Final Project

Machine Learning – CPSC 550 **Christina Morgenstern**

1. Motivation

As a former research scientist working in basic cancer research, I am aiming to apply my newly acquired Machine Learning skills to the wealth of biomedical data in the future in order to move precision medicine for early diagnosis and defined treatment options forward.
For that reason, I searched the CrowdFlower, Kaggle and UCI Machine Learning Repository websites for a biomedical dataset and found the *Breast Cancer Wisconsin Data Set* on Kaggle (and UCI Machine Learning Repository), which caught my interest.

2. Background

Breast cancer is a devastating disease, that affects 2.1 million women a year and contributes to the greatest number of cancer-related deaths among women. Early diagnosis is critical in improving breast cancer outcomes and survival. Evaluated cancer screening tools such as breast examination and mammography along with targeted biopsies are common techniques in order to monitor women in the first place as well as to predict the outcome of the cancerous lesion, i.e. benign or malignant, in the long term [1].

3. The Dataset

The Breast Cancer Wisconsin (Diagnostic) Data Set contains data donated by 569 breast cancer patients and was established 1995. Each patient has undergone a breast cancer biopsy and digitalized images thereof were characterized based on 32 features. The following features describing the appearance of the cell nuclei in these tissues are part of and correspond to the individual columns in the dataset:

1) ID number
2) Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error (labelled: se) and worst (mean of the three largest values) of features a-j were computed for each image, resulting in 30 features in total. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.
The class distribution (column: diagnosis) is with 357 benign (B) cases and 212 malignant (M) cases.

Credit must be given to Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian, the creators of the Breast Cancer Wisconsin Data Set. See references 2-7 for original publications on the generation and analysis of the dataset.

4. The Problem

The Breast Cancer Wisconsin Data Set frames a classification problem in supervised learning, where known instances (benign or malignant cancer cases) serve as labeled input to the prediction model. My goal is to generate several models based on this dataset, that are capable of predicting accurately, if a new unseen instance of breast tissue is benign or malignant. Individual models will be evaluated against each other and the best performing model will be chosen for further prediction analysis.

5. Data Analysis Approach

For the data analysis and the building of a machine learning model, I choose to work with the programming language Python (Version 2) using the Jupyter notebook environment.

5.1. Data Preprocessing

After loading the data into the programming environment of Jupyter, I inspected the columns and rows of the dataset in terms of completeness. Missing data was deleted, such as the last column (column 32) in the dataset that didn´t have any values. Also, the ID column (column 1) was removed from the analysis because it doesn´t provide any useful information.
In order to prepare the data for the analysis algorithm, I had to transform the class labels M (malignant) and B (benign) to integer values 1 and 0, respectively, which were stored in the form of an array using variable *y*. The remaining 30 feature columns were initially stored together as an array in the variable *X*.

In this type of machine learning, also known as supervised learning, the entire dataset is divided into a training set and a testing set, that the computer uses as example data to train the model and to ensure generalization, respectively. Therefore, 80% of the original data was used for training the model, and 20% for the model test.

Both training and testing data sets were further standardized, meaning that the resulting data is centered around 0 with a standard deviation of 1and numerical columns were further normalized with the aim to bring the features onto the same scale.

5.2. Feature selection

Feature selection refers to choosing relevant features from the original feature set, in this case 30, in order to find features that are most relevant to the problem and to further reduce the complexity of the model. There are several algorithms, that help with feature selection, like the *Sequential Backward Selection (SBS)* or the *Random Forest Algorithm*, both of which were applied in my analysis. The goal of feature selection is always tied to feature reduction in order to obtain simpler models that are easier to work with and are more interpretable.

The Sequential Backward Algorithm showed that with only 8 features, out of 30, the model is still close to 100% accurate (Figure 1).

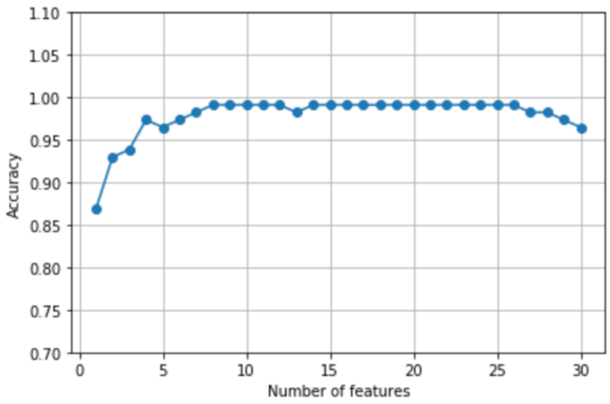


Figure 1. Result of the Sequential Backward Algorithm on the Breast Cancer Wisconsin Data Set.

