University of Hertfordshire UH

School of Computer Science

BSc (Hons) Computer Science (Artificial Intelligence)

6COM1044 - Machine Learning and Neural Computing

Experimental Report

April 2024

Dhyan Nilesh Patel

Module Leader: Prof Volker Steuber

1. Task 1- Data Exploration

a. In this task, the Python functions from the Pandas library are employed to facilitate the loading and manipulation of the datasets. The primary function utilized is pd.read_csv(), which serves the purpose of reading CSV (Comma Separated Values) files into a DataFrame. Specifically, within this task, pd.read_csv() is utilized twice: once to load the training set and once to load the test set. The values provided to this function are the file paths of the CSV files containing the respective datasets.

The next step is to extract the features and label columns after the datasets have been loaded. This is achieved using the iloc[] function, applied to both the train_data and test_data DataFrames. The range of columns that include the features is represented by the selection of all rows and columns from index 2 to 31 (inclusive) using the function iloc[:, 2:32]. Additionally, iloc[:, 1] is employed to select all rows and only the second column (index 1), which is the label column (Diagnosis).

The choice of these specific column indices for feature and label extraction is based on the structure of the datasets. In this case, the second column (index 1) is selected for the label column extraction. Furthermore, the specific range of columns (index 2 to 31) is selected for feature extraction which contains the desired 30 features for analysis.

b. The scatter plot in **Figure 1** shows the relationship between the mean radius and worst radius of breast cancer cells. Each data point represents a single cell, and the colour of the point corresponds to whether the cell is malignant (cancerous) or benign (non-cancerous).

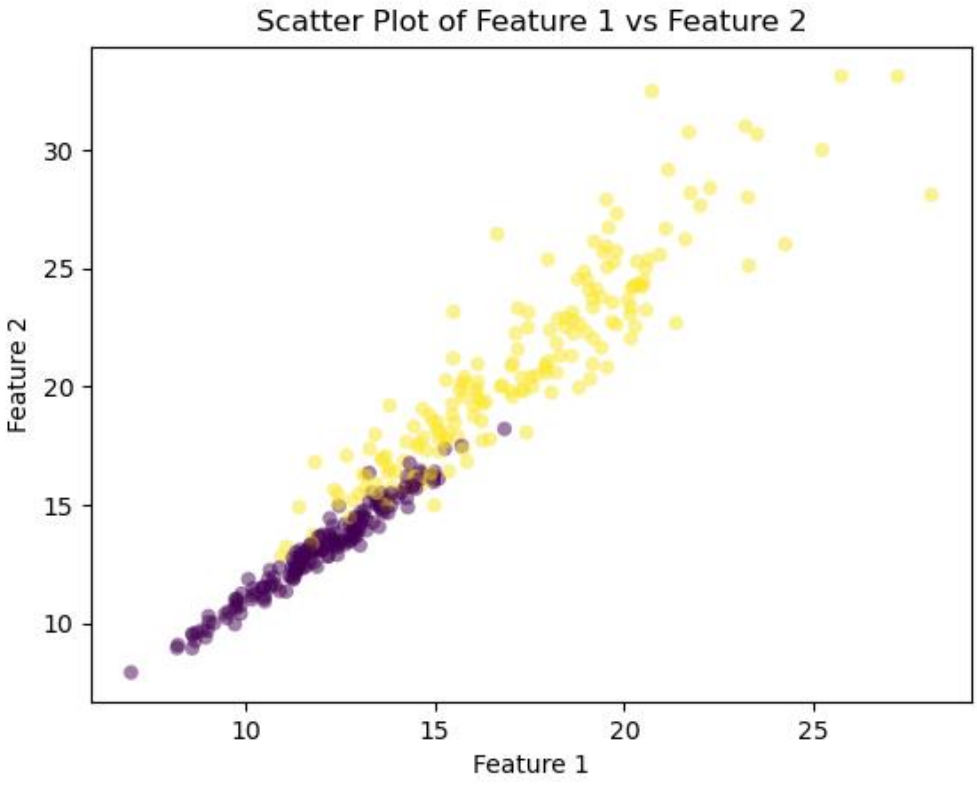

Figure 1

In the scatter plot, we can see that there is a positive correlation between the mean radius (Feature 1) and the worst radius (Feature 2). This means that cells with a larger mean radius also tend to have a larger worst radius. However, there is also a significant amount of overlap between the two classes. This suggests that mean radius alone is not sufficient to distinguish between malignant and benign cells. Moreover, there are some malignant cells (purple points) that have a smaller mean radius than some benign cells (yellow points). The spread of the data is larger for the worst radius than for the mean radius. This suggests that the worst radius may be more variable than the mean radius.

