JOM RESEARCH

Basic concepts in machine learning

Tengku Muhammad Hanis Bin Tengku Mokhtar, PhD

October 22, 2024

Contents

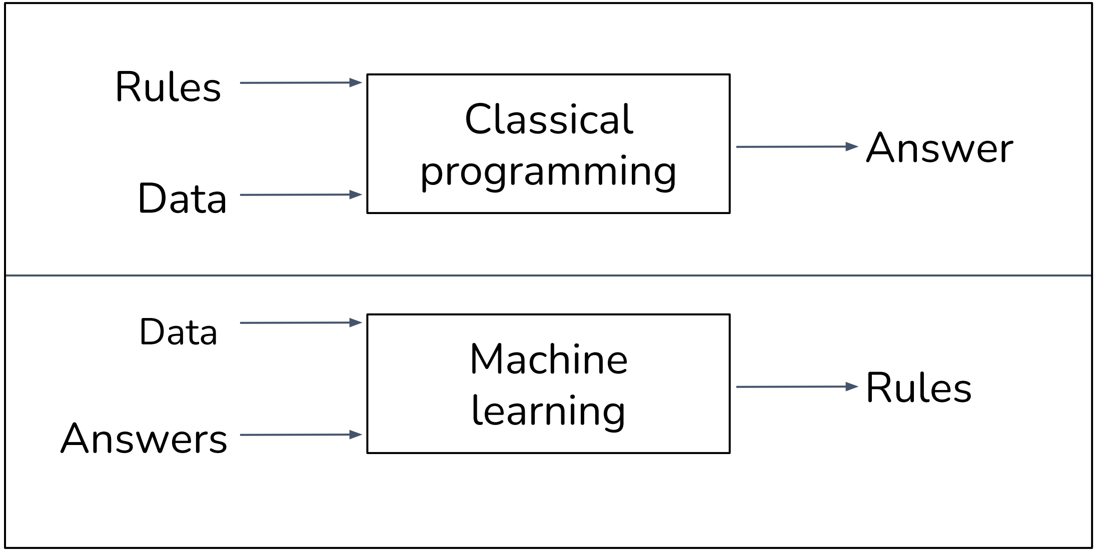
- Material
- Machine learning (ML)
- Regression
- Classification
- Dimension reduction
- Clustering
- Deep learning
- Basic concepts in ML
- ML workflows (MLOps)
- Suggested readings

Material

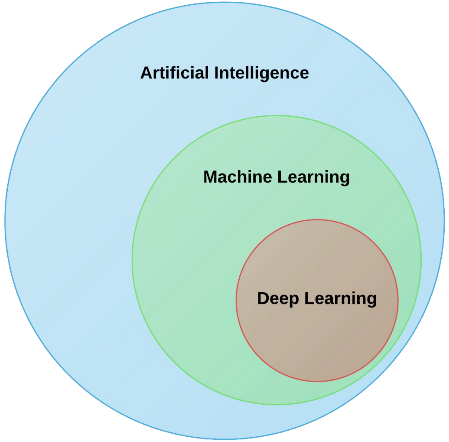
- Available at: https://drive.google.com/drive/folders/1NorSOjOzrSiL2otd_nTAnKoH9CzAi-YV
- All materials are in GitHub repositories:
  - Material for Posit Cloud
    - Link: https://github.com/tengku-hanis/MLiM
    - No need to download (or download it if you want to save the material)
  - Material for kaggle
    - Link: https://github.com/tengku-hanis/MLiM_kaggle
    - Download this, we going to upload the material to kaggle

