Evaluating Classifier Models: A Comparison of Machine Learning Algorithms

Caolán Maguire

Abstract—This project explores the effectiveness of various machine learning algorithms in classifying fashion-based images. The dataset includes 70,000 grayscale images of different fashion items, including ankle boots, bags, coats, dresses, pullovers, sandals, shirts, sneakers, trousers, and t-shirt-tops. The goal of this project is to evaluate which of three different classifiers is the most accurate. The project tests the accuracy of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and k-Nearest Neighbours (kNN). The images have been preprocessed and are then flattened into feature vectors for training the models. The dataset is split into training, validation, and test sets, with class distribution ensured through stratified sampling. Models are trained under the training set, and performance is evaluated based on its accuracy and cross-validation. The results are visually illustrated. The final model's performance is evaluated on the test set to assess its generalisation ability.

1 Introduction

This project focuses on evaluating which of three different machine learning classifiers is more accurate for the task of classifying images of fashion-based items. The dataset used in this project includes grayscale images of fashion items such as ankle boots, t-shirts, trousers, dresses, and coats, which are processed and evaluated to determine which of the three algorithms is the most effective classification model. The algorithms compared in this project are Support Vector Machines (SVM) [2], Logistic Regression [1], and k-nearest Neighbours (kNN) [3].

The dataset is first preprocessed by converting images to grayscale, resizing them, and then flattening them into arrays. Each image is then labelled according to its category and split into training, validation, and test sets. The models are trained on the training set, evaluated on the validation set, and then the best-performing or most accurate model is tested on the test set to assess its generalisation ability.

To ensure a thorough evaluation, a chart presenting the accuracy, classification report, and confusion matrix is generated for each classifier. Cross-validation is performed to validate the consistency of each model's performance [4]. Bar charts are used to compare the performance metrics of each model, and cross-validation results are presented.

Finally, based on the above tests, the most accurate model is chosen and printed for the user.

2 Features and Data Pre-processing

The dataset used in this project consists of approximately 70,000 grayscale images of fashion items, sourced from Kaggle. It is based on the Fashion-MNIST dataset, which is widely used for benchmarking machine learning models [5]. Each category is assigned a numerical label, allowing for supervised classification. The images are processed to ensure uniformity in size and format before being used in the machine learning process.

Each image was pre-converted to grayscale and resized to a standard resolution. This ensures consistency across all

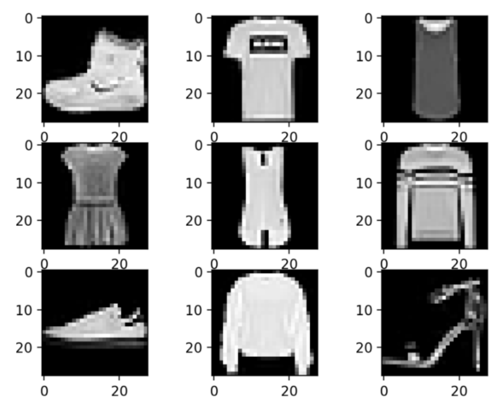

Fig. 1. gray-scaled images of fashion items in the fashion dataset

samples and reduces the complexity needed in our application. The images are flattened into one-dimensional arrays, transforming them into numerical feature vectors that can be used as input for classification models. The dataset is structured with corresponding labels, where each image is mapped to its respective class (e.g., coat, shoe, trousers, etc.).

The dataset is split into three subsets:

- Training Set
- Validation Set
- Test Set

The training set is used to train the machine learning models and consists of the majority of the data. The validation set comprises a small amount of the data; this set is used to tune hyperparameters and assess model performance. The test set is comprised of the remaining data and is reserved for final model evaluation, providing an unbiased assessment of each classifier's generalisation ability.

