Digit Classification Report

By Dimitris Markopoulos

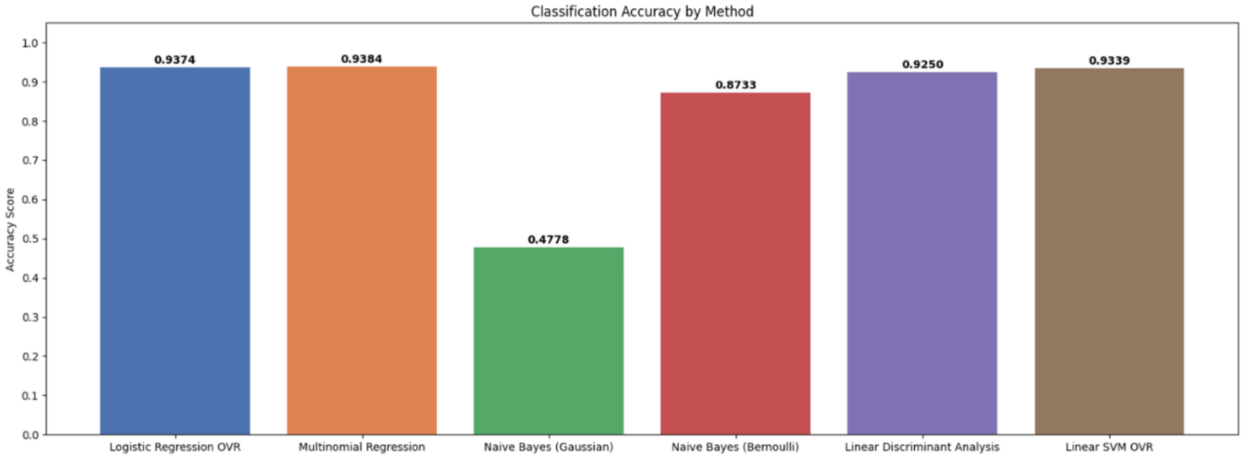
Machine Learning
Columbia University

Dear reader,

This report briefly summarizes some of the results found in MNIST_ImageClassification.ipynb .

Which method performed best in terms of accuracy?

Strictly in terms of accuracy Logistic Regression OVR, Multinomial Regression, LDA and Linear SVM OVR performed very similar; the differences in the accuracy are negligible. To answer the question in absolute terms, Multinomial Regression performed best in terms of accuracy.



But we should also consider runtime because these things matter for efficiency of implementation. If we were to perform this on a larger dataset efficiency could even be prioritized over negligible differences in accuracy.

