2.1 OOP in ML Pipeline Refactoring

POSTGRAD
MNA

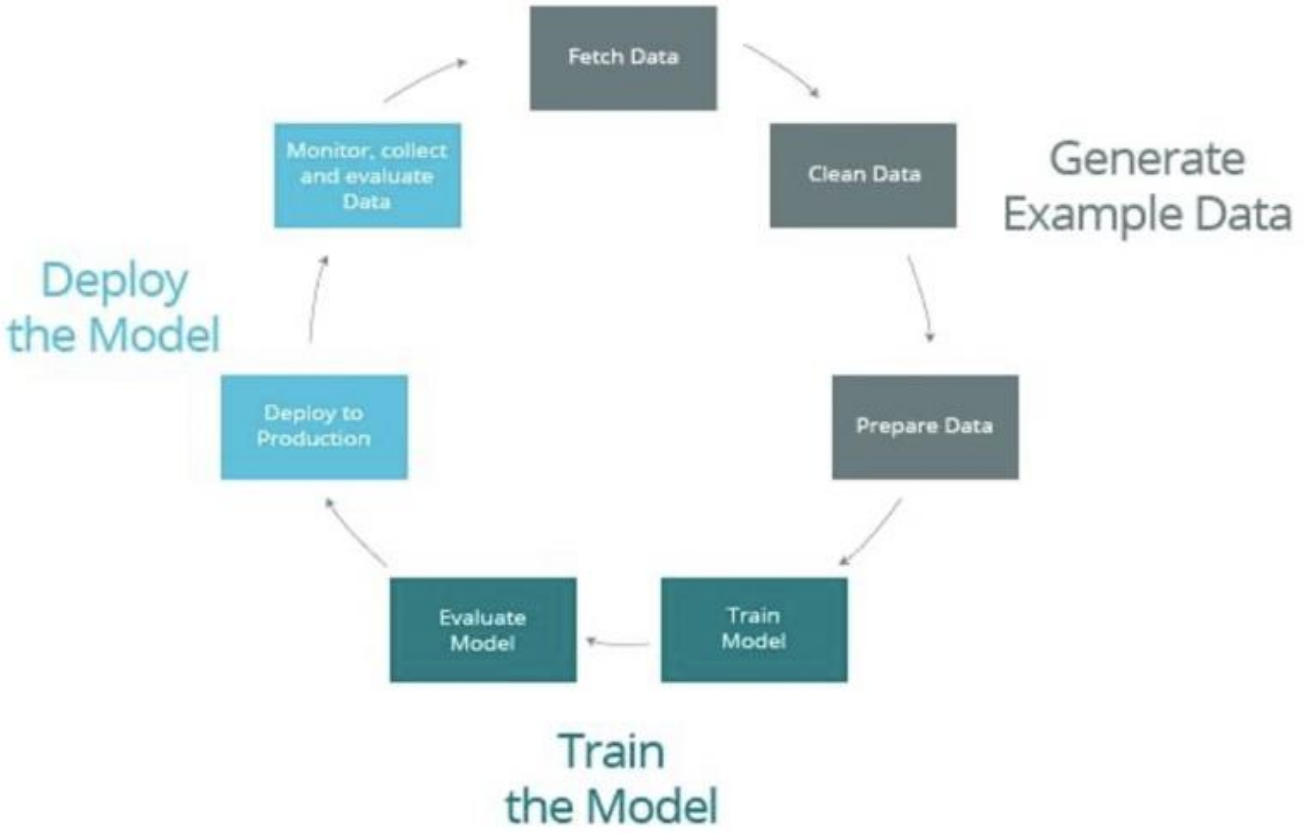
Introduction

Code refactoring is crucial in machine learning projects because it enhances the quality, maintainability, and scalability of the codebase. In ML development, initial models are often created quickly, focusing on experimentation and testing hypotheses. However, as projects grow, this approach can lead to code that is difficult to manage, test, or scale. Refactoring transforms this prototype-level code into modular, reusable, and efficient components, improving performance and making it easier to integrate new features or models. Additionally, clean and structured code simplifies debugging and testing, ensures consistent performance across environments, and enhances collaboration, especially when working with large teams or deploying to production environments. In the context of MLOps, refactoring supports best practices like version control, reproducibility, and automated pipelines, enabling smoother transitions from research to production.