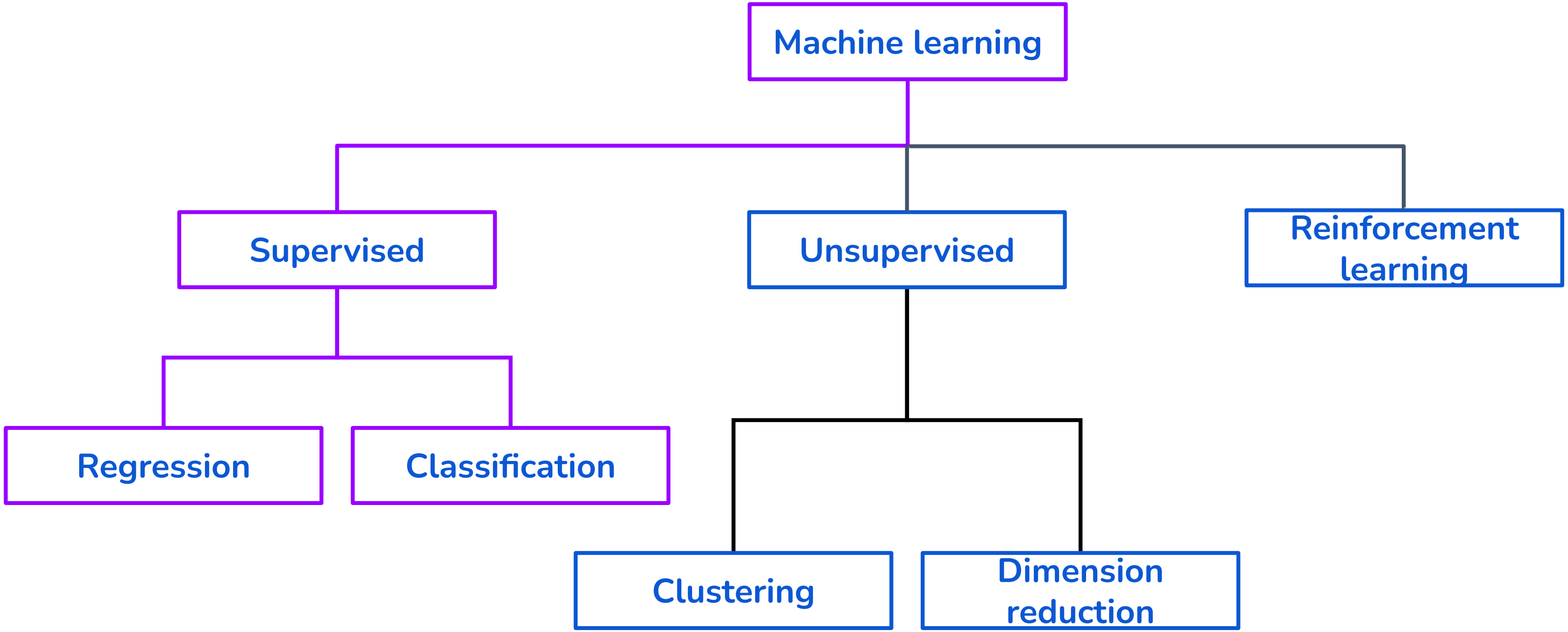
Machine learning (ML)

- A branch of artificial intelligence (AI)
- ML algorithms learn from data to make predictions or decisions without being explicitly programmed



- Deep learning is considered as a subfield of ML
- Nowadays, due to advance in DL (such as large language model (ChatGPT, BERT, etc), generative AI, etc), this subfield can be considered as its own field





Machine learning

Supervised

Unsupervised

Reinforcement learning

Regression

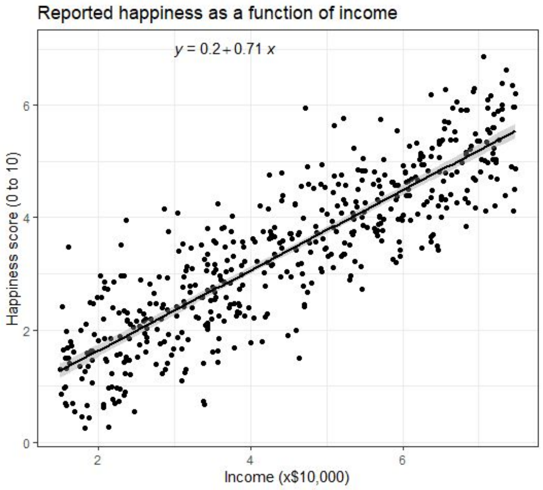
Classification

Clustering

Dimension reduction

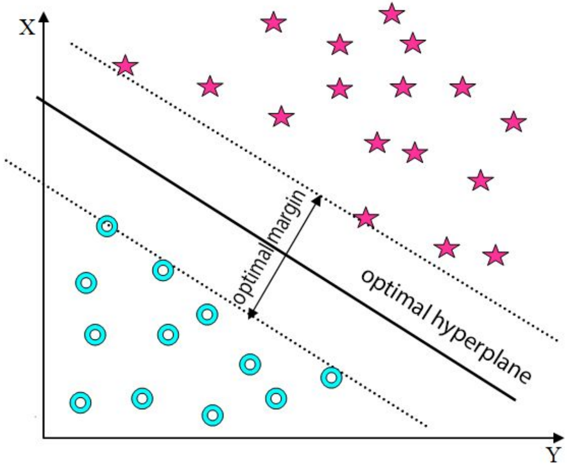
Regression

- Regression algorithms aim to predict the numerical/continuous outcome
- Example of regression problem/ data:
  - House price prediction
  - Medical cost prediction
  - Patient length of stay prediction
- Regression can be:
  - Normal regression
  - Censored/truncated regression