Overall, the scatter plot suggests that there is a relationship between mean radius and worst radius, but it is not a perfect relationship. This means that other features are likely needed to accurately classify breast cancer cells.

c. The data was normalized using the StandardScaler from scikit-learn library. The scaler was fit solely on the training features to learn the mean and standard deviation of each feature. This prevents any information leakage from the testing set and avoid overfitting. Both the training and test features were transformed using the scaler. This centres the data (subtracts the mean) and scales it to unit variance (divides by the standard deviation).

   The mean and standard deviation of the first feature in the normalized test set were calculated. These values are -0.38 (mean) and 0.84 (standard deviation), indicating that the first feature in the test data is now centered around a mean of -0.38 and has a standard deviation of 0.84.

d. **Subplot 1** Projection onto PC1and PC2

   This subplot visualizes the data points from the training set projected onto the first two principal components (PC1 and PC2). Each data point is coloured according to its class label (malignant or benign) in the training set. From the plot, we can observe some separation between the two classes. The purple points (malignant) tend to concentrate in a specific region, while the yellow points (benign) show a broader distribution. However, there is also some overlap between the classes, indicating that perfect separation is not achievable using only the first two principal components.

   Subplot 2 Scree Plot

   The scree plot shows the explained variance ratio by each principal component. The x-axis represents the number of principal components, and the y-axis represents the percentage of variance explained by that component. We can see a sharp decrease in explained variance ratio after the first two principal components. The first principal component explains a significant portion of the variance, followed by a smaller contribution from the second principal component. The remaining components contribute a much smaller proportion of the variance.

   Findings:

   By analysing both subplots together, we can gain valuable insights into the dimensionality of the data and the effectiveness of PCA for dimensionality reduction. The projection plot suggests that the first two principal components capture a meaningful amount of information about the class separation, even though perfect distinction isn't achieved. The scree plot indicates that the first two principal components might be sufficient to retain most of the relevant information for classification purposes. Moreover, including more principal components might not provide

significant additional benefits, and could introduce redundancy. However, we can use the explained variance ratio from the scree plot to decide on an appropriate number of principal components to retain for further analysis.

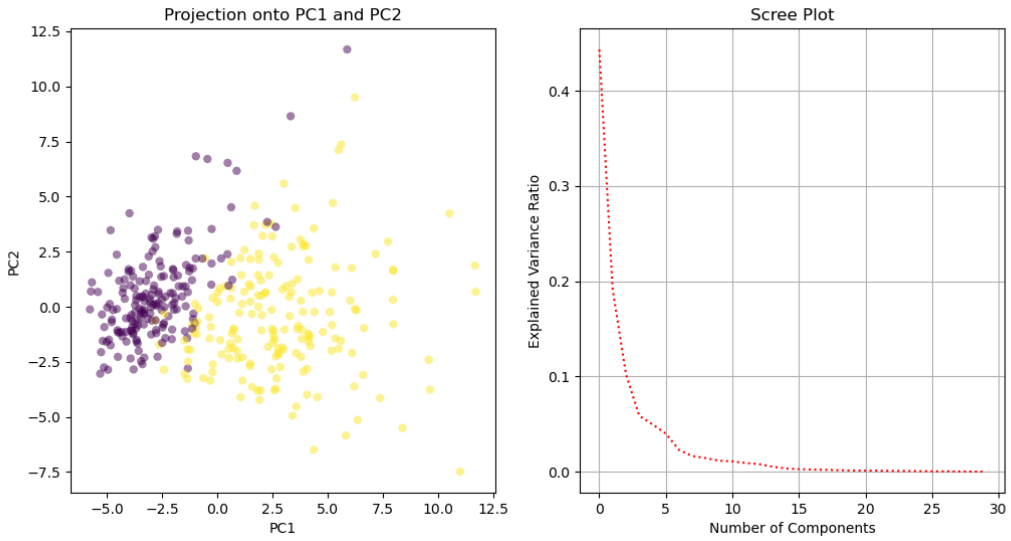

Figure 2

2. Task 2 – Data Preparation

a. To divide the training dataset into a smaller training set (II) and a validation set, the 'train_test_split' function from the scikit-learn library was utilized. A ratio of 75% for the smaller training set and 25% for the validation set was chosen, ensuring a substantial portion of the data is used for training while still having a reasonable amount for validation.