2.1.1 Structure of ML Projects

ML project stages/workflow

- Data preparation
- Exploratory Data Analysis
- Feature Engineering
- Model Training
- Model Evaluation
- Deployment
- Monitoring



ML Cookie Cutter

- Provides structure for ML projects.

```
├── LICENSE          <- Open-source license if one is chosen
├── Makefile         <- Makefile with convenience commands like `make data` or `make train`
├── README.md        <- The top-level README for developers using this project.
├── data
│   ├── external     <- Data from third party sources.
│   ├── interim      <- Intermediate data that has been transformed.
│   ├── processed    <- The final, canonical data sets for modeling.
│   └── raw          <- The original, immutable data dump.
│
├── docs             <- A default mkdocs project; see www.mkdocs.org for details
│
├── models           <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks        <- Jupyter notebooks. Naming convention is a number (for ordering),
│                       the creator's initials, and a short `-` delimited description, e.g.
│                       `1.0-jqp-initial-data-exploration`.
│
├── pyproject.toml   <- Project configuration file with package metadata for
│                       mlops and configuration for tools like black
│
├── references       <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports          <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures      <- Generated graphics and figures to be used in reporting
│
├── requirements.txt <- The requirements file for reproducing the analysis environment, e.g.
│                       generated with `pip freeze > requirements.txt`
│
├── setup.cfg        <- Configuration file for flake8
│
└── mlops  <- Source code for use in this project.
    │
    ├── __init__.py          <- Makes mlops a Python module
    │
    ├── config.py            <- Store useful variables and configuration
    │
    ├── dataset.py           <- Scripts to download or generate data
    │
    ├── features.py          <- Code to create features for modeling
    │
    ├── modeling
    │   ├── __init__.py
    │   ├── predict.py       <- Code to run model inference with trained models
    │   └── train.py         <- Code to train models
    │
    └── plots.py             <- Code to create visualizations
```

https://github.com/jeannefukumaru/cookiecutter-ml

Importance of repeatable workflows for scaling ML

- **Consistency in Processes:** Ensures data preparation, model training, and evaluation are reproducible.
- **Automation for Scalability:** Automating workflows allows projects to grow without requiring manual intervention.
- **Efficient Experimentation**: Repeatable workflows make it easier to test and compare different models or datasets.
- **Collaboration & Debugging**: Team members can rely on consistent workflows to debug and collaborate effectively.
- **Smoother Production Deployment**: Clear, repeatable workflows reduce risks and errors when moving from development to production.

Challenges of using notebooks in production

- **Lack of Modularity**: Notebooks often combine code, data, and logic in a single place, making it difficult to reuse or modify specific components independently.
- **Poor Version Control**: Tracking changes in notebooks is challenging since outputs and code cells can be modified out of order, complicating collaboration and code history.
- **Difficulty in Testing and Debugging:** Notebooks are designed for interactive exploration, making it hard to write unit tests or systematically debug complex code.
- **Scalability Issues:** Notebooks are not well-suited for handling large-scale data processing or training models at scale due to their linear execution flow and memory management.
- **Hard to Integrate with CI/CD Pipelines:** Continuous integration and deployment pipelines rely on modular, scriptable code, which is harder to implement in the notebook format.
- **Inconsistent Environments:** Running notebooks across different environments can lead to inconsistencies in libraries or dependencies, causing reproducibility problems in production.

2.1.2 OOP Fundamentals

Introduction to Object Oriented Programming (OOP)

- **Programming Paradigm:** OOP is a programming paradigm that organizes software design around objects, which represent real-world entities and data, instead of focusing solely on functions and logic.
- **Invented in the 1960s:** The concept of OOP was introduced in the 1960s with the creation of the programming language Simula, designed for simulations and modeling complex systems.
- **Popularized by Languages like C++, Java:** OOP gained widespread use in the 1980s and 1990s, with languages like C++ and Java, making it the dominant approach in software engineering.
- **Supported on Python:** Allowing developers to create classes and objects and implement OOP principles.
- **Key Concept Objects and Classes:** Objects represent real-world entities or concepts, and classes define the blueprint for these objects. This makes it easier to model complex systems.
- **Shift from Procedural Programming:** OOP emerged as an evolution from procedural programming, which focuses on functions and procedures. OOP allows for better data management and modularity.
- **Advantages of OOP:** Provides better code organization, promotes reusability, and simplifies debugging and maintenance, making it ideal for large-scale software projects.

