Procesos de Machine Learning

Profesor: Juan Gamarra Moreno

Proceso de Machine Learning

- Un **proceso de Machine Learning** es un conjunto de pasos estructurados que permiten a una máquina aprender a partir de datos y realizar predicciones o tomar decisiones sin estar explícitamente programada para cada tarea.

Proceso y Metodología

- La diferencia entre un **proceso** y una **metodología** radica en la naturaleza y propósito de cada término, lo que también explica por qué se usa "proceso" en lugar de "metodología" en el contexto de Machine Learning.

Proceso y Metodología

1. Proceso:

- Un **proceso** es una secuencia de pasos **estructurados** y **predefinidos** que deben seguirse para lograr un objetivo específico.
- En el caso de Machine Learning, el proceso se refiere a las **etapas claramente definidas** (como la recolección de datos, el entrenamiento, la evaluación, etc.) que deben realizarse de manera sistemática y secuencial para desarrollar y desplegar un modelo de Machine Learning.
- Los **procesos son repetibles**: Puedes seguir el mismo conjunto de pasos una y otra vez para diferentes proyectos de Machine Learning.
- **Ejemplo**: El proceso de CRISP-DM para proyectos de Machine Learning, que define pasos desde la comprensión del negocio hasta el despliegue del modelo.

Proceso y Metodología

2. Metodología:

- Una **metodología**, por otro lado, es un **marco teórico** o enfoque que guía cómo se deben realizar las cosas. Es una estrategia general que te indica qué pasos **deberían** tomarse, pero no te da detalles estrictos de cada paso.
- Las metodologías tienden a ser más **flexibles** y más **conceptuales**. Te proporcionan una filosofía o enfoque para realizar el trabajo, pero no dictan una serie específica de acciones.
- **Ejemplo**: La metodología ágil en desarrollo de software te da principios generales para gestionar proyectos, pero no especifica un conjunto fijo de pasos como un proceso.

Proceso y Metodología

¿Por qué se llama "proceso" en Machine Learning?

- El término "proceso" es más adecuado en Machine Learning porque se trata de un conjunto de **pasos bien definidos y estructurados** que son **repetibles** y **específicos** para la creación de modelos. Cada proyecto sigue más o menos la misma secuencia de tareas (recolección de datos, preprocesamiento, entrenamiento, etc.), independientemente del tipo de problema.
- Si se utilizara "metodología", implicaría que el enfoque es más **abstracto** o **filosófico**, lo cual no refleja con precisión el carácter práctico, estructurado y repetitivo de los proyectos de Machine Learning.

Proceso y Metodología

Resumen:

- Se usa "proceso" en lugar de "metodología" en Machine Learning porque los pasos son **claros, secuenciales, repetibles** y están orientados a una **acción específica** (desarrollar y desplegar un modelo). Una "metodología" sería más teórica y flexible, mientras que un proceso garantiza que se sigan ciertas etapas para obtener resultados consistentes.

Etapas del Proceso de Machine Learning

1. Definición del Problema

El primer paso es comprender y formular el problema que se quiere resolver con Machine Learning. Esto implica identificar las metas del negocio o investigación, los datos disponibles, y qué tipo de problema se aborda (clasificación, regresión, clustering, etc.).

Etapas del Proceso de Machine Learning

2. Recolección y Preparación de Datos

- Requiere recopilar datos relevantes para el problema. Esto incluye datos históricos o datos generados específicamente para el modelo.
- **Preprocesamiento de datos**: Los datos crudos suelen contener errores, valores faltantes o características irrelevantes. El preprocesamiento incluye la limpieza de datos, normalización, codificación de variables categóricas, y la división de los datos en conjuntos de entrenamiento, validación y prueba.
- Este enfoque garantiza un desarrollo iterativo y robusto en la creación de modelos de Machine Learning.

Etapas del Proceso de Machine Learning

3. Selección del Modelo

- Se elige el algoritmo o modelo que mejor se ajusta al problema. La elección puede depender del tipo de datos y de la naturaleza del problema (modelos supervisados, no supervisados, de refuerzo, etc.). Ejemplos de algoritmos incluyen regresión lineal, árboles de decisión, redes neuronales, entre otros.

Etapas del Proceso de Machine Learning

4. Entrenamiento del Modelo

- En esta etapa, el modelo se entrena utilizando los datos de entrenamiento. El algoritmo ajusta sus parámetros internos para minimizar una función de pérdida, que mide la precisión del modelo en la tarea específica.

Etapas del Proceso de Machine Learning

5. Evaluación del Modelo

- Se evalúa el rendimiento del modelo utilizando los datos de validación o prueba. Métricas comunes incluyen la precisión, recall, F1-score, o el error cuadrático medio, dependiendo del tipo de problema. Esta etapa puede revelar si el modelo está sufriendo de sobreajuste o subajuste.