b. Normalization of both the smaller training set (II) and the validation set was carried out to ensure that all features had a mean of 0 and a standard deviation of 1, facilitating model convergence and performance. We utilized the **StandardScaler** from the **sklearn.preprocessing** module to normalize the sets. The normalization process involved fitting the scaler on the smaller training set and then transforming both the smaller training and validation sets using the parameters learned from the smaller training set.

   After normalization, the mean value and standard deviation of each feature in the normalized training set were compared to those of the corresponding feature in the normalized validation set. In general, the mean values of each feature in the normalized validation set differ slightly from those in the normalized training set. This discrepancy is expected due to the finite size of the datasets and the random nature of the data split. In addition, the standard deviation values of each feature in both sets are relatively consistent, indicating that the scaling applied to the training set has been effectively applied to the validation set as well.

3. Task 3 – SVM Classification

a. In this task, I aimed to understand how the performance of Support Vector Machine (SVM) models varies with different values of the regularization parameter *C* and different kernel functions (linear and radial basis function, RBF), including the parameter *γ* (gamma) for RBF kernels.

Linear Kernel Models:

For SVM models with a linear kernel, we examined three different values of *C* ranging from 2 to 52. These models are suitable for linearly separable data or when the number of features is large compared to the number of samples.

1. Model 1 (C=3)

   This model achieved an accuracy of 0.978, indicating that it correctly classified 97.8% of the validation set samples. The precision, recall, and F1-score for both classes (benign and malignant) were high, suggesting a well-balanced classification performance without significant bias towards any specific class.

2. Model 2 (C=20)

   The accuracy of this model was slightly lower at 0.967 compared to Model 1.

3. Model 3 (C=50)

   This model exhibited a further decrease in accuracy to 0.956. Despite the higher value of *C*, the model's performance did not improve compared to Model 1. This might indicate that increasing the regularization parameter beyond a certain point does not necessarily lead to better generalization.

RBF Kernel Models:

For SVM models with an RBF kernel, we explored three different combinations of *C* and *γ* values. RBF kernels are more flexible and can capture complex relationships between features.

4. Model 4 (C=3, gamma=0.01)

   This model achieved the highest accuracy of 0.989 among all models evaluated. By including the *γ* parameter, the RBF kernel with a relatively low value of *γ* and moderate *C* effectively captured the non-linear relationships in the data while avoiding overfitting.

5. Model 5 (C=20, gamma=1.0):

   Despite the higher value of *C* and *γ*, this model exhibited a significantly lower accuracy of 0.700. The increase in *C* and *γ* might have led to overfitting, causing the model to perform poorly on the validation set.

6. Model 6 (C=50, gamma=10):

   This model showed the lowest accuracy of 0.489 among all models evaluated. With both high *C* and *γ* values, severe overfitting likely occurred, leading to poor generalization performance.

b. The selection of the kernel type and its corresponding parameter values was based on the performance metrics obtained from Task 3(a).

For Linear Kernel

Among the three models with varying C values, Model 1 achieved the highest accuracy (0.978), indicating good performance without overfitting. Therefore, Model 1 with C=3 was selected as the optimal choice.

For RBF Kernel

Model 4 with C=3 and γ=0.01 demonstrated the highest accuracy (0.989), suggesting a well-fitted model with optimal parameter values. In contrast, Models 5 and 6, with higher C and γ values, exhibited significantly lower accuracies, indicating overfitting and poor generalization performance.

Based on these findings, the selection of the kernel type and its parameter values was guided by maximizing accuracy while mitigating the risk of overfitting. Consequently, Model 4 with an RBF kernel was identified as the preferred choices due to their superior performance on the validation set and it had the highest accuracy rate.

c. We used the RBF kernel with gamma=0.01 and C=3 to train a final SVM model on the entire normalized training set. The model was then used to predict labels for the unseen test data.