Class

- A class contains the blueprints or the prototype from which the objects are being created.
- It is a logical entity that contains some attributes and methods.
- Classes are created by keyword class.
- Attributes are the variables that belong to a class.
- Attributes are always public and can be accessed using the dot (.) operator. i.e. Class.Attribute
- In Python, class methods must have an extra self first parameter in the method definition. No value for this parameter is given when the method is called, it is provided by Python. It is an auto reference.
- Clases have __init__ method similar to constructors in C++. It is run as soon as an object of a class is instantiated. The method is useful to do any initialization for the object.

```
#Python Class declaration
#example:

class ClassName:
  # Statement-1
  .
  .
  .
```

Object

- The object is an entity that has a state and behavior associated with it.
- It resembles a real-world object abstraction.
- Consists of state, behavior and ID.
- It is the instance of a class.

```
# Python program to
# demonstrate defining
# a class and
# instantiating it

class SomeClass:
    pass

object = SomeClass()
```

Inheritance

- Allowing new classes to derive from existing ones, promoting code reuse and extensibility.
- The class that derives properties is called the derived class or **child** class and the class from which the properties are being derived is called the base class or **parent**.
- Inheritance could be simple (inherit from a single class) or multiple (inherit from multiple classes simultaneously).

```
#Python single inheritance
class DerivedClassName(BaseClassName):
    <statement-1>
    .
    .
    .
    <statement-N>
```

```
#Python multiple inheritance
class DerivedClassName(Base1, Base2, Base3):
    <statement-1>
    .
    .
    .
    <statement-N>
```

Polymorphism

- Enabling objects of different types to be accessed through the same interface.
- Supporting flexibility in program design.
- Example shows how subclasses can override methods defined in their parent class to provide specific behavior.

```
#Python example of polymorphism and
#Inheritance

class Bird:

   def intro(self):
      print("There are many types of birds.")

   def flight(self):
      print("Most of the birds can fly but some cannot.")

class sparrow(Bird):

   def flight(self):
      print("Sparrows can fly.")

class ostrich(Bird):

   def flight(self):
      print("Ostriches cannot fly.")
```

https://www.geeksforgeeks.org/python-oops-concepts/

Encapsulation

- Bundling data and methods together in a single class, hiding internal details and exposing only what's necessary.
- It describes the idea of wrapping data and the methods that work on data within one unit.
- Puts restrictions on accessing variables and methods directly and can prevent the accidental modification of data.
- An object’s private variable can only be changed by an object’s method.

```
# Python program to
# demonstrate private members
# "__" double underscore represents private attribute.
# Private attributes start with "__".

# Creating a Base class
class Base:
    def __init__(self):
        self.a = "GeeksforGeeks"
        self.__c = "GeeksforGeeks"

# Creating a derived class
class Derived(Base):
    def __init__(self):

        # Calling constructor of
        # Base class
        Base.__init__(self)
        print("Calling private member of base class: ")
        print(self.__c) #Will rise error
```

https://www.geeksforgeeks.org/python-oops-concepts/

Abstraction

- Hiding unnecessary complexity by exposing only essential features.
- Reducing the burden on developers to understand implementation details.

```
class Rectangle:
    def __init__(self, length, width):
        self.__length = length  # Private attribute
        self.__width = width    # Private attribute

    def area(self):
        return self.__length * self.__width

    def perimeter(self):
        return 2 * (self.__length + self.__width)


rect = Rectangle(5, 3)

# Output: Area: 15
print(f"Area: {rect.area()}")

# Output: Perimeter: 16
print(f"Perimeter: {rect.perimeter()}")

# print(rect.__length)
# This will raise AttributeError
# as length and width are private attributes
```

https://www.geeksforgeeks.org/python-oops-concepts/

Why is OOP important for refactoring?

