Darshil Shah: 1002084331
Assignment 2

Question/Answer

Question 1

ANSWER-) **Maximum Likelihood Estimates for Gaussian Naive Bayes (GNB) Parameters**:

1. **Class Means**:
   - The maximum likelihood estimate for the mean of each feature in each class is computed as the sample mean of the feature values in that class.
2. **Class Variances**:
   - The maximum likelihood estimate for the variance of each feature in each class is computed as the sample variance of the feature values in that class.
3. **Class Priors**:
   - The maximum likelihood estimate for the prior probability of each class is computed as the proportion of samples belonging to that class in the training dataset.

Invariance to Image Type (Grayscale vs. RGB):

- These estimates do not change if we use grayscale images instead of RGB images.
- The estimation process for GNB relies solely on the feature values (pixel intensities), treating each pixel intensity as a separate feature.
- Grayscale images and RGB images provide the same feature values (pixel intensities), so the estimation process remains unchanged.

Conclusion:

- The estimates for class means, variances, and priors in Gaussian Naive Bayes remain consistent regardless of whether grayscale or RGB images are used.
- This consistency arises because the estimation process depends solely on the feature values and their distributions within each class, which are not affected by the type of image used.

Question 2

ANSWER-) **Factors Contributing to the Difference in Performance between QDA using RGB and Grayscale Images**:

1. **Covariance Matrices**:
   - QDA assumes that each class has its own covariance matrix.
   - For RGB images, where each pixel represents a separate feature, the covariance matrix for each class becomes large and complex due to the high dimensionality of the feature space.

- This complexity can lead to overfitting, especially when the number of training samples is limited, resulting in decreased accuracy.

2. **Number of Parameters**:
   - With RGB images, the number of parameters (mean vectors, covariance matrices) that need to be estimated for each class increases significantly compared to grayscale images.
   - This increased parameterization raises the risk of overfitting, especially with limited training samples, leading to reduced generalization performance.

3. **Assumption of Quadratic Decision Boundary**:
   - QDA assumes that the decision boundary between classes is quadratic.
   - This assumption may not hold true for RGB images where the relationship between pixel values across channels may not exhibit quadratic behavior.
   - Grayscale images, being single-channel, may have simpler decision boundaries that align better with the quadratic assumption.

4. **Correlation Between Channels**:
   - In RGB images, there may be correlations between pixel values across different channels (e.g., red, green, blue), violating the assumption of independence between features (pixels) in each class made by QDA.
   - Grayscale images, being single-channel, do not have this issue.

Overall, the assumptions of QDA, particularly regarding covariance matrices and decision boundaries, may not hold true for RGB images, leading to lower accuracy compared to grayscale images where these assumptions are more likely to be met.

Question 3

ANSWER-) **Reasons for Reduced Test Accuracy on Grayscale Images Compared to RGB Images in LDA and Gaussian Naive Bayes**:

1. **Feature Representation**:
   - Grayscale images only have one channel representing intensity values, while RGB images have three channels representing color intensities (red, green, blue).
   - The reduced dimensionality of grayscale images may result in loss of information compared to RGB images, affecting the models' ability to discriminate between classes.

2. **Model Assumptions**:
   - LDA assumes that the data follow a Gaussian distribution with the same covariance matrix for all classes.
   - Gaussian Naive Bayes assumes that the features are conditionally independent given the class label.
   - Grayscale images may not satisfy these assumptions as well as RGB images, leading to reduced accuracy.

3. **Complexity of Decision Boundary**:
   - RGB images often contain more detailed information due to the representation of color, leading to potentially more complex decision boundaries between classes.
   - If the decision boundaries in the grayscale image space are more complex or less well-defined, the models may struggle to accurately separate classes, resulting in reduced accuracy.

4. **Model Flexibility**:
   - The models' flexibility or expressiveness may also play a role.

- If the models are not flexible enough to capture the nuances present in the grayscale images, they may underperform compared to RGB images.
- For example, if the decision boundaries are inherently non-linear in the grayscale image space, linear models like LDA may not perform as well.

5. **Sample Size and Variability**:
   - The size and variability of the dataset, particularly the grayscale dataset, could impact model performance.
   - If the grayscale dataset is smaller or less diverse than the RGB dataset, the models may not generalize as well, leading to reduced accuracy.

Overall, the reduced test accuracy on grayscale images compared to RGB images could be attributed to a combination of factors related to the data, the models, and how they interact. Further analysis and experimentation would be needed to pinpoint the specific reasons in a given context.

Question 4

ANSWER-) **LDA**:
For LDA, the parameters estimated include the class means and the shared covariance matrix.

- For each class mean, there are n parameters (where n is the number of features).
- The shared covariance matrix is symmetric and has $\times$ $n \times n$ parameters.
- In total, the number of parameters estimated for LDA depends on the number of classes and the number of features.

Naive Bayes:
For Naive Bayes, the parameters estimated include the class means, class variances, and class priors.

- For each class mean and class variance, there are n parameters (where n is the number of features).
- The class priors are estimated for each class, resulting in C parameters (where C is the number of classes).
- In total, the number of parameters estimated for Naive Bayes depends on the number of classes and the number of features.

Image Types:

- For RGB images, the number of features depends on the number of channels (typically 3 for red, green, and blue).
- For grayscale images, there is only one channel, so the number of features is reduced compared to RGB images.

To calculate the total number of parameters estimated for each model and each image type, we need to consider the specific dataset used (number of classes and number of features). Since the dataset details are not provided in the code snippets, we cannot give an exact count of parameters. However, we can infer that the total number of parameters estimated will depend on the dataset's characteristics and the model's complexity.