Marlena Kuhn

Homework 2

Part A

A.1)

	summation form	matrix form
$SSE(\beta)$	$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$\|X\beta - y\|^2$
$MSE(\beta)$	$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$\frac{1}{n} \|X\beta - y\|$
$\mathcal{L}(\beta)$	$\frac{1}{2n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$\frac{1}{2n} \|X\beta - y\|^2$

A.2)

The minimizer is the input that causes the function to be its minimum. The chain rule states that the derivative of the MSE is $\frac{2}{n} X^T(X\beta - y)$ and that of the loss function is $\frac{1}{n} X^T(X\beta - y)$. The multiplying by a positive constant only impacts $\frac{2}{n}$, which is, itself, a constant. As the slope of a function at a minimum is 0, the derivative is 0. Scaling the loss function only impacts a non-zero constant (with the same sign) in the derivative, so the argument at both functions' minimum is the same, and the minimizer is not changed.

A.3)

The rate of change of all β is $\frac{d\mathcal{L}}{d\beta}$, so it is $\frac{1}{n} X^T(X\beta - y)$.

A.4) $\frac{d\mathcal{L}}{d\beta} = \frac{1}{n} X^T(X\beta - y) = 0$

$\times n$ $0 = \frac{1}{n}(X^T X\beta - X^T y) \times n$

$+X^T y$ $0 = X^T X\beta - X^T y$ $+X^T y$

$$\boxed{X^T y = X^T X\beta}$$

Part C

C.1)

GCLM produced the best results for the Naive Bayes (as it had the highest accuracy of the 3 extraction methods and the highest F-1 score for accuracy). PCA reduces the number of features, which may clash with Naive Bayes' assumption of independent predictors. The histogram only represents color (which may not be enough information).

PCA produced the best logistic regression results (highest accuracy and F-1 score of accuracy). The dimension reduction might have greatly helped overfitting.

LDA in this collab document does not use a seperate feature extraction.

C.2) Logistic regression under PCA feature extraction performed the best.

C.3)

LDA performed better than the Naive Bayes, and most of the logistic regression, but better than the logistic regression with PCA. LDA may have difficulty maximizing the separation of categories from raw pixels rather than using extracted features, leading to a noisy and poorly performing model.

C.4) (The ViT model could not run for me.) The advantages are a better performing model. The disadvantages are the features may not be independent, and the resources going into the ViT may not make the Naive Bayes that much better (low ROI).

C.5) Feature engineering and model performance vary from the quirks of each model type. In this case, logistic regression with PCA was best.