Machine Learning in Medicine and Biology - Homework 2

Part A — Math (show your work) 3pts

A.1 Write SSE(β), MSE(β) and $(\beta) = \frac{1}{2n}||X\beta - y||^2$ in both summation and matrix forms.

SSE(β):

- Summation form:

$$\text{SSE}(\beta) = \sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

- Matrix form:

$$\text{SSE}(\beta) = (y - X\beta)^\top (y - X\beta) = |y - X\beta|_2^2$$

MSE(β):

- Summation form:

$$\text{MSE}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

- Matrix form:

$$\text{MSE}(\beta) = \frac{1}{n}(y - X\beta)^\top (y - X\beta) = \frac{1}{n}|y - X\beta|_2^2$$

L(β):

- Summation form:

$$L(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

- Matrix form:

$$L(\beta) = \frac{1}{2n}(y - X\beta)^\top (y - X\beta) = \frac{1}{2n}|y - X\beta|_2^2$$

A.2 Explain in one paragraph why scaling the loss by a positive constant does not change the minimizer. Think in terms of the chain-rule. You don't have to formally derive it.

Scaling the loss function $L(\beta)$ by a positive constant (c > 0) — that is, replacing it with $cL(\beta)$ — does not change the minimizer β^*. Both $L(\beta)$ and $cL(\beta)$ attain their minima at the same point since scalar multiplication preserves the location of extrema for positive scalars. By the **chain rule**, if we take the derivative of $cL(\beta)$ with respect to β, we get $c \cdot \frac{dL}{d\beta}$, confirming that scaling only affects step size, not the solution.

A.3 Look closely at the Jacobian definition. $\frac{\partial}{\partial\beta} = \frac{1}{n}X^\top(X\beta - y)$ The rate of change of all of our β's is what?

The rate of change of all of our β's is given by the gradient vector

$$\frac{\partial}{\partial\beta} = \frac{1}{n}X^\top(X\beta - y)$$

Each component of this vector tells us how much the loss would increase or decrease if we changed one of the β_j's slightly, while keeping the others fixed. Specifically, the j-th entry represents the **average correlation** between the prediction error $X\beta - y$ and the j-th feature column of X. If this value is large and positive, increasing β_j will increase the loss; if it is negative, increasing β_j will decrease the loss. This gradient guides how each β_j should be updated in gradient descent.

A.4 Show that setting the gradient $\frac{\partial}{\partial\beta} = \frac{1}{n}X^\top(X\beta - y)$ to zero yields $X^\top X\beta = X^\top y$.

Set the gradient to zero:

$$\mathbf{0} = \frac{1}{n}X^\top(X\beta - y).$$

Multiply both sides by (n) (a positive number, does not change the equation):

$$\mathbf{0} = X^\top(X\beta - y) = X^\top X, \beta - X^\top y.$$

Transposing the terms yields the normal equation:

$$X^\top X, \beta = X^\top y.$$

Part C — Classification Methods Comparison Exercise 3pts

C.1 Which feature extraction method produced the best results for each classifier? Why do you think this is the case?

Feature Method	Naive Bayes Accuracy	Logistic Regression Accuracy
PCA	71.02%	**81.63%** (Best for LR)
Histogram	61.74%	63.07%
GLCM	**75.00%** (Best for NB)	75.19%

The best feature extraction method for **Naive Bayes** was **GLCM**, achieving 75% accuracy. GLCM captures texture information, which is a strong indicator for differentiating skin cancer lesions. Naive Bayes benefits from lower-dimensional, well-separated features, making GLCM a good fit.

For **Logistic Regression**, **PCA** yielded the highest accuracy at 81.6%. PCA transforms high-dimensional pixel data into orthogonal components that preserve the most variance, helping linear models like logistic regression perform better.

The Histogram feature performs worst because it loses spatial structure and only contains information about brightness distribution.

C.2 Compare the performance of Naive Bayes, Logistic Regression, and LDA. Which model performed best and under what conditions?

Model	Features Used	Accuracy
Logistic Regression	PCA	**81.63%** (Best overall)
LDA	Raw pixels	77.08%
Naive Bayes	GLCM	75.00%

Logistic Regression with PCA features outperformed both Naive Bayes and LDA, highlighting the synergy between effective dimensionality reduction and flexible linear decision boundaries.

C.3 How does LDA perform on the raw pixel data compared to the other models using extracted features? Explain why this might be a poor strategy for images.

LDA applied directly to raw pixel data achieved a respectable 77.1% accuracy, worse than Logistic Regression with PCA features (81.63%), but better than any other models. But in reality, this is generally a poor strategy for image classification:

1. Raw pixels form a very high-dimensional input space, making it prone to overfitting and numerical instability when estimating class covariances.
2. Unlike feature extraction methods such as PCA or GLCM, raw pixels fail to capture patterns like texture or shape, which are often critical for visual tasks.
3. LDA is essentially a linear projection and can easily capture noise in image data.

4. The images used in this assignment are compressed, and LDA would easily fail if using large amounts of raw pixel data.

C.4 The ViT model is a complex deep learning model. What are the advantages and disadvantages of using it solely as a feature extractor for a simple model like Naive Bayes?

Advantages

- ViT gives you very informative features, often much better than raw pixels or hand-crafted features, even from small datasets.

Disadvantages

- Naive Bayes assumes that features are conditionally independent — but ViT outputs highly entangled, abstract features.
- We're not using ViT's classification head or end-to-end training — just freezing it as extractor, which may not bring out its true powerful performance.
- ViT features are high-dimensional and abstract — hard to understand or visualize, unlike GLCM or histogram.

C.5 Summarize your findings on the relationship between feature engineering and model performance for this image classification task.

Effective feature engineering is often more impactful than model complexity.

- Using raw pixel data (e.g., with LDA) can yield acceptable results but is generally suboptimal due to high dimensionality and lack of semantic structure. Feature extraction methods like PCA and GLCM offer compact, structured representations that align better with classical models' assumptions.
- Modern deep features from ViT provide rich semantic representations. However, mismatches in model assumptions (e.g., feature independence in Naive Bayes) may limit their full potential.