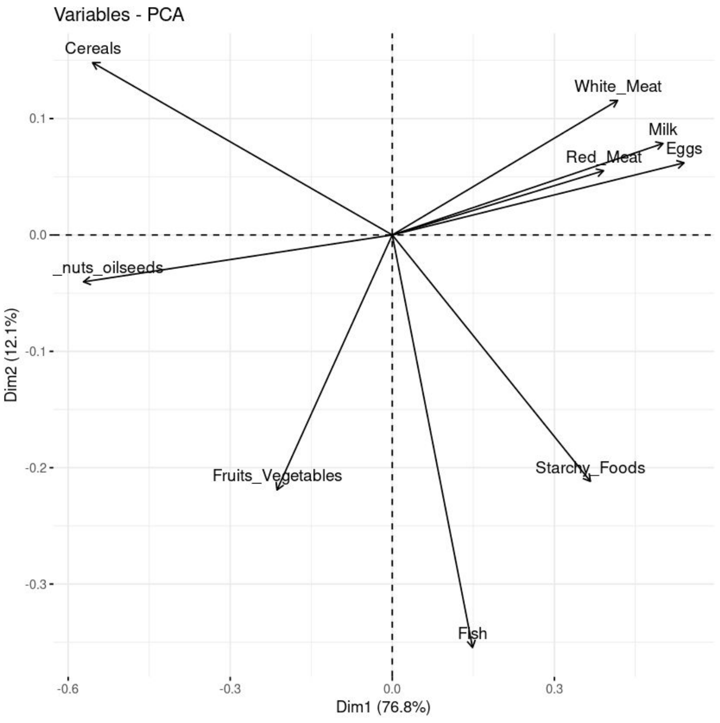
Classification

- Classification algorithms aim to predict the categorical outcome
- Example of classification problem/ data:
  - Breast cancer prediction (yes/no, normal/malignant, normal/ benign/ malignant)
  - Email spam detection (yes/no)
  - Death due to a disease (survived/death)
- Classification can be:
  - Binary - two groups
  - Multiclass - more than two groups



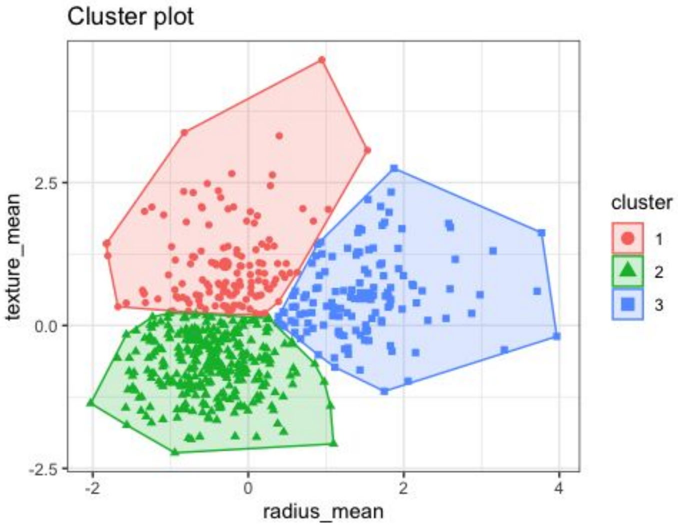
Dimension reduction

- Dimension reduction algorithms aim to reduce the number of features in a dataset
- Examples of algorithm:
  - Principal Component Analysis (PCA)
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Uniform Manifold Approximation and Projection (UMAP)
  - Linear Discriminant Analysis (LDA)
  - Etc

