**Name: Arshia Jamil**

**Manual 2 Task 1**

**Detailed Notes on Classification (Chapter 3)**

**Date: June 21, 2025**

**1. MNIST Dataset**

**Explanation:**

The MNIST dataset includes 70,000 handwritten digit images (0–9), each 28x28 pixels, resulting in 784 features (pixel intensities from 0 = white to 255 = black). Each image is labeled with the digit it represents. Known as the “hello world” of machine learning, it’s a standard for testing classification algorithms due to its manageable size and clear structure.

**Purpose:**

MNIST provides a standardized dataset to evaluate and compare classification models. Its moderate complexity makes it ideal for beginners to experiment with algorithms, visualize results, and understand challenges like distinguishing similar digits (e.g., 3 vs. 5).

**Usage in Machine Learning:**

- **Testing Algorithms**: Researchers benchmark new classification algorithms, as MNIST’s size and complexity reveal model strengths and weaknesses.

- **Learning and Prototyping**: Beginners use it to learn model training, evaluation, and tuning, as it’s available in libraries like Scikit-Learn.

- **Visualization**: The image-based nature allows visual inspection of predictions, aiding in understanding model behavior.

**Short Notes:**

The MNIST dataset contains 70,000 handwritten digit images (28x28 pixels, 784 features) with labels (0–9). Split into 60,000 training and 10,000 test images, it’s shuffled for consistent cross-validation. Used as a benchmark to test and compare classification models, it’s ideal for learning due to its simplicity and availability.

**2. Binary Classification**

**Explanation:**

Binary classification predicts whether an instance belongs to one of two classes (e.g., 5 vs. not-5 in MNIST). A model learns to distinguish between a positive class (e.g., 5) and a negative class (e.g., not-5) using algorithms like Stochastic Gradient Descent (SGD) Classifier.

**Purpose:**

Binary classification solves problems with two outcomes, like spam detection or disease diagnosis. In MNIST, a “5-detector” simplifies the problem to focus on one digit, helping understand basic classification mechanics.

**Usage in Machine Learning:**

**- Real-World Applications**: Used in spam filters (spam vs. not spam), medical diagnosis (disease vs. no disease), and fraud detection (fraud vs. legitimate).

- **Model Training:** Algorithms like SGD Classifier process one instance at a time, making them efficient for large datasets.

- **Foundation for Complex Tasks:** Binary classification is a building block for multiclass or multilabel classification.

**Short Notes:**

Binary classification predicts one of two classes (e.g., 5 vs. not-5). In MNIST, a “5-detector” uses an SGD Classifier to label images as True (5) or False (not-5). It’s used for tasks like spam detection, leveraging efficient algorithms like SGD for large datasets.

**3. Performance Measures**

**Explanation:**

Performance measures evaluate classifier effectiveness. Key metrics include:

- **Accuracy**: Ratio of correct predictions, but misleading for imbalanced datasets (e.g., 90% accuracy if only 10% of MNIST images are 5s).

- **Confusion Matrix**: Shows true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- **Precision**: TP / (TP + FP), accuracy of positive predictions.

- **Recall**: TP / (TP + FN), proportion of positives detected.

- **F1 Score**: Harmonic mean of precision and recall, balancing both.

**Purpose:**

These metrics assess model performance and guide improvements. In MNIST, high accuracy may hide poor performance on 5s if most images are non-5s. The confusion matrix reveals specific errors, while precision, recall, and F1 score provide nuanced insights.

**Usage in Machine Learning:**

- **Model Evaluation**: Metrics like precision (e.g., 72.9% for 5-detector) and recall (75.6%) evaluate trade-offs in applications like medical diagnosis (high recall) or spam filtering (high precision).

- **Cross-Validation**: K-fold cross-validation (e.g., 3 folds) ensures reliable performance estimates.

- **Decision Making**: Metrics guide model selection and tuning, ensuring alignment with project goals.

**Short Notes:**

Performance measures evaluate classifiers: accuracy (correct predictions), confusion matrix (TP, TN, FP, FN), precision (accuracy of positives), recall (positives detected), and F1 score (balance of both). Cross-validation ensures reliable metrics. In MNIST, they reveal model strengths and weaknesses, guiding improvements.

