Invoice recognition has the unique challenge of extracting structured information from unstructured documents, with unpredictable templates and choices of keywords. For this project, “bags of potential features” are generated to capture aspects of invoice layout, and then evaluated in multiple models to reveal the key properties that identify specific fields of interest. 8 fields of interest (including a negative class) are trained from 97 different templates, with 4.58% training error and 19.40% test error.

Three models (Naive Bayes, logistic regression, and SVM) are evaluated against the feature set:

Naive Bayes: \[ \phi_{y=k} = \sum_{i=1}^{m} I(y^{(i)} = k) + 1 \]

Logistic Regression: \[ h_{\theta}(x) = \frac{1}{1 + \exp(-\theta^T x)} \]

SVM Loss Function: \[ L(z, y) = \max\{1 - yz, 0\} \]

In the input matrix to the model, the \( (i, j) \)-entry of this matrix represents whether the \( j \)-th distinct feature appears for the \( i \)-th training data. Each text token has the potential to be included in the input matrix multiple times, to represent different feature selection rules (for instance, “invoice_nearby” and “invoice_halign” are two distinct features for text token “invoice”, indicating whether another training data is close by or horizontally aligned with it).

Regardless of the model used, performance is inevitably affected by limitation in image quality and OCR and hidden text extraction results. More work is needed to handle overfitting, which includes regularization, training data size expansion, and “Wrapper” model feature selection uptake for more precise results. Further analysis might be worthwhile on the effectiveness of the current set of feature selection rules in capturing the inherent nature of invoice layout, and perform rule-level feature selection on a larger pool of potential feature generation rules. Additionally, current results are largely affected by poor OCR and pdf text extraction performance, which suggests investigation on more optimized OCR and text extraction solutions or collection of higher quality invoice images.