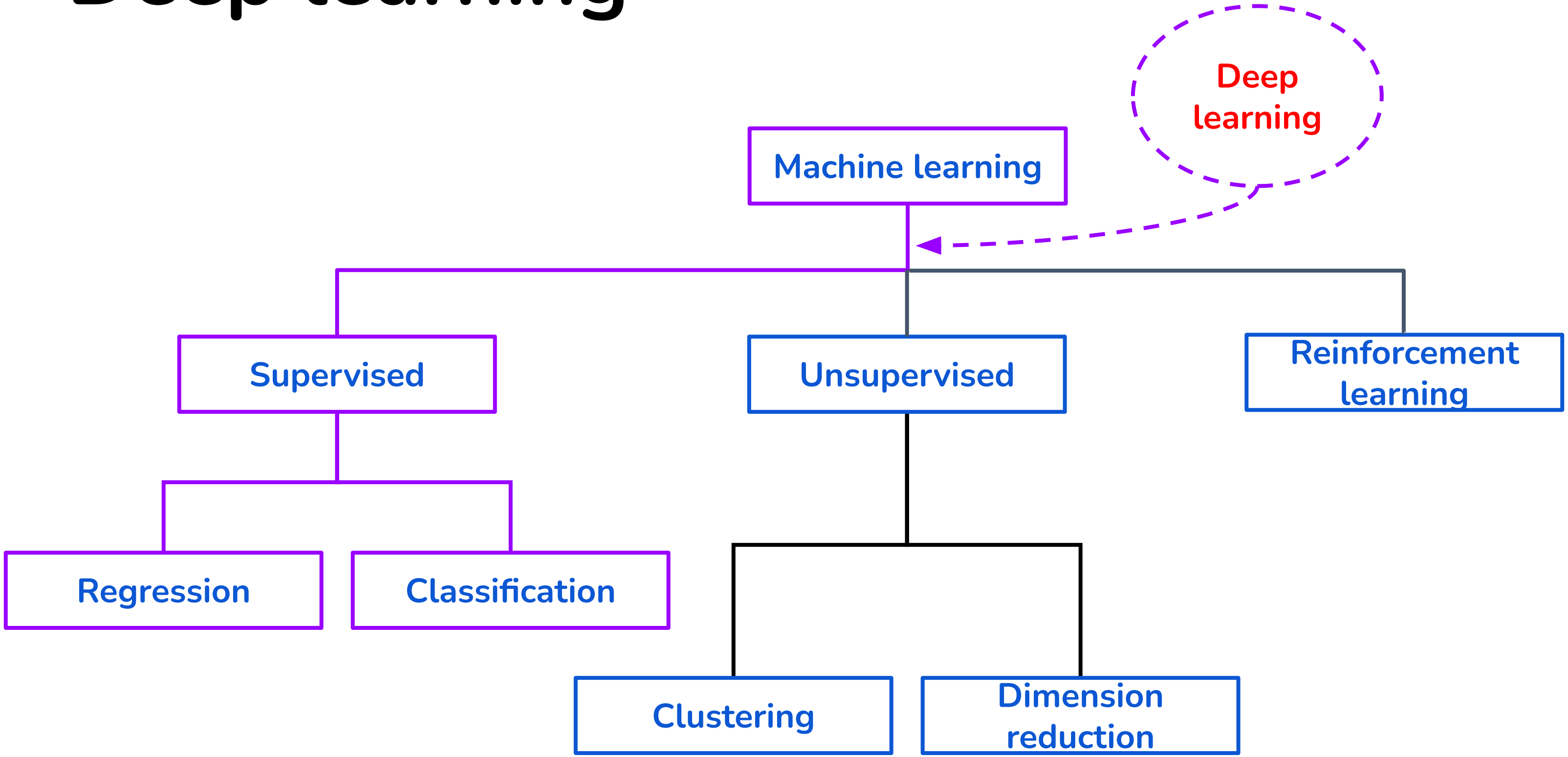


Clustering

- Clustering algorithms aim to group similar data points into a few groups based on their characteristics or features
- Examples of algorithm:
    - k-mean clustering
    - Hierarchical clustering
    - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
    - Etc



Deep learning



Basic concepts in ML

- Training and testing datasets
- Data leakage
- Feature engineering
- Resampling
- Cross validation
- Performance metrics
- Loss function
- Overfit vs. underfit
- Curse of dimensionality
- Parameter vs hyperparameters
- Hyperparameter tuning

Training and testing datasets

- In developing the ML model, the dataset is split into training and testing
- Training dataset - dataset used for training the ML model
    - Main purpose in training phase is to find the best hyperparameters for the ML model (hyperparameter tuning)
- Testing dataset - dataset used for validating the ML model