**4. Precision/Recall Trade-off**

**Explanation:**

Classifiers assign a score to each instance, predicting the positive class if the score exceeds a threshold. Raising the threshold increases precision (fewer false positives) but lowers recall (more false negatives). In MNIST, a 90% precision threshold for the 5-detector yields 43.7% recall.

**Purpose:**

The precision/recall trade-off tailors a classifier to project needs. High precision is critical for safe video detection (avoiding harmful content), while high recall is vital for shoplifting detection (catching most culprits).

**Usage in Machine Learning:**

- **Threshold Selection:** Plotting precision vs. recall helps choose a balance (e.g., 90% precision for safety-critical tasks).

- **Application-Specific Tuning**: High precision for spam filters, high recall for medical screening.

- **Model Optimization:** Adjust thresholds to optimize performance based on error costs.

**Short Notes:**

The precision/recall trade-off adjusts a classifier’s threshold to balance precision (correct positives) and recall (detected positives). In MNIST, a high threshold gives 90% precision but 43.7% recall. Used to tailor models for tasks like spam filtering (high precision) or medical diagnosis (high recall).

**5. ROC Curve**

**Explanation:**

The ROC curve plots true positive rate (recall) vs. false positive rate (FPR = FP / (FP + TN)) for various thresholds. A good classifier’s curve is close to the top-left corner, with an area under the curve (AUC) near 1 (0.5 for random). In MNIST, a Random Forest classifier (AUC 0.998) outperforms SGD (AUC 0.961).

**Purpose:**

The ROC curve and AUC compare classifier performance across thresholds, especially when false positives are critical. It’s less useful for imbalanced datasets, where the precision/recall curve is preferred.

**Usage in Machine Learning:**

**- Classifier Comparison**: AUC quantifies model quality (e.g., Random Forest vs. SGD).

- **Performance Analysis**: ROC curves visualize trade-offs between sensitivity and specificity.

- **Decision Support:** Used in fraud detection, where minimizing false positives is key.

**Short Notes:**

The ROC curve plots recall vs. false positive rate, with AUC measuring classifier quality (1 = perfect, 0.5 = random). In MNIST, Random Forest (AUC 0.998) beats SGD (AUC 0.961). Used to compare models and analyze performance in tasks like fraud detection.

**6. Multiclass Classification**

**Explanation:**

Multiclass classification distinguishes between multiple classes (e.g., digits 0–9 in MNIST). Strategies include:

**- One-vs-Rest (OvR)**: Train one binary classifier per class (e.g., 0 vs. not-0).

- **One-vs-One (OvO):** Train a classifier for each pair of classes (e.g., 0 vs. 1).

Scikit-Learn applies OvR or OvO automatically based on the algorithm (e.g., OvO for SVM, OvR for SGD).

**Purpose:**

Multiclass classification handles problems with more than two classes, like digit recognition or object classification. It extends binary classification to real-world scenarios with multiple categories.

**Usage in Machine Learning:**

- **Applications**: Used in image recognition (e.g., digits), text classification (e.g., sentiment categories), and medical diagnosis (e.g., disease types).

**- Algorithm Flexibility**: Algorithms like SGD and Random Forest handle multiclass natively, while binary algorithms (e.g., SVM) use OvR/OvO.

- **Performance Boost**: Scaling inputs (e.g., StandardScaler) improves accuracy (e.g., from 84% to 89% for SGD in MNIST).

**Short Notes:**

Multiclass classification predicts multiple classes (e.g., 0–9 in MNIST) using OvR (one classifier per class) or OvO (one per class pair). Scikit-Learn applies OvR/OvO automatically. Scaling inputs boosts accuracy. Used in digit recognition, text classification, and more.

**7. Error Analysis**

**Explanation:**

Error analysis examines a classifier’s mistakes using a normalized confusion matrix to compare error rates across classes. In MNIST, many digits are misclassified as 8s, and 3s/5s are often confused due to visual similarity. Visualizing errors helps identify patterns.

**Purpose:**

Error analysis identifies specific weaknesses in a model, guiding improvements like gathering more data or engineering new features. In MNIST, it reveals confusion between 3s and 5s, suggesting preprocessing to reduce errors.

**Usage in Machine Learning:**

- **Model Improvement**: Identifies misclassifications (e.g., 3s as 5s) to guide data collection or feature engineering (e.g., counting closed loops).

