1 y 2

TEMAS

MÓDULO:
PROJECT MANEGEMENT

VALIDACIÓN E IMPLEMENTACIÓN DE MODELOS

PARFAIT ATCHADE ADELOMOU
MBAs y Ingeniero Superior de Telecomunicaciones.
PhD Candidate en Computación cuántica

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Entender la ingeniería y la validación de los modelos.
- Entender las fases de Diseño de aplicaciones impulsadas por Machine Learning, su experimentación, implementación, las operaciones asociadas con un buen despliegue y su posterior mantenimiento.

1. VISIÓN GENERAL DE GESTIÓN DE PROYECTOS

Los conceptos necesarios para la gestión de proyectos basada en ML (La **Integración Continua** (CI)) es el proceso encargado de probar y validar el código y los componentes, y de, probar y validar datos, esquemas de datos y modelos.

La **Entrega Continua** (CD) es el sistema (una canalización/pipeline de entrenamiento de ML) que debe implementar otro servicio (servicio de predicción del modelo) de manera automática.

El **Entrenamiento Continuo** (CT) es el proceso definido para los sistemas de ML que se ocupa de volver a entrenar y entregar los modelos de forma automática.

2. INGENIERÍA DEL APRENDIZAJE AUTOMÁTICO

EXTRACCIÓN DE DATOS

Selecciona y, luego, integra los datos relevantes de varias fuentes de datos para la tarea de ML

ANÁLISIS DE DATOS

Realiza un análisis de datos exploratorios (EDA) para identificar los datos disponibles con el fin de compilar el modelo de ML. La compresión de las características y el esquema de datos que espera el modelo. La identificación de la preparación de datos y la ingeniería de atributos que se necesitan para el modelo.

PREPARACIÓN DE DATOS

La limpieza de datos, en la que se dividen los datos en conjuntos de entrenamiento, validación y pruebas.

ENTRENAMIENTO DE MODELOS

El científico de datos implementa algoritmos diferentes con los datos preparados para entrenar varios modelos de ML. Además, debe someter a los algoritmos implementados al ajuste de hiperparámetros para obtener el modelo de ML de mejor rendimiento, es decir, conseguir tener un modelo entrenado.

EVALUACIÓN DE MODELOS

El modelo se analiza en un conjunto de pruebas de exclusión para evaluar la calidad del modelo y obtener las métricas para evaluar la calidad del modelo.

VALIDACIÓN DE MODELOS

Se confirma que el modelo es adecuado para la implementación si su rendimiento predictivo es mejor que la referencia.

ENTREGA DE MODELOS

Se implementa el modelo validado en un entorno de destino a fin de entregar predicciones: microservicios con una API de REST para entregar predicciones en línea. Un modelo incorporado a un borde o dispositivo móvil parte de un sistema de predicción por lotes.

SUPERVISIÓN DEL MODELO

Se supervisa el rendimiento predictivo del modelo para realimentar, en caso necesario, el proceso de ML.

3. MONITORIZACIÓN DEL RENDIMIENTO

En cualquier proyecto de ML, después de definir el caso de uso empresarial y **establecer los KPIs de éxito,** el **proceso de entrega** (a producción) implica **definir unos pasos** que pueden ser **manuales** o **automáticos** en función de la madurez de las aplicaciones.

4. CONCLUSIÓN Y TOMA DE DECISIÓN

La implementación de ML en un entorno de producción no solo implica implementar su modelo como una API para la predicción, en realidad, significa implementar una canalización de ML que pueda automatizar el reentrenamiento y la implementación de modelos nuevos.

La configuración de un sistema de CI/CD permite probar y establecer, de forma automática, implementaciones de canalización nuevas.

Este sistema permite lidiar con los cambios rápidos en los datos y el entorno empresarial.

No es necesario mover, de inmediato, todos los procesos de un nivel a otro.

Lo adecuado es poder implementar estas prácticas de forma gradual para que el científico de datos pueda mejorar la automatización del desarrollo y la producción del sistema de ML.

IDEAS CLAVE

- El concepto de gestión de la operacionalización del modelo de aprendizaje automático **(MLOps)** proporciona un proceso de desarrollo de aprendizaje automático de extremo a extremo para **diseñar, construir y administrar software reproducible, comprobable y evolutivo** impulsado por ML. El flujo de trabajo de aprendizaje automático típico consta de tres fases principale**s.**
  - **Ingeniería de datos: adquisición y preparación de datos.**
  - **Ingeniería de modelos de ML: capacitación y servicio de modelos de ML.**
  - **Ingeniería de código: integración del modelo ML en el producto final.**
- **Entender la naturaleza del problema: regresión vs. clasificación. Variables dependientes e independientes, necesidad de feature engineering.** Entender la precisión requerida por el problema, búsqueda de soluciones simples, uso de algoritmos con alta capacidad de generalización (Redes Neuronales, XGBoost u otros Boosted trees) y tener como referencia el **cheat sheets.**
- Dominar las métricas de validación de modelos basados en regresión y clasificación. Y saber en todo momento cuál usar.