Data leakage

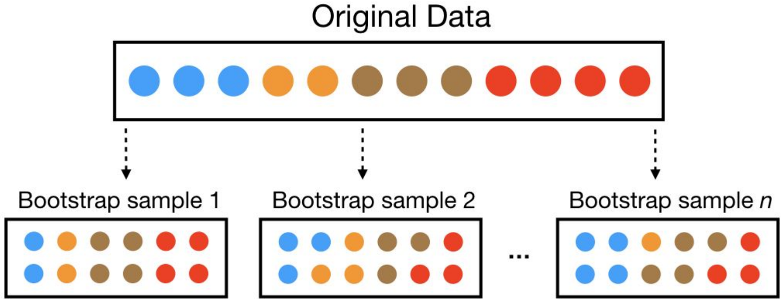
- Occurs when information from outside the training dataset is inadvertently used to build the model
- Leads to overly optimistic performance estimates because the model "cheats" by accessing information it should not have during training
- Most common data leakage:
  - Train-test contamination
  - Feature leakage

Feature engineering

- The process of transforming raw data into meaningful features that improve the performance of a ML model
- Any forms of feature engineering should be done using training set only
- Example:
  - Create new meaningful feature - BMI instead of weight and height
  - Handling missing value - imputation
  - Normalisation or scaling

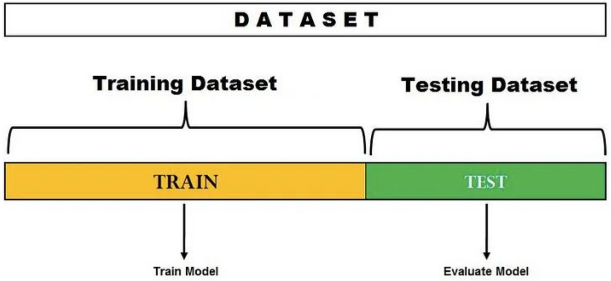
Resampling

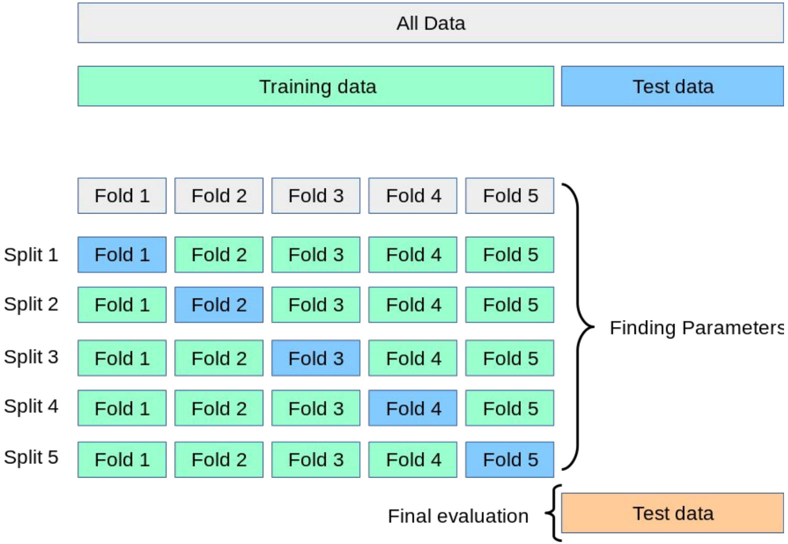
- Repeatedly and randomly drawing samples from the dataset to create a new dataset
- Resampling techniques are applied to the training datasets:
    - Bootstrap - repeatedly sampling with replacement from the available dataset to create multiple bootstrap samples
    - Cross-validation (CV)
    - Etc



Cross-validation (CV)

- One of the resample techniques
- Most common type:
  - Holdout CV
  - k-fold CV
  - Leave-one-out CV (LOO-CV)
  - Etc