- **Modularity**: OOP breaks code into modular components (classes), making it easier to maintain and update.
- **Reusability**: Common functionality is encapsulated in reusable classes, reducing redundancy and simplifying updates.
- **Scalability**: OOP structures allow ML pipelines to scale by separating concerns into distinct objects.
- **Testability**: OOP makes code easier to test by isolating components, enabling targeted unit tests.

Benefits of OOP in ML Projects

- **Improved Code Organization**: OOP helps organize code by separating concerns into classes, making it more structured and maintainable.
- **Enhanced Collaboration**: Teams can work on separate modules or classes independently, improving productivity.
- **Easier Transition to Production**: OOP facilitates moving from experimentation to production by offering modular and testable code.
- **Integration with MLOps Tools**: OOP fits well into CI/CD pipelines and integrates smoothly with tools like MLFlow for tracking experiments and models.

2.1.3 Refactoring

Why is important to refactor for ML Pipelines

- Improving structure without changing behavior.
- Cleaner, **modular code**.
- Easier to maintain, test, and deploy.
- Scalability!!!
- When deploying a model, there are several possible scenarios:
  - One model deployed on one server
  - One model deployed on multiple servers
  - Multiple versions of a model deployed on one server
  - Multiple versions of a model deployed on multiple servers
  - Multiple versions of multiple models deployed on multiple servers

Why Object Oriented Programming for Refactoring?

- Modularity, flexibility, reusability, and scalability naturally provided by OOP.
- Code organization: Preprocessing, training, evaluation in separate classes.
- Cleaner, more resilient code.
- Easier maintenance, testing.
- Version control for models and different pipeline components.
- Improve reproducibility and experiment management.

Challenges of Refactoring Jupyter Notebooks

- **Non-modular Code**: Jupyter notebooks often mix data, logic, and presentation, making refactoring challenging.
- **Linear Execution Flow**: Notebooks are designed for interactive workflows, which don’t align well with production needs.
- **Difficult to Debug and Test**: Debugging and unit testing are more complex due to the lack of structured code in notebooks.
- **Dependency Management**: Notebooks can suffer from inconsistent dependencies, making reproducibility difficult.

Converting Notebooks into Functions

- Refactor reusable code blocks into standalone Python functions.
- Separating by functionality based on ML Workflow:
  - Data preparation
  - Exploratory Data Analysis
  - Feature Engineering
  - Model Training
  - Model Evaluation
  - Deployment
  - Monitoring

Modular Design: Breaking Down into ML Pipeline Components

- Use different code clases for different ML pipeline components: preprocessing, training, evaluation, etc.
- Modular design allows easy extension, scaling and reuse.

Identifying Code Smells in ML Notebooks

- **Common anti-patterns of Jupyter Notebooks**:
  - Redundant cells
  - Hardcoding
  - Copy-pasting
  - Complex execution sequence (individual cells)
  - Possible mixed dependencies (in different cells)
  - Sequential programming paradigm (in many cases)

- **Problems**: Difficult to scale, maintain, and debug

Appling OOP concepts to ML pipelines

- Designing a Simple OOP Structure:
  - Sample class design for ML pipeline: DataLoader, Preprocessor, Model, Evaluator, etc.
  - Relationships between classes and pipeline stages.
- Applying Encapsulation:
  - Encapsulating data preprocessing inside a specific class.
  - Encapsulating model training inside a specific class.
  - Improved code readability, separation of logic.
- Handling Configuration with OOP:
  - Use a Config class to manage hyperparameters and file paths
  - I.e. refactor a hardcoded script to use a configuration object that contains hyperparameters for hyperparameter tunning.

Example of Refactoring

- Refactoring a Jupyter notebook into OOP-based code.
- Break down a notebook into classes for different ML pipeline states.

(Demonstration of Activity code)

For Next Week:
MLFlow: tool for ML life cycle management (Pyton pipeline management)

- https://mlflow.org/docs/latest/recipes.html#recipe-template-structure

Supporting Material

- https://docs.python.org/3/tutorial/classes.html
- https://mlflow.org/docs/latest/introduction/index.html#core-components-of-mlflow
- Book: “Practical MLOps, O’Reilly, Chapter 11.