Machine Learning(ML) Basics

Tessema Mengistu (Ph.D.)

mengistu@cs.vt.edu

Outline

- Overview of ML
- Classification of ML
- Issues in ML

Overview of ML

- Machine learning
  - Study of algorithms and statistical models to solve problems by inference rather than instructions
  - Enables a system to autonomously learn and improve using **neural networks** and **deep learning**, without being explicitly programmed, by feeding it large amounts of data
  - A subset of **Artificial Intelligence**

Overview of ML

- A typical ML process follows the following steps – **ML pipeline**:
  - Problem framing
    - Define ML problems from business projects
      - What is the success measurement of the problem?
      - What are the traditional ways of solving the problem?
      - Is there enough quality data to solve the problem using ML?
      - What is the best ML approach to solve the problem?
  - Data collection
    - Collect data from various sources, which may involve data labeling
    - Make sure that the datasets represent the real ML problem and are in the right format for ML model training
  - Data evaluation
    - Examine the data using statistical tools
    - to sample, balance, and scale datasets and handle missing values and outliers in the datasets

Overview of ML

- Feature engineering
  - Select and create model features and targets
  - Extract, construct, and transform features
- Model selection
  - choose the appropriate machine learning algorithm(s) based on the problem type (e.g., classification, regression), data characteristics, and performance requirements

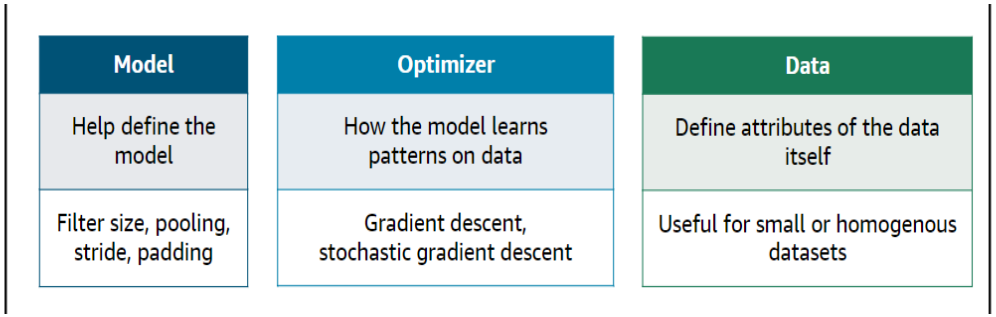
Overview of ML

- Model training
  - Train the model with the training dataset
  - Minimize the gap between the forecasted target value and the actual target value - the **loss function** (also called the **cost function**)
    - Mean Square Error (MSE) - A popular loss function for regression

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Overview of ML

- Model verification
  - Verify the model with the verification dataset
  - Improving the performance of models
    - Model parameter tuning



Model	Optimizer	Data
Help define the model	How the model learns patterns on data	Define attributes of the data itself
Filter size, pooling, stride, padding	Gradient descent, stochastic gradient descent	Useful for small or homogenous datasets

    - Combining different models with diverse strengths -ensemble

Overview of ML

- Model testing
  - Test the model with the testing dataset
- Model deployment
  - Deploy the ML model to production
  - **Inference** - applying the trained model to unseen data