The 8 features sufficient to yield good model performance were the following:

- Radius_mean
- Texture_mean
- Smoothness_mean
- Concavity_mean
- Symmetry_mean
- Radius_se

- Concavity_se
- Concavity_mean

Another way of assessing feature importance was done using the random forest algorithm. The result from this analysis ranks all features according to their relative importance (Figure 2) with five features (concave points_worst, perimeter_worst, radius_worst, area_worst and concave points_mean) apparently sufficient to explain the entire dataset.

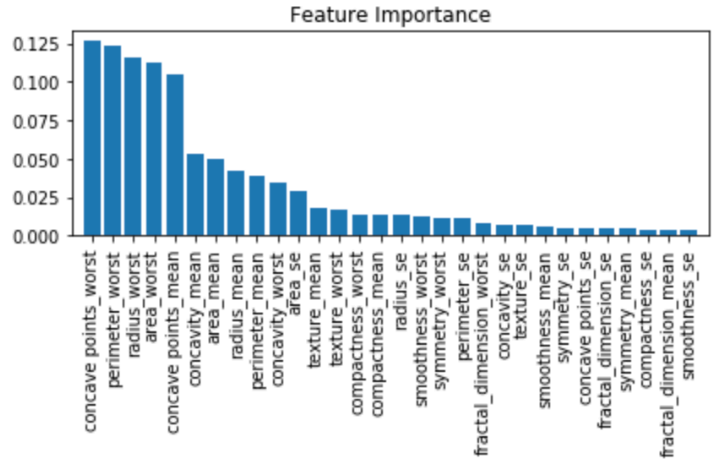


Figure 2. Result of Random Forest Algorithm for feature selection on the Breast Cancer Wisconsin Data Set.

In conclusion, feature selection and reduction analysis can help in finding relevant features in the dataset as well as reduce the dimensionality, however, as seen in my case, different techniques will yield a different set of important features. This is because of the inner workings of the individual algorithms.

6. Modelling Step

In order to find a useful model, I used quick prototyping and assessed the accuracy of several different supervised learning algorithms, which are described in the following.

6.1. K-nearest neighbors (KNN)

In the KNN algorithm, new instances are classified based on their closeness/relatedness to the nearest instances of the training data.

Executing the KNN algorithm on the original standardized and normalized dataset containing all features yielded the following accuracy scores for training and testing data run: 98% and 96%, respectively.

```
'Training accuracy:', 0.9802197802197802
'Test accuracy:', 0.9649122807017544
```

Applying the KNN algorithm to the reduced dataset with only 8 features from the Sequential Backward Selection (see 5.2) yielded reduced prediction accuracy with 96% and 95% for training and testing accuracy, respectively.

```
'Training accuracy:', 0.9692307692307692
'Test accuracy:', 0.956140350877193
```

Thus, using only about a quarter of the data didn´t improve the performance of the KNN algorithm, indicating that these 8 features do not provide less discriminatory information than the original dataset.

6.2. Random Forest Classifier

A random forest classifier makes use of an ensemble of decision trees where after each iteration a sample from the training set is drawn and a decision tree is grown, selecting features randomly and performing a split at each node. For this classifier, the original dataset was split into 70% training and 30% testing data because the 80/20 split yielded lower accuracy.
The resulting prediction accuracy of the random forest classifier is 97%.

```
Accuracy is: ', 0.9707602339181286
```

The confusion matrix is as follows:

```
array([[106,   2],
       [  3,  60]])
```

Out of 171 samples, 106 samples were correctly classified as class 0 (benign) with 2 false positives, 60 instances were correctly classified as class 1 (malignant) with 3 cases misclassified.

6.3. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis, which is also a dimensionality reduction technique, makes predictions by estimating the probability that a new input belongs to either class. Consequently, the class with the highest probability is the output class.
With 97% accuracy, the LDA algorithm did a good job on predicting the classes (Figure 3).

The confusion matrix states that out of 171 samples, 106 samples were correctly classified as class 0 (benign) with 2 false positives, 57 instances were correctly classified as class 1 (malignant) with 6 cases misclassified.

