Determine classifier strengths

MVE440 Statistical learning for big data

Ella Guiladi
guiladi@student.chalmers.se

Project aim

- Comparison of two classifiers on two dataset
  - Logistic regression vs. k-Nearest Neighbours (kNN)
  - The two chosen dataset: iris and wine

- All the training and testing of the model was done with sklearn.datasets package in python.

Methods

- Standardized the datasets
  - The wine dataset has features with different unit and in different scales.
  - Specially important to standardise the data when performing kNN, since it operates on the distance between the data points

- Stratification
  - **Stratify** wine dataset over the classes, since it is unbalanced .
    - Use build in stratification function “StratifiedKFold” from sklearn.datasets package .

- Methods for model validation
  - **Holdout method for the wine dataset**
    - Split into training data 70% and test data 30%. Since both datasets are pretty small the test set might be too small, which is why the holdout method might not be relevant in this case.
  - **5-fold cross-validation for the iris dataset**
    - Chose cross-validation for both datasets since it works better for smaller datasets.
    - Use build in cross –validation functions from sklearn.datasets package.

- Methods for analysing the classifier strength
  - Training and predicting the model on all features.
    - Look at **Accuracy**
    - Analyse **Confusion Matrix**
    - **Classification report** which can be calculated from the confusion matrix (includes precision, recall, f1-score and accuracy).

  - Training and predicting the model on two features at the time.
    - **Decision boundaries**

Datasets

Datasets are chosen from sklearn.datasets package in python.

IRIS DATASET

Multivariate balanced dataset, which consists of 50 samples from each of three species of Iris flower.

- **Attribute Information (features):**
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
- **Species (classes):**
  - Iris-Setosa
  - Iris-Versicolour
  - Iris-Virginica

WINE DATASET

Multivariate unbalanced dataset, which consists of 178 observations with the class distribution: 59, 71, 48.

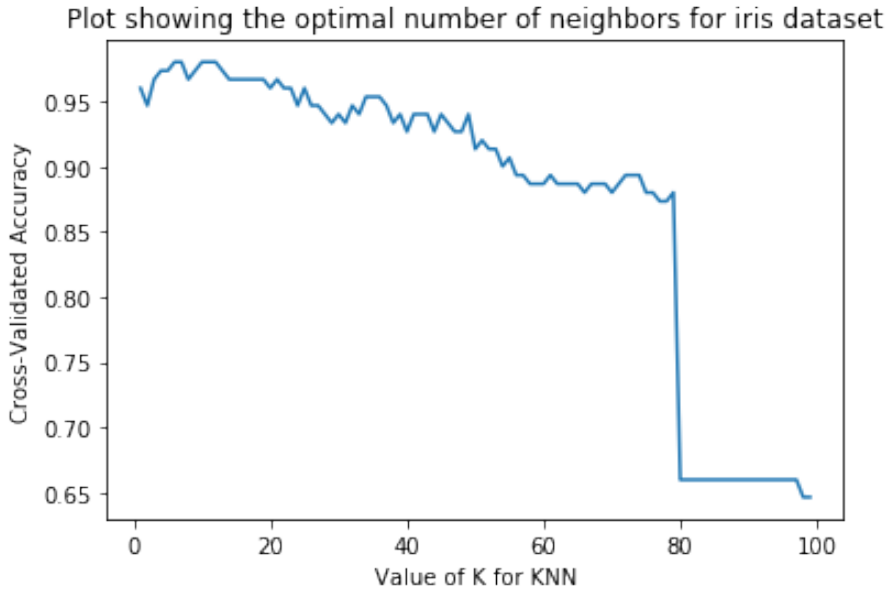
- **Attribute Information (features):**
- 13 features {Alcohol, Malic acid, Ash…}

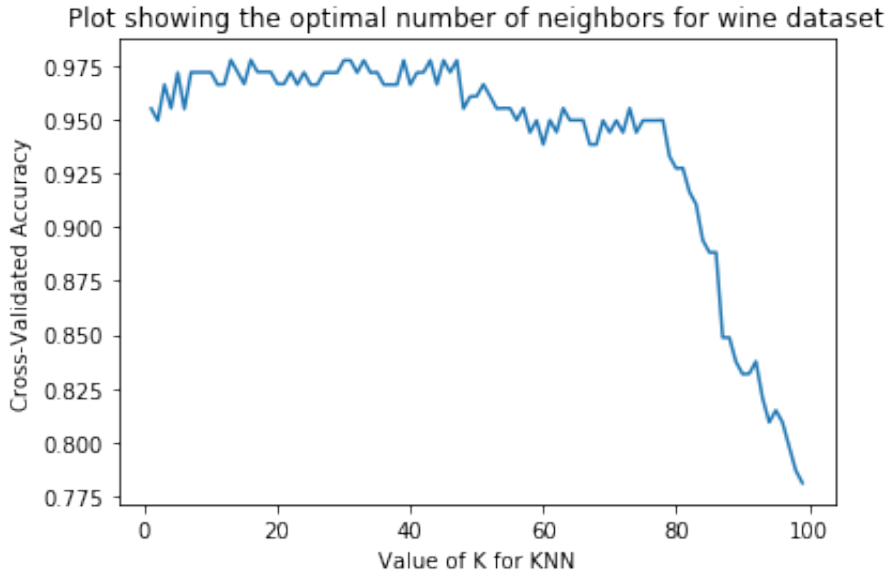
- **Cultivators in the same region (classes):**
- Class 1
- Class 2
- Class 3