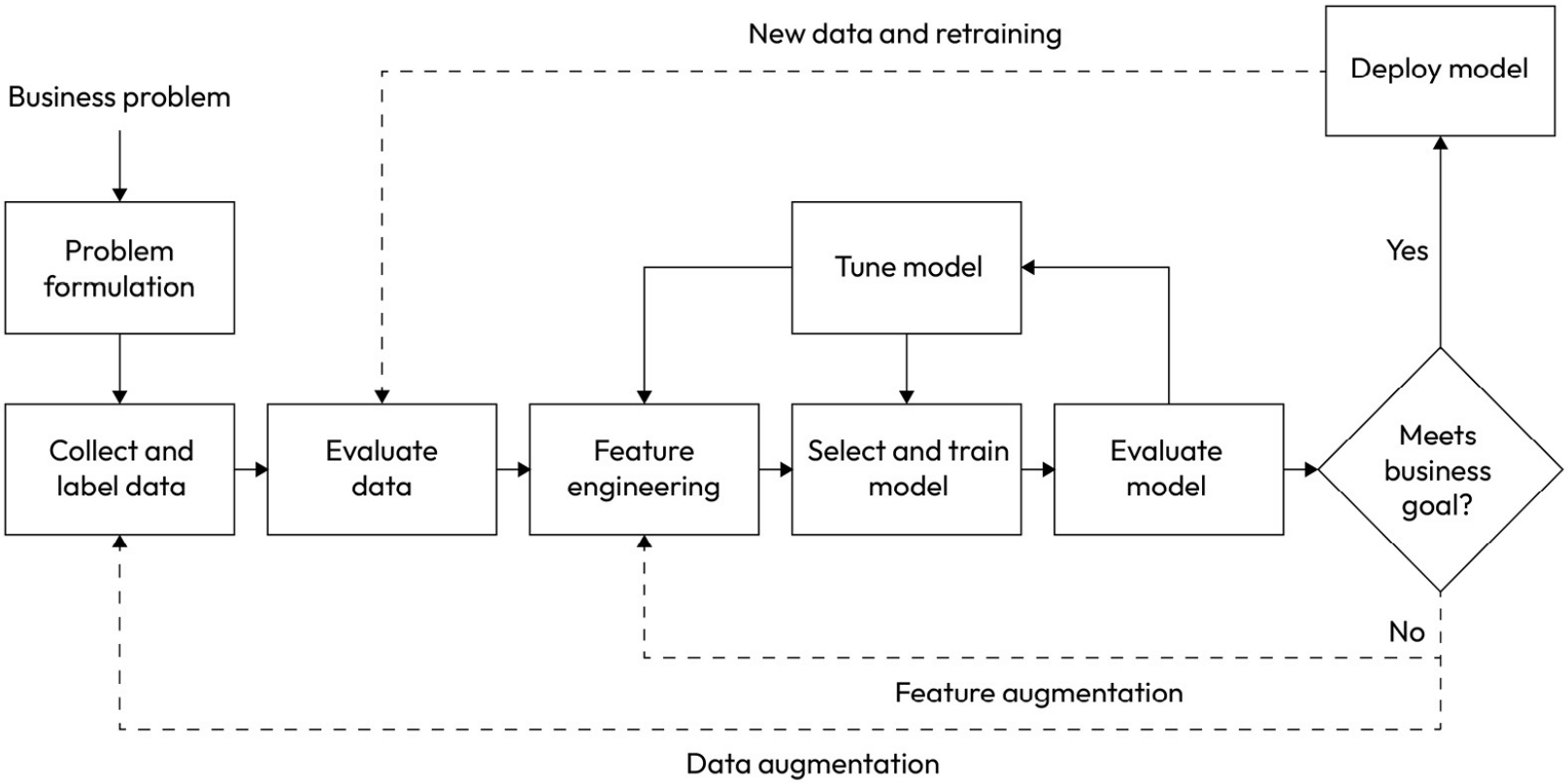
Classification of ML

- Three major classes based on the type of learning
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Classification of ML

- Supervised learning
  - A machine predicts the class of unknown objects based on prior class-related information of similar objects
  - When you have a labeled dataset that you can use to train your model
  - Broadly categorized:
    - Classification
      - Credit card fraud detection
    - Regression
      - Predicting stock value based on historical trend
  - Example Algorithms: Naïve Bayes, Decision tree, and k-Nearest Neighbors, etc.

Classification of ML



Classification of ML

- Confusion Matrix:
  - An NxN table that summarizes the number of correct and incorrect predictions that a classification model made
  - **True positive (TP)** - an outcome where the model correctly predicts the positive class
  - **False positive (FP)** - is an outcome where the model incorrectly predicts the positive class
  - **False negative (FN)** - an outcome where the model incorrectly predicts the negative class
  - **True negative (TN)** - is an outcome where the model correctly predicts the negative class

Classification of ML

- Evaluation
  - Classification
    - Model **accuracy**:

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

    - **Precision** - the proportion of positive identifications that are correct

$$= \frac{TP}{TP + FP}$$

    - **Recall** (sensitivity) - the proportion of actual positives that were identified correctly

$$= \frac{TP}{TP + FN}$$

Classification of ML

- Unsupervised learning
  - A machine finds patterns in unknown objects by grouping similar objects together
  - Training a model on unlabeled data to identify patterns and relationships
    - Pattern discovery or knowledge discovery
  - Categorized as:
    - Clustering
    - Association analysis
  - Example: K-means, DBSCAN