- **Preprocessing Strategies**: Suggests centering or rotating images to reduce errors.

- **Debugging**: Helps understand why a model fails, especially for linear models sensitive to small feature differences.

**Short Notes:**

Error analysis uses a normalized confusion matrix to identify misclassifications (e.g., 3s as 5s in MNIST). It guides improvements like adding data or preprocessing (e.g., centering images). Used to enhance model performance in tasks like digit recognition.

**8. Multilabel Classification**

**Explanation:**

Multilabel classification assigns multiple binary labels to each instance (e.g., for a digit: [large (7–9), odd]). In MNIST, a KNeighborsClassifier predicts whether a digit is large and odd, outputting [False, True] for 5.

**Purpose:**

Multilabel classification handles tasks where instances belong to multiple categories, like tagging images with multiple objects or labels. It extends binary classification to complex scenarios.

**Usage in Machine Learning:**

- **Applications**: Used in face recognition (multiple people in an image), text tagging (multiple topics), and medical diagnosis (multiple conditions).

- **Evaluation**: Metrics like average F1 score (e.g., 97.6% in MNIST) assess performance, with weighted averaging for imbalanced labels.

- **Algorithm Support**: Algorithms like KNeighborsClassifier support multilabel tasks.

**Short Notes:**

Multilabel classification assigns multiple binary labels (e.g., [large, odd] for digits). In MNIST, a KNeighborsClassifier predicts [False, True] for 5. Evaluated with average F1 score. Used in face recognition, text tagging, and more.

**9. Multioutput Classification**

**Explanation:**

Multioutput classification generalizes multilabel classification, where each label can have multiple values. In MNIST, a system to denoise images outputs 784 pixel intensities (0–255) per image, predicting a clean image from a noisy one.

**Purpose**:

Multioutput classification handles complex tasks where outputs are multidimensional, like image denoising or pixel-level predictions. It combines classification and regression elements.

**Usage in Machine Learning:**

- **Applications**: Used in image denoising, super-resolution, and semantic segmentation (labeling each pixel).

**- Model Training:** Algorithms like KNeighborsClassifier predict multiple outputs (e.g., pixel intensities in MNIST).

- **Evaluation**: Assessed by comparing predicted outputs to target values, often visually or with regression metrics.

**Short Notes:**

Multioutput classification predicts multiple labels with multiple values (e.g., 784 pixel intensities in MNIST for denoising). It generalizes multilabel classification, used in image denoising and segmentation. Algorithms like KNeighborsClassifier handle such tasks.

**Create comparison tables for SGD Classifier vs Random Forest performance**

|  |  |  |
| --- | --- | --- |
| Aspect | SGD Classifier | Random Forest |
| Type | Linear model with stochastic gradient descent | Ensemble of decision trees |
| Training Speed | Fast, suitable for large datasets | Slower, especially with many trees |
| Interpretability | Moderate (coefficients available) | Lower, but feature importance can be extracted |
| Handling Non-linearity | Poor (linear decision boundary) | Excellent (non-linear, complex boundaries) |
| Overfitting | Prone if not regularized | Less prone due to averaging multiple trees |
| Hyperparameters | Learning rate, penalty, alpha, epochs | Number of trees, max depth, min samples split |
| Memory Usage | Low | Higher due to many trees |
| Performance on Noisy data | Sensitive | Robust |
| Multiclass strategy | Naturally support OvR | Naturally support OvR |
| Use cases | Large sparse datasets, text classification | Tabular data with complex feature interactions |

**Create comparison tables for OvR vs OvO strategies for multiclass**

|  |  |  |
| --- | --- | --- |
| Aspect | One-vs-Rest (OvR) | One-vs-One (OvO) |
| How it works | Trains one classifier per class vs all others | Trains classifier for every pair of classes |
| Number of classifiers | Equal to number of classes (K) | K \* (K - 1) / 2 classifiers |
| Training time | Faster, fewer classifiers | Slower due to many classifiers |
| Prediction time | Faster | Slower, combines many predictions |
| Performance | Good for large number of classes | Often better with fewer classes or very similar classes |
| Handling Class Imbalance | Can be sensitive | Usually more balanced per pair |
| Complexity | Simpler to implement | More complex due to multiple classifiers |
| Use cases | Large-scale problems with many classes | Problems with small to medium class counts |