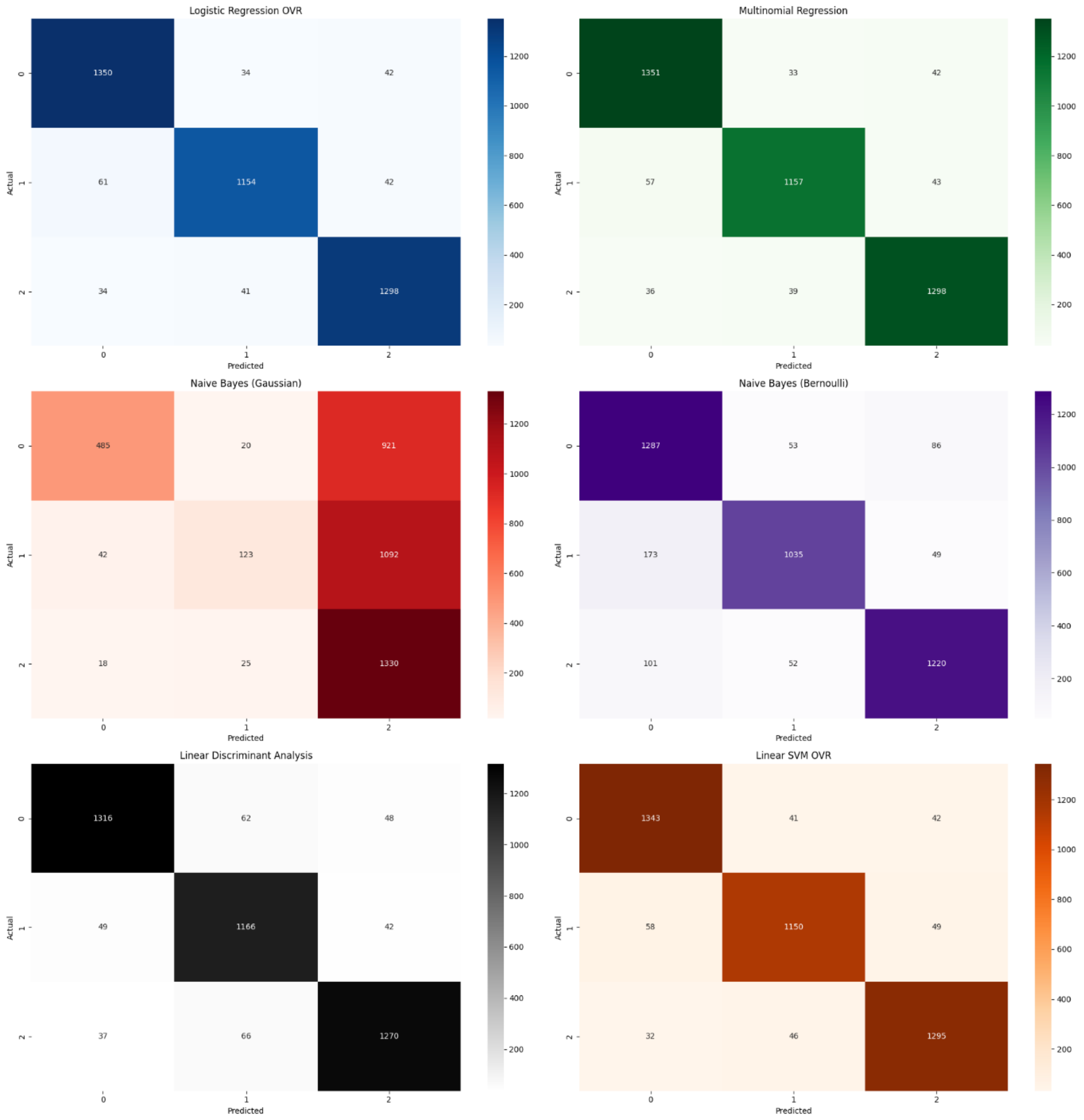
Why these results? (explained).

Multinomial Regression achieved superior accuracy (93.84%) when classifying MNIST digits 3, 5, and 8 primarily because it directly models multi-class problems without decomposing them into binary sub-problems, allowing it to better capture the subtle distinctions between these visually similar digits. Unlike Naive Bayes methods, it doesn't assume feature independence between pixels, which is crucial when distinguishing these digits' shared curved structures and enclosed spaces. The poor performance of Gaussian Naive Bayes (47.78%) confirms that pixel distributions don't follow Gaussian patterns, while the strong showing of other linear models (Logistic Regression OVR at 93.74% and Linear SVM OVR at 93.39%) demonstrates that despite their similarities, these digits remain reasonably linearly separable in high-dimensional pixel space. Multinomial Regression's ability to handle the probability distribution across multiple classes without restrictive assumptions makes it particularly well-suited for this challenging subset of handwritten digits.

Confusion Matrix. Which digit is often misclassified?

Note: Index 0: Digit 3 (with 1426 samples in the support column), index 1: Digit 5 (with 1257 samples in the support column), index 2: Digit 8 (with 1373 samples in the support column).

Now we can use the confusion matrix below to identify the digit that is misclassified the most. In general, by inspecting the confusion matrix across all methods digit 5 appears to be misclassified the most (LDA misclassifies digit 3 more but there are more samples for digit 3 so the misclassification rate can also be considered).



Reflect on results.

The classification results reveal significant performance disparities across methods when distinguishing between MNIST digits 3, 5, and 8. While Multinomial Regression achieved the highest accuracy (93.84%), followed closely by Logistic Regression OVR (93.74%), Naive Bayes methods demonstrated a clear trade-off between speed and accuracy. Naive Bayes algorithms offer computational efficiency with linear time complexity, making them substantially faster than iterative methods like SVM or Logistic Regression, especially for high-dimensional data like images. However, this speed advantage comes at a considerable cost to performance, particularly for Gaussian Naive Bayes which achieved only 47.78% accuracy due to its inappropriate distributional assumptions for pixel data. Bernoulli Naive Bayes performed better (87.33%) by treating pixel values as binary rather than continuous, but still lagged behind linear methods. This highlights the fundamental trade-off in machine learning: Naive Bayes offers rapid training and prediction with minimal computational resources but sacrifices accuracy by making the restrictive assumption of feature independence, which is particularly problematic for image data where pixel relationships are critical for classification.