TOPIC THREE

LAUNCHING INTO MACHINE LEARNING

SUMMARY NOTES (ET0732)

Machine Learning Phases and Data Quality

Phases in Machine Learning:

- **Training Phase**: Model development.
- **Inference Phase**: Model deployment and predictions.

ML Project Steps:

1. Define the use case and success criteria.
2. Deliver the ML model through manual or automated steps.

Key Steps Related to Data:

1. **Data Extraction**: Retrieve data from diverse sources (structured/unstructured).
2. **Data Analysis**: Use Exploratory Data Analysis (EDA) to identify trends and anomalies.
3. **Data Preparation**:
    - **Data Transformation**: Modify format/structure.
    - **Data Cleansing**: Remove duplicates and irrelevant records.
    - **Data Type Correction**: Fix errors and convert types.

Data Quality Attributes:

- **Accuracy**: Data should reflect real-world events.
- **Timeliness**: Measure the time from capture to availability.
- **Completeness**: Ensure all intended data is present.

Improving Data Quality:

- Address missing values and unwanted characters.
- Properly format date/time features.
- Use One-hot Encoding for categorical features.

Iterative Process: Data exploration and cleaning are continuous, improving data quality over time.

Importance of Data Quality: High-quality data enhances the predictive power of ML models.

Machine Learning in Practice

Module Overview: Focus on real-world problem-solving with data and ML algorithms.

Supervised vs. Unsupervised Learning:

- **Supervised Learning**: Involves labeled data (e.g., predicting tips).
- **Unsupervised Learning**: Works with unlabeled data (e.g., clustering).

Types of Supervised Learning Problems:

1. **Regression**: Predicts continuous values (e.g., tips based on bill).
2. **Classification**: Predicts discrete classes (e.g., gender based on features).

Experimentation: Involves testing different models and techniques.

Data Types:

- **Structured Data**: Organized in rows and columns.
- **Unstructured Data**: Includes images, audio, etc.

Model Training and Optimization

Loss Functions: Measure model performance against actual values.

Gradient Descent: Optimizes model parameters by minimizing loss functions.

Learning Rate: A hyperparameter that affects the size of the steps taken in gradient descent.

Generalization vs. Overfitting: Generalization measures performance on unseen data, while overfitting shows a model's poor performance on validation datasets.

Data Splitting:

- **Training**: For model training.
- **Validation**: For tuning hyperparameters.
- **Test**: For final evaluation.

Cross-Validation: Maximizes data usage for robust model evaluation.

Key Points on Model Performance

Confusion Matrix: Assesses classification performance through TP, FP, TN, FN metrics.

Metrics:

- **Precision**: Accuracy of positive predictions.
- **Recall**: Ability to identify actual positives.

Evaluation: Use performance metrics to assess model accuracy and generalization.

Final Considerations: Regularly monitor predictions for bias and optimize model performance based on evaluation metrics.