crear un modelo con la sentencia CREATE MODEL, como puede ser el modelo ARIMA ya que se está tratando con datos temporales, dando la posibilidad de que se genere un modelo automático, o ajustar los hiperparámetros manualmente. Con el modelo generado, se puede usar la función ML.ARIMA_EVALUATE para comprobar las métricas de todos los modelos que se evaluaron mediante el ajuste de hiperparámetros si se escogió el automático o si se incluyeron diferentes combinaciones de hiperparámetros, pudiendo profundizar posteriormente en los parámetros del modelo escogido.

El siguiente paso sería el objetivo por el cuál se construye el modelo, que es la predicción futura de los datos, para lo cual se puede usar la función ML.FORECAST, la cuál predice los

valores de series temporales futuras con un intervalo de predicción mediante el modelo generado.

De esta forma, se puede generar un modelo para series temporales con cualquier tipo de dato temporal y conseguir realizar predicciones futuras con el histórico de datos, con multitud de casos de uso como, por ejemplo:

- En el campo de la economía predecir la tasa de inflación o el PIB.
- Predicción de temperaturas máximas o mínimas, precipitaciones diarias, etc.
- Predicción de emisiones de gases contaminantes anuales.

3. REFERENCIAS

Las principales referencias consultadas han sido:

- Apuntes del curso Google Cloud Platform Big Data and Machine Learning Fundamentals que realicé en Mayo de 2020: https://www.coursera.org/learn/gcp-big-data-ml-fundamentals
- Documentación introductoria de BigQuery ML: https://cloud.google.com/bigquery-ml/docs/introduction
- Tutorial Prever una serie temporal a partir de datos de Google Analytics: https://cloud.google.com/bigquery-ml/docs/arima-single-time-series-forecasting-tutorial