Performance metrics

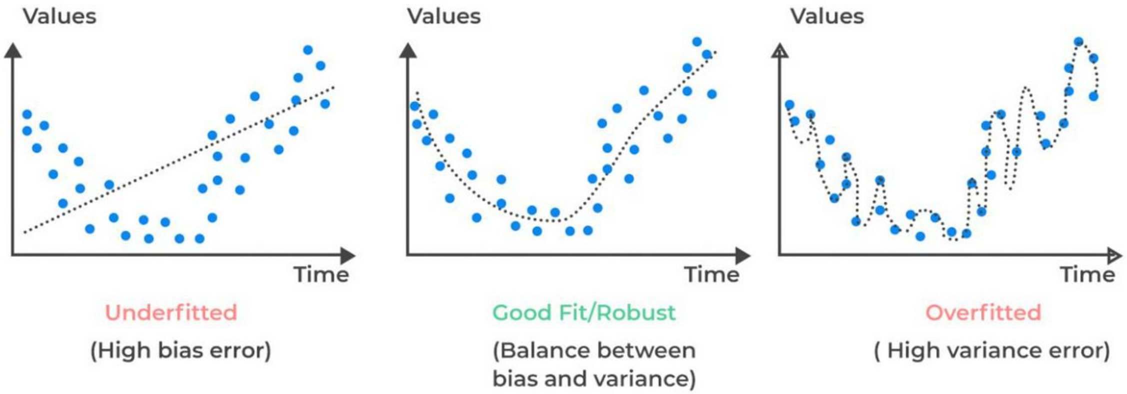
- How we measure the performance of our ML models
- It differs according to types of algorithm:
  - Regression: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc
  - Censored regression: Concordance index (C-index), Brier Score, etc
  - Classification: accuracy, precision, Receiver Operating Characteristic Area Under Curve (ROC-AUC), confusion matrix, etc
  - Clustering: silhouette score, Davies-Bouldin Index, etc

Loss function

- It reflects how well the ML model performs and it signals how the model’s hyperparameters supposed to be tuned
- Commonly used loss function:
  - Regression: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc
  - Classification: binary cross-entropy, log loss, hinge loss, etc

Overfit vs. underfit

- Overfit: when a model learns the training data too well and fail to generalise to a new data
- Underfit: when a model fails to learn a training data, thus, fails to performs on a new data as well



Curse of dimensionality

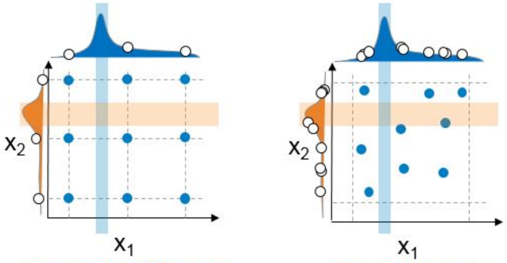
- As the number of features (dimensions) increases, the volume of the space grows exponentially, making data points sparse
- This sparsity
  - Makes it difficult for algorithms to find meaningful patterns
  - Increase computational costs
  - Reducing overall performance - models are more likely to overfit because they capture noise instead of the underlying pattern in the data

Parameters vs hyperparameters

Aspect	Parameters	Hyperparameters
Definition	Values learned automatically by the model during training	Values set manually before training to control the learning process
Examples	Weights in neural networks, coefficients in linear regression	Learning rate, number of layers in a neural network, regularization value
Learning Process	Determined based on the data and optimized during model training	Cannot be learned by the model; must be set by the user or through tuning methods like grid search
Purpose	Capture patterns from the training data to make predictions	Guide how the model should be trained and influence its overall performance
Tuning	Typically adjusted automatically by the learning algorithm	Manually tuned or adjusted using techniques like cross-validation or random search

Hyperparameter tuning

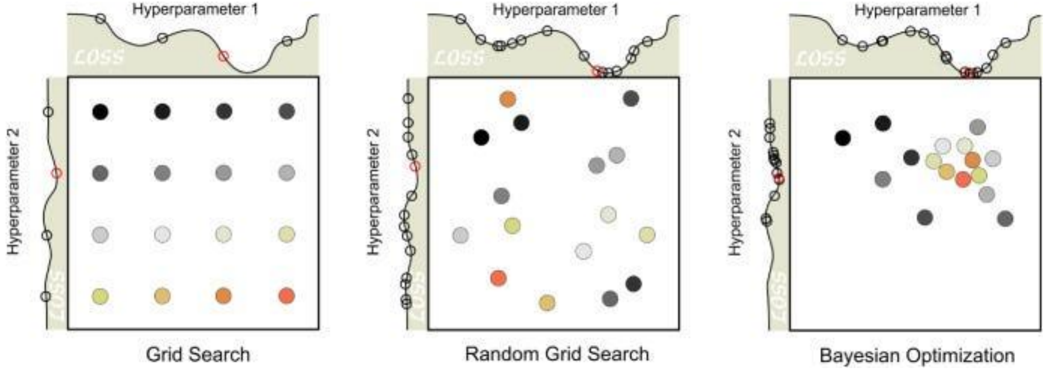
- Involves the process of selecting the best hyperparameters for a ML model
- How to come up with a set of hyperparameter combination?
    - Grid search:
        - Regular grid search - explore each set of predefined combination of hyperparameters
        - Non-regular grid search
            - Random grid search - explore random set of combination of hyperparameters
            - Etc