Classification of ML

- Reinforcement learning
  - Interacts with its environment and take actions to maximize rewards
  - A machine learns to act on its own to achieve the given goals
  - Agents learn from the environment by trial and error
  - Example Application
    - self-driving cars
  - Algorithms
    - Q-learning, Sarsa

Issues in ML

- Issues in Machine Learning
  - Data quality
  - Privacy
  - Bias and ethical issues

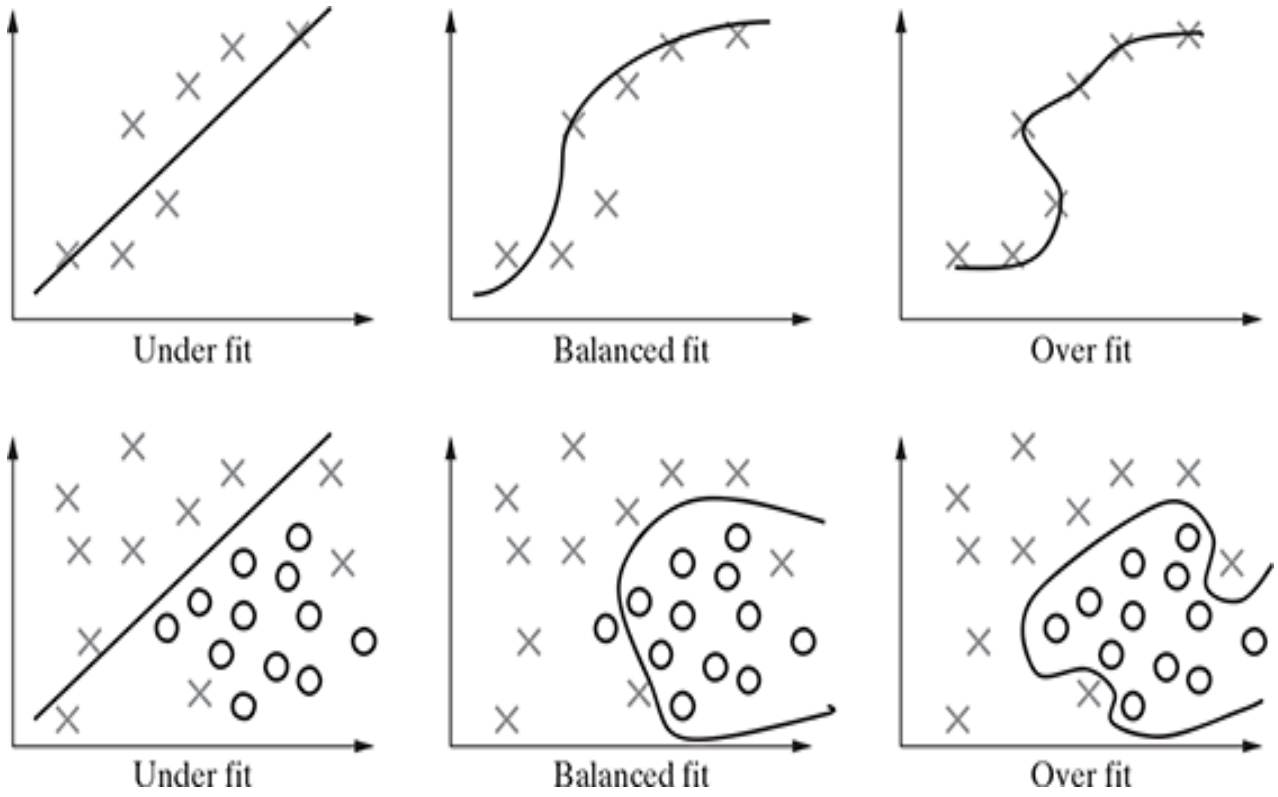
Issues in ML

- Model underfitting
  - If the target function (model) fails to capture the essential characteristics of the underlying data well
  - Results in both poor performance with training data as well as poor generalization to test data
  - Solution
    - Using more training data
    - Reducing features by effective feature selection

Issues in ML

- Model overfitting
  - Model has been designed in such a way that it emulates the training data too closely
  - Results in wrong classification in the test data set
  - Solution
    - Using validation

Issues in ML


Under fit
Balanced fit
Over fit
Under fit
Balanced fit
Over fit

References

- The Self-Taught Cloud Computing Engineer: A comprehensive professional study guide to AWS, Azure, and GCP. Logan Song. Packt Publishing, 2023
- Machine Learning. S. Chandramouli, S. Dutt, A. K. Das, Pearson Education India, 2018