Department of computer science
Umeå University

March 6, 2021

Machine Learning
5DV194 Spring 2021
Assignment 1

Name	**Email**
Klas Holmberg	hed16khg@cs.umu.se

Teacher: Lili Jiang

Contents

Department of computer science
Umeå University March 6, 2021

1 Introduction

Machine Learning is a diverse field of algorithms based on the notion of learning from data. The purpose of the assignment is to explore this space of algorithms a bit and to get a basic understanding of how they work and what results they produce, and also to get familiar with the `Scikit-learn` toolkit used for simulations. The data that is used as a basis for the algorithms is the Wisconsin Breast Cancer dataset [1]. Mainly there are four algorithms that where explored.

2 Algorithm 1: KNN

The `KNN` (K-Nearest Neighbours) algorithm is a supervised learning algorithm that tries to classify data through a method of looking for clusters of data where the points are near each other within the searchspace in some sense. By dividing up the searchspace into zones where datapoints reside and having these zones define what the test data shall be classified as. The fitting of the zones is what makes this algorithm tunable, essentially, how many (`K`) close neighbours are needed to define a zone of classification. Different values of `K` was employed, the results of which can be seen in Table 1.

Table 1 – The results of the KNN algorithm for the wisconsin dataset, without any preproccessing done on the training data.

K	Accuracy on training data	Accuracy on testing data
1	1.00	0.92
2	0.97	0.90
3	0.95	0.92
4	0.95	0.92
5	0.95	0.93
6	0.94	0.93
7	0.94	0.93
8	0.94	0.93
9	0.94	0.93
10	0.94	0.93

3 Algorithm 2: Logistic Regression

The second Machine Learning algorithm explored is Logistic Regression, an algorithm that tries to learn the correlation between input and output as a function that outputs true or false, and then tries to predict the output based on the input and the function that has been computed. In Table 2 the results of the logistic regression model can be seen, these are computed with different values of `C` (constant for regularisation strength) ranging from 70 to 110 and 0.007 to 0.011.

Table 2 – The results of the Logistic Regression algorithm with varying values of C, maximum iterations set to 5000 and solver set to 'liblinear'.

C	Accuracy on training data	Accuracy on testing data
70	0.972	0.958
80	0.977	0.965
90	0.972	0.958
100	0.977	0.965
110	0.974	0.965
0.007	0.955	0.930
0.008	0.957	0.930
0.009	0.957	0.930
0.010	0.957	0.937
0.011	0.955	0.937

4 Algorithm 3: SVM

The 3rd algorithm is SVM (Support Vector Machines), a supervised algorithm which uses kernel functions to discern a hyperplane that splits the data points into groups that then are used to classify the input. The two kernel function can be seen in Code 1. The results of using the two different functions on the same dataset can be seen in Table 3. What is interesting here is how the Guassian kernel function has roughly 20% lower accuracy in comparison to the Linear function on the training set but only 15% lower on the test set.

Code 1

```
# kernel=linear
def my_kernel_linear(xi,xj):
    return np.dot(xi, np.transpose(xj))

# kernel=Gaussian = RBF
def my_kernel_gaussian(xi,xj):

    sigma = 0.485
    x = np.dot(xi, np.transpose(xj))
    res = -(x**2)/(2*sigma**2)
    return np.exp(res)
```

Table 3 – The results of the SVM model using different types of kernel functions. The Guassian- (rbf) and Linear- kernel functions are depicted in Code 1

Kernel	Accuracy on training set	Accuracy on testing set
Default	0.985	0.974
Linear	0.980	0.965
Gaussian	0.787	0.807

5 Algorithm 4: Decision tree

The last algorithm is the Decision tree algorithm. This was employed on a different set of depths to see when results started to converge, the best results on this setting was achieved on the depth of 5, see Table 4.

Table 4 – The results of the decision tree model with different maximum depth settings.

Depth	Accuracy on training set	Accuracy on testing set
2	0.955	0.923
3	0.962	0.937
4	0.974	0.937
5	0.988	0.951
6	0.995	0.944
7	1.0	0.930

The feature of highest importance to get good results was discerned to be 'Worst Perimeter', see Figure 1

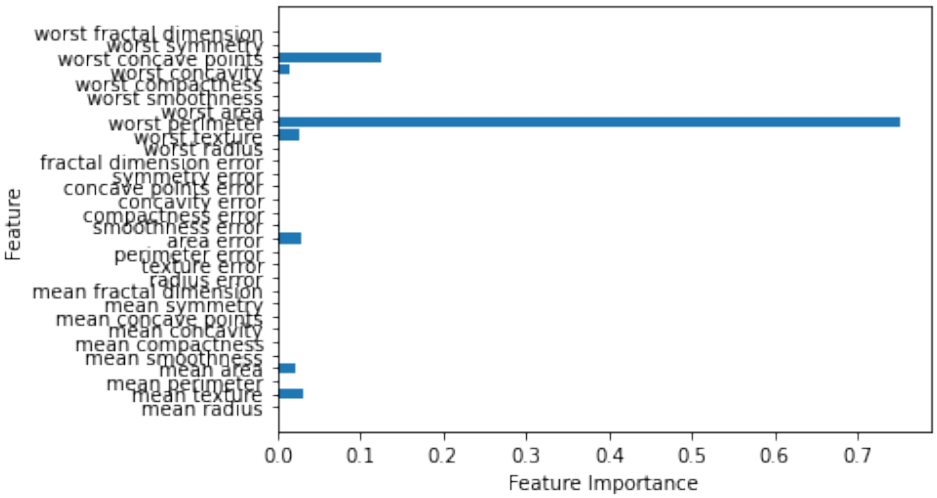

Figure 1 – A graph depicting the features against their importance to achieve good results.

6 Reflections

It was interesting to see the differences in the algorithms on such a basic level, but it is also hard to actually discern how and why they are better/worse on this type of input dataset. The results are clear, but why the results are what they are is harder to understand.

The toolkit is brilliant in my opinion for school assignments, i wish computer assignments always had this kind of format. It is clear what is to be done and why. The sklearn and

the libraries used have alot of support online so getting help if one is stuck is usually really easy. It was really hard to write a report that was only 1 - 1.5 pages if one wanted to add tables of the results though, i really tried to keep it short.

References

[1] *Wisconsin Breastcancer Database*
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)