FML Module 1 Part A

1. Imagine you're working on a machine learning project, and the dataset contains a significant amount of noisy data. How would you identify and handle noisy data to ensure the robustness and reliability of the model? Discuss various techniques for noise detection, such as outlier detection, and strategies for data cleaning and preprocessing.

Answer:

- **Identifying Noisy Data:**
  - Outlier Detection: Statistical methods like Z-score or IQR can identify outliers.
  - Visualization: Plotting data helps identify anomalies or inconsistencies.
  - Domain Knowledge: Understanding the data domain helps spot unrealistic values.
- **Handling Noisy Data:**
  - Removing Outliers: Exclude extreme values that don't align with the rest of the data.
  - Imputation: Replace missing values with a measure like mean, median, or mode.
  - Data Transformation: Apply log transformation or normalization to stabilize variance.
  - Ensemble Methods: Use models like Random Forests that are robust to noise.

These techniques ensure the robustness and reliability of the model by reducing the impact of noisy data.

2. Consider a scenario where you have access to limited labelled data for training a machine learning model, but acquiring additional labelled data is costly or time-consuming. How would you improve model performance?

Answer:

- **Techniques to Improve Model Performance with Limited Data:**
  - Data Augmentation: Generate synthetic data by applying transformations like rotation or flipping.
  - Transfer Learning: Pretrain a model on a large dataset and fine-tune it on the limited dataset.
  - Semi-Supervised Learning: Utilize unlabeled data in conjunction with limited labeled data.
  - Active Learning: Selectively query labels for instances that the model is uncertain about.
  - Ensemble Learning: Combine predictions from multiple models trained on different subsets of data.

These approaches help enhance model performance even with limited labeled data, mitigating the challenges of data scarcity.

---

3. Suppose you're building a machine learning model for a critical application, such as healthcare or finance, where model interpretability and explainability are essential. How would you ensure that the model's predictions are transparent and understandable to end-users or stakeholders?

Answer:

- **Ensuring Model Transparency and Explainability:**
  - Feature Importance: Use techniques like SHAP values or permutation importance to explain feature contributions.

- Model Interpretation: Employ interpretable models like decision trees or linear regression instead of black-box models.
- Local Interpretability: Provide explanations for individual predictions using methods like LIME or SHAP.
- Visualizations: Present model behavior through intuitive visualizations like partial dependence plots or decision trees.

These strategies ensure that stakeholders can trust and understand the model's predictions, crucial for critical applications like healthcare and finance.

---

4. If you were to design an experiment to determine the best predictive model for a dataset with multiple features and a continuous target variable, how would you evaluate and contrast the predictive capabilities of distinct algorithms, such as linear regression, decision trees, and support vector machines, in order to ascertain their respective effectiveness in modeling a dataset with multiple features and a continuous target variable?

Answer:

- **Experiment Design for Model Evaluation:**
  - Split Data: Divide dataset into training, validation, and test sets.
  - Train Models: Train linear regression, decision tree, and support vector machine models on the training set.
  - Evaluate Performance: Measure performance metrics like RMSE, MAE, or R-squared on the validation set for each model.
  - Select Best Model: Choose the model with the best performance on the validation set.
  - Validate on Test Set: Assess the selected model's performance on the test set to ensure generalization.

By systematically comparing different algorithms' performance metrics, we can determine the most effective model for the given dataset.

5. Once you've trained a predictive model on historical data, what steps would you take to implement the model in a production environment and continuously track its performance as time progresses?

Answer:

- **Steps for Model Deployment and Monitoring:**
  - Integration: Integrate the model into the production environment using APIs or containerization.
  - Performance Monitoring: Continuously monitor the model's performance using metrics like accuracy or F1-score.
  - Error Handling: Implement mechanisms to handle errors and exceptions gracefully.
  - Version Control: Maintain version control for the deployed model to track changes and updates.
  - Feedback Loop: Incorporate feedback from users and real-world data to retrain and improve the model periodically.

These steps ensure that the deployed model remains effective and reliable in real-world scenarios, adapting to evolving data and user needs.

---

6. Compare and contrast the advantages and disadvantages of scanning and emailing images versus utilizing an optical character reader (OCR) to send text files. Under what circumstances would one approach be more advantageous than the other?

Answer:

- **Scanning and Emailing Images:**
  - Advantages: Preserves visual layout and formatting, suitable for handwritten or complex documents.