Etapas del Proceso de Machine Learning

6. Ajuste del Modelo (Optimización de Hiperparámetros)

- Se ajustan los hiperparámetros del modelo para mejorar su rendimiento. Los hiperparámetros no se aprenden durante el entrenamiento y deben seleccionarse a través de métodos como la búsqueda en cuadrícula o la búsqueda aleatoria.

Etapas del Proceso de Machine Learning

7. Despliegue del Modelo

- Una vez que el modelo tiene un rendimiento aceptable, se despliega para su uso en entornos de producción donde puede hacer predicciones o clasificaciones en nuevos datos.

Etapas del Proceso de Machine Learning

8. Monitoreo y Mantenimiento

- Los modelos de Machine Learning deben monitorearse de manera continua para garantizar que su desempeño no se degrade con el tiempo, debido a cambios en los datos o en las condiciones bajo las cuales fueron entrenados. Esto puede requerir el retraining periódico.

Etapas del Proceso de Machine Learning

Resumen del Proceso:

1. Definición del problema
2. Recolección y Preparación de datos
3. Selección del modelo
4. Entrenamiento
5. Evaluación
6. Optimización
7. Despliegue
8. Monitoreo

Este enfoque garantiza un desarrollo iterativo y robusto en la creación de modelos de Machine Learning.

Procesos Estandarizados

- Existen varios procesos y marcos de trabajo que se han estandarizado para guiar el desarrollo de proyectos de Machine Learning de manera estructurada. Presentaremos algunos de los más conocidos y utilizados.

Procesos Estandarizados

1. CRISP-DM (Cross-Industry Standard Process for Data Mining)

- **Fases principales**:
  - **Entendimiento del negocio**: Definir los objetivos del negocio y formular preguntas clave.
  - **Entendimiento de los datos**: Inspección inicial de los datos para familiarizarse con ellos.
  - **Preparación de los datos**: Limpieza, transformación y selección de los datos relevantes.
  - **Modelado**: Selección y aplicación de algoritmos de Machine Learning.
  - **Evaluación**: Validar el rendimiento del modelo para asegurarse de que cumple con los objetivos del negocio.
  - **Despliegue**: Implementar el modelo en un entorno de producción.
- CRISP-DM es uno de los enfoques más populares en la industria y es aplicable a proyectos de Machine Learning y minería de datos.

Procesos Estandarizados

2. KDD (Knowledge Discovery in Databases)

Este proceso, centrado en el descubrimiento **de conocimiento** en bases de datos, es uno de los más antiguos y se utiliza comúnmente en proyectos de minería de datos y Machine Learning.

- **Fases principales**:
  - **Selección**: Extraer datos relevantes de una base de datos más grande.
  - **Preprocesamiento**: Limpieza y preparación de datos.
  - **Transformación**: Reducción o proyección de los datos para adaptarlos a los modelos.
  - **Data Mining**: Aplicación de algoritmos de Machine Learning para encontrar patrones.
  - **Interpretación/Evaluación**: Evaluar y extraer el conocimiento útil de los resultados.

Procesos Estandarizados

3. TDSP (Team Data Science Process)

- Es un proceso desarrollado por Microsoft para proporcionar un marco estructurado a los equipos de ciencia de datos.
- **Fases principales**:
  - **Planificación**: Definir los objetivos del proyecto.
  - **Adquisición y preparación de datos**: Obtener, limpiar y transformar datos.
  - **Modelado**: Construir y entrenar los modelos de Machine Learning.
  - **Implementación**: Desplegar el modelo en un entorno de producción.
  - **Monitoreo y mantenimiento**: Asegurar que el modelo sigue siendo efectivo con datos nuevos.

Procesos Estandarizados

4. OSEM (Operational Stages of Enterprise Machine Learning)

- Es un enfoque de IBM basado en la operación y el monitoreo de modelos de Machine Learning.
- **Fases principales**:
  - **Desarrollo**: Construcción y entrenamiento del modelo.
  - **Validación**: Verificación de que el modelo cumple con los requisitos.
  - **Despliegue**: Implementar el modelo en un entorno de producción.
  - **Monitoreo**: Supervisar el rendimiento del modelo.
  - **Actualización**: Retrain y ajuste del modelo cuando sea necesario.

Procesos Estandarizados

5. Google Cloud AI Platform Workflow

- Google ofrece su propio flujo de trabajo para proyectos de Machine Learning, especialmente en la nube.
- **Fases principales**:
  - **Definir el objetivo**: Entender el problema que se intenta resolver.
  - **Preparar los datos**: Limpieza y transformación.
  - **Construir y entrenar modelos**: Seleccionar modelos y entrenarlos con datos.
  - **Evaluar y ajustar**: Medir el rendimiento y ajustar el modelo.
  - **Implementar y monitorizar**: Desplegar el modelo y realizar monitoreo constante.