Standard Grid Search

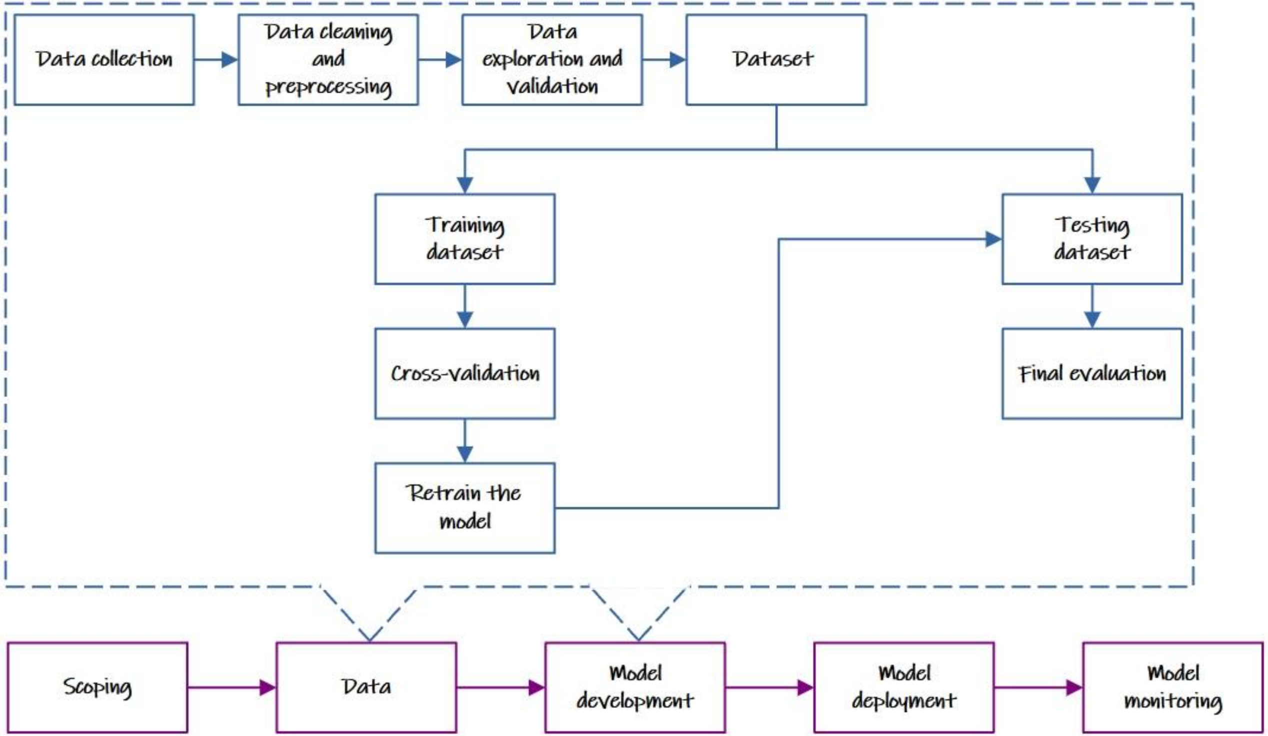
Random Search

- Iterative search
  - Bayesian optimization - explore the next best combination based on the performance of the previous combination
  - Simulated annealing
  - Etc



JOM RESEARCH

ML workflows (MLOps)



Suggested readings/references

- Burger, S. V. (2018). Introduction to machine learning with R: Rigorous mathematical analysis (First edition). O’Reilly Media.
- Kuhn, M., & Silge, J. (2022). Tidy Modeling with R: A Framework for Modeling in the Tidyverse. O’Reilly Media.

Any question?

tengkuhanismokhtar@gmail.com
jom.research.malaysia@gmail.com