Machine Learning (Problem set 1)

Ali Alipour, *University of Tehran*

This report addresses solutions to questions 1 and 2 from the Machine Learning homework. In question 1, we implement a Naïve Bayes classifier, explain its theory, and compare manual results with those from Scikit-learn, evaluating performance through accuracy, precision, recall, and confusion matrices. In question 2, we design a binary classifier to distinguish between sea and forest images based on color features, analyzing classification accuracy and misclassifications.

Question 1

Explanation of Naïve Bayes Classifier

The main difference between the Naïve Bayes classifier and the Optimal Bayes classifier is that Naïve Bayes assumes that the features in the dataset are independent from each other, which results in zero covariance. Due to the assumption of feature independence, the Naïve Bayes classifier simplifies the Bayes rule as follows:

$$P(y|x_1, x_2, ..., x_n) = \frac{p(y)\prod_{i=1}^{n} p(x_i|y)}{p(x_1, x_2, ..., x_n)}$$

Even though the independence assumption of features in Naïve Bayes is relatively simplistic, it often performs well in practice, even when the features are not truly independent. The formula for Optimal Bayes is as follows:

$$P(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

If the data distribution is normal, the relevant normal distribution formula should be used in place of conditional probability, and the covariance matrix must be considered in cases where features are not independent.

Naïve Bayes Algorithm Implementation

1. First, the specified dataset is loaded using the `pd.read_csv()` function.
2. Data preprocessing is carried out to handle missing values (`NaN`) and remove rows with incomplete data.
3. To ease computations, two helper functions were defined. One for feature normalization and another to compute the prior probability. These functions are shown in Figures 3 and 4.
4. The dataset is split into training and testing sets with a ratio of 70% training and 30% testing data.
5. For each class, the mean and variance of features are computed using the `groupby` function.
6. Using the computed means and variances, the normal distribution for each feature is calculated, and the probability for each class is computed.
7. The results are combined, and for each test sample, the class with the highest probability is selected.
8. A confusion matrix is generated to compare predicted and actual values.

Naïve Bayes Implementation Using Scikit-learn

Using the `GaussianNB` from the `Scikit-learn` library, the Naïve Bayes classifier is implemented, and its accuracy is compared with the manually implemented version. The results indicate that both methods yield the same accuracy.

Question 2

In this problem, a binary classifier is designed to distinguish between sea and forest images using color

features (blue for the sea and green for the forest). The dataset is loaded, and the RGB values are extracted for each image.

1. The path to the images is defined, and the list of images is obtained using `list_dir`.

2. For each image, the mean color values are calculated, and predictions are made based on the dominant color (blue for the sea, green for the forest). The confusion matrix is then calculated.

References

[Alipour Fraydani, 2024] Alipour Fraydani, A. (2024). Homework on Machine Learning problem set 1, University of Tehran. *Unpublished Manuscript*, Department of Electrical Engineering, University of Tehran.