Machine Learning Concepts

Essential Terminology

- **Model**: A mathematical representation of a real-world process, trained on data.
- **Feature** (X): An independent variable or input used for prediction.
- **Label** (y): The target or output we want to predict.
- **Training**: The process of finding the optimal parameters of a model using data.
- **Inference**: Using a trained model to make predictions on new data.
- **Loss Function**: A function that measures how far the model's prediction is from the actual label.
- **Optimization**: The process of minimizing the loss function (e.g., Gradient Descent).

Types of Learning

- **Supervised Learning**: Learning from labeled data (Input-Output pairs).
  - **Regression**: Predicting continuous values (e.g., house prices).
  - **Classification**: Predicting categories (e.g., spam vs. not spam).
- **Unsupervised Learning**: Finding patterns in unlabeled data.
  - **Clustering**: Grouping similar data points (e.g., customer segmentation).
  - **Dimensionality Reduction**: Reducing features while preserving info (e.g., PCA).
- **Reinforcement Learning**: Learning through trial and error to maximize a reward.

Model Evaluation

- **Train / Test Split**: Dividing data to evaluate performance on unseen data.
- **Overfitting**: Model performs perfectly on training data but poorly on test data (memorization).
- **Underfitting**: Model is too simple to capture patterns in the data.
- **Metrics (Classification)**:
  - `Accuracy`: Correct predictions / Total.
  - `Precision`: Correct positive predictions / Total predicted positive.
  - `Recall`: Correct positive predictions / Total actual positive.
  - `F1 Score`: Harmonic mean of Precision and Recall.
- **Metrics (Regression)**:
  - `MSE` (Mean Squared Error), `MAE` (Mean Absolute Error), `R-squared`.

Common Algorithms

- **Linear Regression**: Simple model for predicting continuous values.
- **Logistic Regression**: Baseline for binary classification.
- **Decision Trees**: Tree-like structure for making decisions.
- **Random Forest**: Ensemble of many decision trees (bagging).
- **XGBoost / LightGBM**: Powerful gradient boosting algorithms.
- **SVM (Support Vector Machine)**: Finding the hyperplane that best separates classes.
- **Neural Networks (Deep Learning)**: Bio-inspired models with multiple layers.

Feature Engineering and Selection

- **Normalization / Scaling**: Adjusting feature ranges (e.g., 0 to 1).
- **Encoding**: Converting categorical data to numbers (One-hot, Label encoding).
- **Imputation**: Handling missing values (replacing with mean/median).
- **Regularization**: Adding a penalty to the loss function to prevent overfitting (L1 Lasso, L2 Ridge).

Modern AI (Deep Learning & LLMs)

- **CNN (Convolutional Neural Network)**: Specialized for images.
- **RNN / LSTM**: Specialized for sequential data (text, time-series).
- **Transformer**: The architecture behind modern LLMs (using Self-Attention).
- **Transfer Learning**: Taking a pre-trained model and fine-tuning it for a specific task.

Best Practices

Data Quality

"Garbage in, garbage out." The quality and quantity of data are often more important than the complexity of the algorithm.

Cross-Validation

Use k-fold cross-validation to get a more robust estimate of model performance.

Keep It Simple

Always start with a simple baseline model before moving to complex deep learning architectures.