Results:

Accuracy

The model achieved an accuracy of 96.65% on the test data, indicating good generalization performance.

Confusion Matrix

The confusion matrix shows 170 true positives (correctly classified benign) and 7 false positives (benign classified as malignant). There were zero false negatives (malignant classified as benign), which is desirable for cancer classification, and 32 true negatives (correctly classified malignant).

Classification Report

The precision and recall for class 2 (benign) are high (1.00 and 0.96), indicating very good performance in identifying benign cases. The precision for class 4 (malignant) is 0.82, meaning that out of the predicted malignant cases, 82% were malignant. The recall is 1.00, indicating the model identified all malignant cases in the test set (no false negatives).

Overall, the chosen RBF kernel with the tuned parameters achieved a good balance between accuracy and identifying malignant cases (no false negatives). This is crucial for cancer classification where correctly identifying malignant cells is essential.

4. Task 4 - SVM classification with features reduced using PCA

a. Based on the scree plot analysis, I have opted to utilize 4 principal components for feature reduction. This choice stems from the observation that these 4 principal components effectively capture a substantial portion of the dataset's variance while mitigating information loss. Furthermore, this decision is supported by the trend in the explained variance ratio, which begins to plateau or exhibit a significant decrease beyond the fifth principal component. This phenomenon indicates that additional principal components beyond this threshold may not significantly contribute to elucidating the variance within the dataset.

b. For feature reduction using PCA, I applied it to both the normalized training set and the normalized test set, utilizing a chosen number of principal components (4). The process involved initializing PCA with 4 principal components, fitting PCA on the normalized training set using 'pca.fit()', transforming the training set features using the fitted PCA, and projecting the normalized test set features onto the same PCA space obtained from the training set using 'pca.transform()'.

c. In this step, classification was carried out using a Support Vector Machine (SVM) with the RBF kernel with gamma=0.01 and C=3. After reducing the features using Principal Component Analysis (PCA), both the training set and the test set were normalized using StandardScaler. Subsequently, an SVM model was trained on the reduced-feature training set. Once the model was trained, it was tested on the corresponding reduced-feature test set.

   The performance of the model was evaluated by calculating the accuracy rate and the confusion matrix. The accuracy rate achieved on the test set with reduced features was approximately 98.08%. This accuracy rate indicates the proportion of correctly classified instances out of the total instances in the test set, showcasing the effectiveness of the SVM classifier with reduced features in accurately predicting the classes of the samples in the test set. Moreover, the confusion matrix showed that the model correctly classified 173 out of 177 benign cases (true positives) and 32 out of 32 malignant cases (true negatives). There were 4 false positives (benign classified as malignant) and 1 false negative (malignant classified as benign).

d. By comparing the classification results from Task 3(c) and Task 4(c) provides valuable insights into the impact of feature reduction on the SVM classifier's performance.

   In Task 3(c), where SVM classification was performed without feature reduction, we achieved an accuracy rate of 96.65%. However, in Task 4(c), where feature reduction using PCA was implemented prior to SVM classification, we observed a slight improvement in accuracy, with an accuracy rate of approximately 98.08%. This improvement in accuracy suggests that reducing the dimensionality of the dataset through PCA helped enhance the SVM classifier's performance. By reducing the number of features while retaining essential information, the model became more efficient in capturing the underlying patterns in the data, leading to better generalization and classification accuracy on unseen test data.

   Furthermore, the confusion matrices from both tasks reveal notable changes in the distribution of correctly and incorrectly classified instances. Task 4(c) exhibits a slightly more balanced distribution, indicating better classification performance across different classes compared to Task 3(c). Moreover, Task 4(c) with PCA-reduced features exhibited slightly better performance in

identifying benign cases with only 4 false positives (benign classified as malignant). While the difference is small (3 cases), it suggests that dimensionality reduction using PCA might have helped the model slightly better distinguish between some benign and malignant cases in this specific dataset.

In critical thinking, these findings underscore the importance of feature reduction techniques like PCA in optimizing the performance of machine learning models. While the improvement in accuracy may be modest, even small enhancements can have significant implications, especially in real-world applications where precise classification is crucial. Therefore, integrating feature reduction methods into the model-building pipeline can contribute to more efficient and effective data analysis and decision-making processes.