Task 1 Report

Andy Ngo
DATA 471
06/05/2025

1 Task

This report addresses a multiclass document classification task in which the objective is to predict the topic category of a document given its frequency vector representation. The input features are provided as sparse matrices in a triplet format, and the labels represent class IDs of the document. The goal is to build a model that performs well on unseen documents from a development set and to select the most effective model for final test set predictions.

2 Methods

We explored two supervised learning methods for this multiclass text classification problem: Logistic Regression and Support Vector Classifier (SVC).
Logistic Regression was tested because it is a well-known method for multiclass classification problems, particularly effective when combined with a multinomial loss function in high-dimensional sparse data settings.
SVC was tested as a comparative method due to its reputation for strong performance in text classification tasks, especially when using a linear kernel on sparse feature spaces.
To measure how effective both of them are, we used the model's accuracy on both the training and development sets. Logistic Regression served as the primary baseline, given its simplicity and efficiency, while SVC was used to examine potential improvements in classification performance.

3 Lead

Andy Ngo was the primary responsible for this task.

4 Submission model details

For this task, two models were implemented and compared: a multinomial Logistic Regression and a Support Vector Classifier (SVC) with a linear kernel, both using scikit-learn. The features were loaded as sparse matrices in triplet format, converted to CSR for training efficiency. Logistic regression used the "saga" solver with multi_class='multinomial' and a maximum of 1000 iterations. The SVC was run with default parameters for multiclass classification using a one-vs-rest strategy [1].
Although the SVC achieved very high training accuracy, it becomes overfit and underperformed on the development set. Logistic regression, though lower accuracy in training set, generalized better and produced higher accuracy development set. Based on this, Logistic Regression was selected for final predictions, but both implementations were retained for reproducibility and future tuning.

[1] `https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/`

5 Result

Model	Train Accuracy	Dev Accuracy
Logistic Regression	0.54623	0.49236
Support Vector Classification	0.92577	0.40109

Accuracy of both methods

Although the Support Vector Classifier (SVC) achieved much higher accuracy in the training set (92.6%), its performance dropped significantly in the development set (40.1%). This decline in accuracy indicates that the model was overfitting the training data, capturing patterns that did not generalize well to unseen examples. At the same time, the Logistic Regression model achieved more balanced performance, with a training set accuracy of 54.6% and a higher development set accuracy of 49.2%. Despite its lower accuracy on the training data, Logistic Regression outperformed SVC on the development set, showing better generalization capabilities.

In Figure 1, the precision per class for the logistic regression model, recorded using Weights and Biases, reveals significant variability between different classes. Although certain classes achieved relatively high precision scores, others lagged considerably behind. This disparity suggests underlying challenges such as class imbalance within the dataset or sparsity in feature representations for particular categories. These issues imply that some classes may have lacked sufficient representative examples or possessed insufficiently distinctive features, making them inherently harder for the classifier to distinguish.

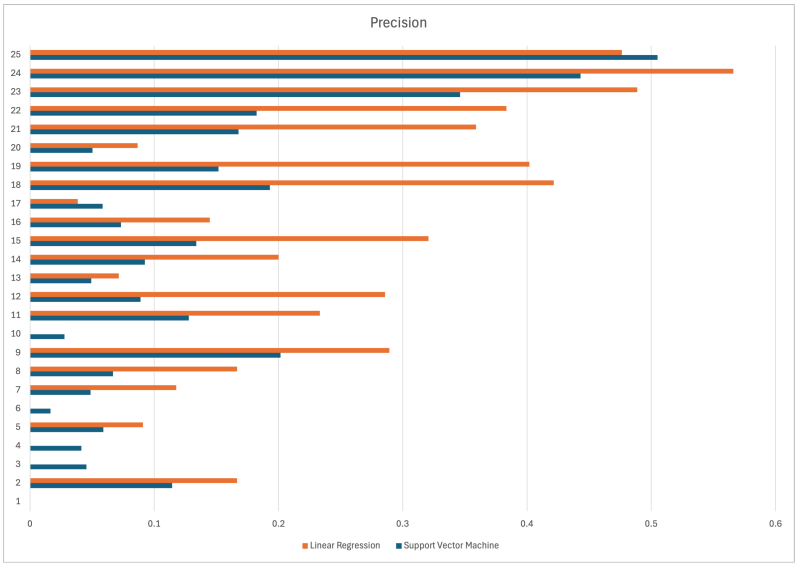

Figure 1: Precision across classes

Error analysis The big gap between SVC's train and dev performance suggests that it memorized the training data patterns without generalizing. Logistic Regression, while not achieving high training accuracy, generalized better, making it more reliable for unseen data. Based on these results, Logistic Regression model was selected for submission, as it offered superior development set performance and better generalization.

6 Distribution of work

Andy Ngo handled the majority of the implementation, debugging, experimentation, result analysis, and report writing. Jenny Sims contributed by helping set up the initial project structure and providing the code for baseline calculation. Additional advice and clarifications on scikit-learn hyperparameters and Weights & Biases logging were gathered through discussions with other group member during work sessions.