Procesos Estandarizados

6. MLOps (Machine Learning Operations)

- MLOps es un enfoque centrado en la **automatización y la operación** continua de modelos de Machine Learning, muy parecido al concepto de DevOps.
- **Fases principales**:
  - **Desarrollo del modelo**: Crear, entrenar y validar los modelos.
  - **Despliegue**: Automatizar el proceso de despliegue de modelos.
  - **Monitorización**: Supervisar el rendimiento del modelo y su comportamiento en producción.
  - **Ciclo de retroalimentación**: Ajustar, entrenar y mejorar continuamente los modelos basados en nuevos datos.

Procesos Estandarizados

7. SEMMA (Sample, Explore, Modify, Model, Assess)

- Es un proceso desarrollado por SAS para proporcionar una estructura clara y eficiente en proyectos de minería de datos y Machine Learning, especialmente cuando se manejan grandes volúmenes de información.
- **Fases principales:**
  - **Sample (Muestreo)**: Extraer un subconjunto representativo de los datos.
  - **Explore (Explorar)**: Realizar un análisis exploratorio de los datos.
  - **Modify (Modificar)**: Limpiar, transformar y preparar los datos para el modelado..
  - **Model (Modelar)**: Construir y entrenar los modelos de Machine Learning.
  - **Assess (Evaluar)**: Evaluar el rendimiento del modelo.
- Aunque SEMMA no incluye explícitamente una fase de monitoreo y mantenimiento, en la práctica, los modelos deben ser monitoreados y ajustados si los datos cambian con el tiempo.

SEMMA

- SEMMA (Sample, Explore, Modify, Model, Assess) es un proceso desarrollado por SAS para guiar los proyectos de minería de datos y Machine Learning. Es una metodología que se centra en el manejo y análisis de grandes conjuntos de datos y sigue una estructura secuencial clara para garantizar que los resultados sean sólidos y accionables.

Fases principales de SEMMA

1. **Sample (Muestreo)**:
   - **Seleccionar un subconjunto representativo de los datos**: Esta fase implica extraer una muestra de los datos, lo que permite a los científicos de datos trabajar con un conjunto más manejable y representar bien el comportamiento general de los datos.
   - **Subconjuntos de entrenamiento y prueba**: Se definen las particiones del conjunto de datos (por ejemplo, 70% para entrenamiento, 30% para pruebas).

Fases principales de SEMMA

2. **Explore (Explorar)**:
   - **Análisis exploratorio de datos**: En esta fase, se realiza un análisis inicial para identificar patrones, relaciones y anomalías en los datos.
   - **Visualización y estadística descriptiva**: Se utilizan gráficos y métricas descriptivas para detectar correlaciones o tendencias en los datos, lo que facilita la comprensión de su estructura.

Fases principales de SEMMA

3. **Modify (Modificar)**:
   - **Preprocesamiento y transformación de datos**: Los datos se limpian y transforman según sea necesario, lo que incluye manejar valores faltantes, crear nuevas variables, eliminar o tratar outliers, y transformar los datos para que sean adecuados para los algoritmos de Machine Learning.
   - **Selección de variables**: Se identifican las variables más importantes o relevantes para el modelado, eliminando las que no aportan valor o son redundantes.

Fases principales de SEMMA

4. **Model (Modelar)**:
   - **Entrenamiento del modelo**: Se seleccionan y aplican algoritmos de Machine Learning adecuados para el problema (por ejemplo, regresión, árboles de decisión, redes neuronales, etc.) utilizando los datos modificados.
   - **Ajuste del modelo**: Los parámetros del modelo se ajustan para mejorar el rendimiento en la fase de evaluación.

Fases principales de SEMMA

5. **Assess (Evaluar)**:
   - **Evaluación del rendimiento del modelo**: Se evalúa la precisión del modelo utilizando métricas como la exactitud, error, precisión, recall, o F1-score. Esto se hace en los datos de prueba o de validación.
   - **Validación del modelo**: Esta fase verifica que el modelo es efectivo y cumple con los objetivos planteados, garantizando que sea robusto y generalizable a datos nuevos.

Fases principales de SEMMA

Beneficios del proceso SEMMA:

- **Eficiencia en el manejo de datos**: Al comenzar con un subconjunto de datos, permite manejar grandes volúmenes de información de manera más eficiente.
- **Enfoque exploratorio**: Dedica una fase completa a la exploración de datos, lo que ayuda a obtener insights valiosos antes del modelado.
- **Flexibilidad en la modificación**: Proporciona un marco claro para transformar los datos y seleccionar las variables más importantes antes de construir los modelos.

Este proceso está enfocado en el análisis de datos y el desarrollo de modelos predictivos sólidos, con énfasis en la preparación y transformación de los datos.