3 Classifier Training

Once the dataset is processed and split into the various sets, the three classification models are trained to distinguish be-

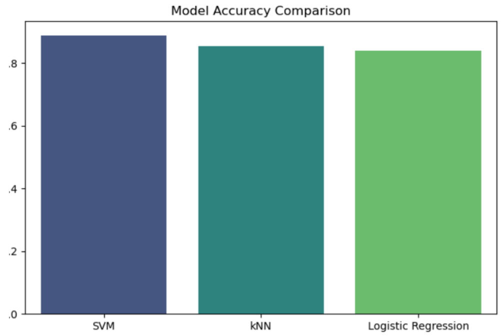

Fig. 2. modal accuracy bar chart

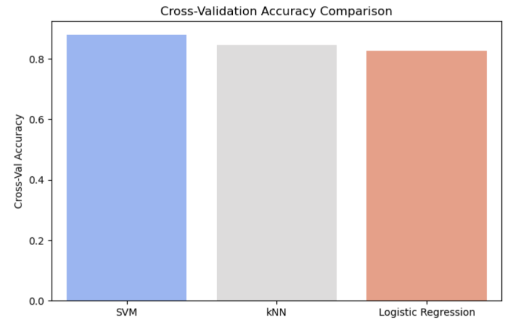

Fig. 3. Cross-validation accuracy comparison

tween the various categories. The goal of classifier training is to evaluate different machine learning algorithms, optimise their performance, and determine the most effective model for the task of classification.

1) **Model Selection**
Three classification algorithms are used in this project:

   - **Support Vector Machines (SVM)** – Uses a radial basis function (RBF) kernel to separate data points in a high-dimensional space [2].
   - **Logistic Regression** – A linear model that estimates probabilities for classification based on weighted input features [1].
   - **k-Nearest Neighbours (kNN)** – A non-parametric model that classifies data points based on the majority class of the k closest neighbours [3].

   Each classifier is initialized with default hyperparameters, and performance is evaluated using validation accuracy.
2) **Training Process**
The models are trained on the training dataset, with each classifier learning to identify patterns in the input images. The validation dataset is used to monitor model performance and fine-tune hyperparameters.
During training:

   - The SVM classifier optimizes a decision boundary using support vectors.
   - Logistic Regression applies gradient descent to minimize classification error.
   - kNN stores training samples and classifies based on Euclidean distance.

3) **Cross-Validation**
To ensure model robustness, 5-fold cross-validation is applied. This technique divides the training set into five subsets, training the model on four and validating it on the remaining fold in each iteration. The process reduces variance and prevents overfitting [4].
4) **Model Evaluation**
After training, each model's accuracy is recorded. The classifier with the highest validation accuracy proceeds to testing. Classification reports and confusion matrices are generated to analyze misclassification rates and error patterns.

4 Results Comparison

The evaluation of multiple classifiers—Support Vector Machines (SVM), Logistic Regression, k-nearest Neighbors (kNN)—was conducted to determine the most effective model for the dataset. The models were assessed based on accuracy, cross-validation performance, confusion matrices, and classification reports.

- **Model Performance**
From the three algorithms tested, SVM with an RBF kernel performed the best, achieving the highest accuracy on the validation and test sets. kNN struggled with feature scaling and was outperformed by the kernel-based methods. Logistic Regression was the weakest model, as it struggled with the dataset's complexity.
- **Accuracy and Cross-Validation**
Cross-validation was performed to guarantee model stability across different variations of data. The SVM model consistently outperformed others in cross-validation scores, indicating strong generalisation capability. The kNN and Logistic Regression had lower and less stable accuracy scores, hence they are less suitability for this task.
- **Confusion Matrix and Misclassification trends**
The SVM model had the lowest misclassification rate. kNN and Logistic Regression struggled to differentiate between certain categories, leading to increased misclassification rates.
- **visualising model comparison**
Included in the project are bar charts that compare the models performance, these bar charts confirm that SVM as the top performer, with the highest accuracy and most consistent cross-validation scores. While kNN and Logistic Regression were significantly weaker in terms of classification accuracy.
- **Best model selection and final testing**
The SVM model was the final chosen classifier suited for this task and project. It was evaluated on the test set, where it achieved the highest accuracy, reinforcing it's suitability for real-world application.

5 Conclusion

The SVM model proved to be the most effective classifier, outperforming kNN, and Logistic Regression. Its ability to capture complex decision boundaries contributed to its superior performance. The results suggest that kernel-based methods, particularly SVM with an RBF kernel, are well-suited for this dataset, making it the optimal choice for classification.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “The elements of statistical learning: data mining, inference, and prediction”. In: *Springer series in statistics* 2 (2009).

[4] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of machine learning research* 12 (2011), pp. 2825–2830.

[5] Han Xiao, Kashif Rasul, and Roland Vollgraf. “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms”. In: *arXiv preprint arXiv:1708.07747* (2017).