The recall score of the LDA model is 90% which means that 90% of the actual cases were identified correctly. However, this means that 10% of people are getting the wrong diagnosis. The F1 score is also a measure of the test´s accuracy and considers both precision and recall measures. In my case the F1 score is 93% which is really good because best performing models reach a F1 score of 1.

Also, the ROC curve (Figure 3) demonstrates a good separability of the two classes using the LDA model.

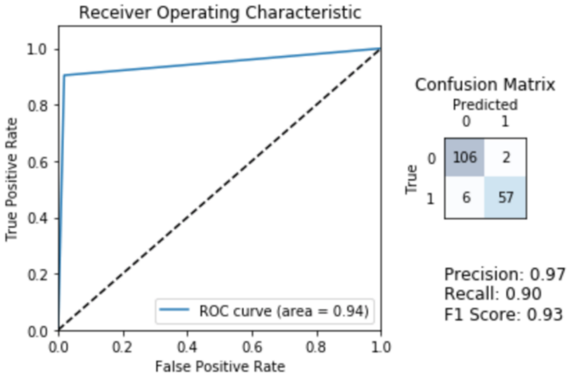
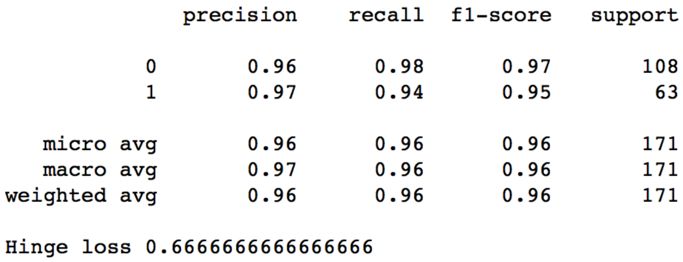


Figure 3. Linear Discriminant Analysis on Breast Cancer Wisconsin Data Set.

6.4. Support Vector Machine (SVM)

The goal of a support vector machine algorithm is to find a hyperplane in an N-dimensional space that classifies the data points.

The result of the SVM model demonstrates also a good working on the Breast Cancer Wisconsin Data Set with 96% of the benign and 97% of the malignant cases being correctly classified (see Figure 4, precision scores). High recall (98% and 94%) and F1-scores (97% and 95%) support the promising model results.



```
             precision    recall  f1-score   support

           0       0.96      0.98      0.97       108
           1       0.97      0.94      0.95        63

   micro avg       0.96      0.96      0.96       171
   macro avg       0.97      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171

Hinge loss 0.6666666666666666
```

Figure 4. Results of the Support Vector Machine Algorithm applied to the Breast Cancer Wisconsin Data Set.

7. Comparison of model performance

When comparing the model performances, one thing to note is that all four classifiers did perform well with an accuracy of greater 90%. The reason for this good performance might be that the underlying data is well described, that there is not too much randomness in the data and that the two classes (benign and malignant) are good separable based on the given features.

When comparing model accuracy metrics among my models, the best results were obtained using the KNN algorithm with 98% training accuracy and 96% testing accuracy.

A real comparison of the models is at this point unsatisfactory because not all models had the same data or train-test-split ratio as input. For example, the random forest classifier performed well on a 70%-30% train-test-split ratio whereas all other models had an 80%-20% split. Also, not all models had standardized and/or normalized data as input which makes comparison also skewed.

8. Conclusion

When comparing my results to the original endeavors of analyzing this dataset [4], the performance accuracies of my algorithms are lagging behind. In Wolberg *et. al* [4], the best algorithm is a SCH-based classifier that demonstrates the highest training accuracy (100%) and test accuracy (99%). Also, the SVM in Wolberg *et. al* [4] outperformed my SVM with 99.2% and 97% training and test accuracy, respectively. In contrast, my LDA analysis showed higher precision than the LDA in the Wolberg paper [4].

Ideally, a "cancer prediction algorithm", should have an accuracy of 100% because misclassifying a case either as false positive (being malignant when actually benign) or false negative (being benign when actually malignant) can have devastating effects for the individual. However, an algorithm that has > 90% accuracy might still perform much better than traditional prognostic options and hence be a valuable tool for clinicians.

9. References

1 https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
accessed on the 3rd of July 10 pm CET

2 W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

3 O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.

4 W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

5 W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April 1995.

6 W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995;130:511-516.

7 W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792--796, 1995.