    - Disadvantages: Large file sizes, not searchable or editable without OCR, limited text extraction accuracy.
- **Utilizing OCR to Send Text Files:**
    - Advantages: Creates searchable and editable text files, reduces file size, enables text analysis and processing.
    - Disadvantages: May lose some formatting or layout information, accuracy depends on OCR quality.

One approach may be more advantageous than the other depending on the requirement for preserving visual layout and the need for searchable and editable text.

---

7. Let us say we are building an OCR and for each character, we store the bitmap of that character as a template that we match with the read character pixel by pixel. Explain when such a system would fail. Why are barcode readers still used?

Answer:

- **Failure of Pixel-Matching OCR System:**
    - Limited Robustness: Susceptible to variations in font, size, and style, leading to incorrect matches.
    - Sensitivity to Noise: Noise or distortions in the image can cause mismatches between templates and characters.
    - Computational Complexity: Pixel-by-pixel comparison is computationally intensive and inefficient.
- **Advantages of Barcode Readers:**
    - Standardized Representation: Barcodes have a structured, standardized format, making them easier to decode accurately.
    - Fast and Reliable: Barcode readers offer fast and reliable scanning, even in challenging environments.

- Error Correction: Barcodes often incorporate error correction techniques to improve reliability in case of damage or distortion.

Barcode readers are preferred for their efficiency, reliability, and error correction capabilities, especially in scenarios where accuracy and speed are critical.

---

8. Assume we are given the task of building a system to distinguish junk email. What is in a junk email that lets us know that it is junk? How can the computer detect junk through a syntactic analysis? What would we like the computer to do if it detects a junk email—delete it automatically, move it to a different file, or just highlight it on the screen?

Answer:

- **Indicators of Junk Email:**
  - Spammy Content: Promotional offers, suspicious links, or irrelevant content.
  - Sender Reputation: Unknown sender, deceptive sender name, or unusual email address.
  - Phishing Attempts: Requests for personal information, urgent action, or financial transactions.
- **Syntactic Analysis for Junk Detection:**
  - Keyword Analysis: Check for common spam keywords or phrases in the email body or subject.
  - Sender Analysis: Verify sender authenticity and cross-check against known spam sender lists.
  - Header Analysis: Inspect email headers for anomalies or suspicious patterns.
- Action for Detected Junk Email
- Options include automatic deletion, moving to a designated junk folder, or flagging for user review.

- Preference may vary based on user preferences and organizational policies regarding email management.

These approaches help computers effectively identify and handle junk email, safeguarding users from potential security threats and cluttering their inbox.

---

9. If a face image is a 100 × 100 image, written in row-major, this is a 10,000-dimensional vector. If we shift the image one pixel to the right, this will be a very different vector in the 10,000-dimensional space. How can we build face recognizers robust to such distortions?

Answer:

- **Techniques for Robust Face Recognition:**
  - Feature Extraction: Use techniques like Principal Component Analysis (PCA) or Convolutional Neural Networks (CNNs) to extract robust features invariant to translations.
  - Data Augmentation: Apply transformations like rotation, scaling, or flipping to increase the diversity of training data and improve model generalization.
  - Spatial Transformer Networks: Incorporate spatial transformer layers in CNNs to learn spatial transformations and enhance model robustness to geometric distortions.

By incorporating techniques that capture intrinsic facial features while being invariant to translations, we can build face recognizers robust to such distortions.

---

10. In basket analysis, we want to find the dependence between two items X and Y. Given a database of customer transactions, how can we find these dependencies? How would we generalize this to more than two items?

Answer:

- **Finding Dependencies between Two Items (X and Y):**

    - Association Rule Mining: Apply algorithms like Apriori or FP-Growth to discover frequent itemsets containing both X and Y.
    - Measure Association Strength: Calculate metrics like support, confidence, and lift to quantify the strength of association between X and Y.
- **Generalizing to More than Two Items:**
    - Extend Association Rule Mining: Discover frequent itemsets containing multiple items (X, Y, Z, etc.) using the same algorithms.
    - Evaluate Multi-Item Associations: Measure association strength for combinations of multiple items to identify complex dependencies.

By analyzing customer transactions and identifying frequent itemsets or patterns, we can uncover dependencies between items and understand purchasing behavior for targeted marketing or recommendation systems.