Choosing optimal k for kNN

- The algorithm for both datasets when using kNN with cross-validation:
  - Choose a random k.
  - Train and test the model with cross validation and loop over a range of different values of k.
  - Then choose the k that gives the optimal accuracy over the range.

- NOTE: in unbalanced dataset, predicting everything as the majority class can still achieve good accuracy. Since the accuracy is used for choosing the optimal k, one can assume that the true accuracy might have a lower value then the resulted one, i.e. the resulted optimal k might not be the true optimal one.





Wine dataset - results

kNN:

- Accuracy = 0.97

Classification report:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	59
1	1.00	0.92	0.96	71
2	0.96	1.00	0.98	48
accuracy			0.97	178
macro avg	0.97	0.97	0.97	178
weighted avg	0.97	0.97	0.97	178

- Confusion matrix using cross-validation

	Predicted class		
True class	59	0	0
	4	65	2
	0	0	48

Logistic Regression:

- Accuracy = 0.98

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	59
1	0.99	0.97	0.98	71
2	0.96	0.98	0.97	48
accuracy			0.98	178
macro avg	0.98	0.98	0.98	178
weighted avg	0.98	0.98	0.98	178

- Confusion matrix using cross-validation

	Predicted class		
True class	59	0	0
	0	69	2
	0	1	47

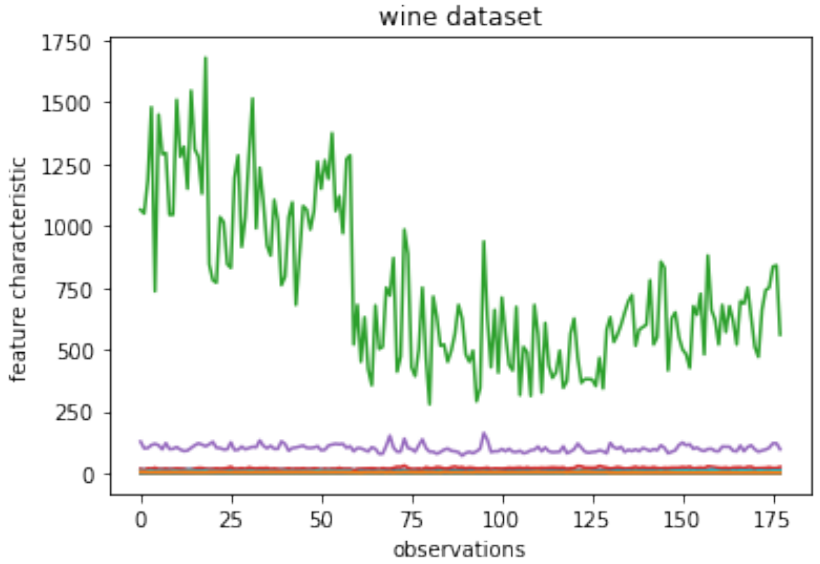
Wine dataset - discussion

- Logistic regression has higher accuracy , better classification report than kNN.

- Can be due to that kNN performs poorly with high dimension p (here p=13), due to the curse of dimensionality.
  - KNN operates on the distance between the data points. The distance of the data points is inversely proportional to the exponential increase in the number of data points, leading to the curse of the dimensionality.

- Also possible to see from the feature plot that the features have different scales and since kNN requires homogenous features this might lead to less accurate performance, which is due to the data not being pre-processed.

- Accuracy of kNN might be lower then the results show since the data set is unbalanced. Since the accuracy for kNN still is lower than for logistic regression, accuracy is still a good metrics for measuring the classifier strength.



Iris dataset - results

kNN:

- Accuracy = 0.97

```
Classification report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        50
           1       0.92      0.98      0.95        50
           2       0.98      0.92      0.95        50

    accuracy                           0.97       150
   macro avg       0.97      0.97      0.97       150
weighted avg       0.97      0.97      0.97       150
```

- Confusion matrix using cross-validation

		Predicted class		
True class	50	0	0
	0	49	1
	0	4	46

Logistic regression:

- Accuracy = 0.96

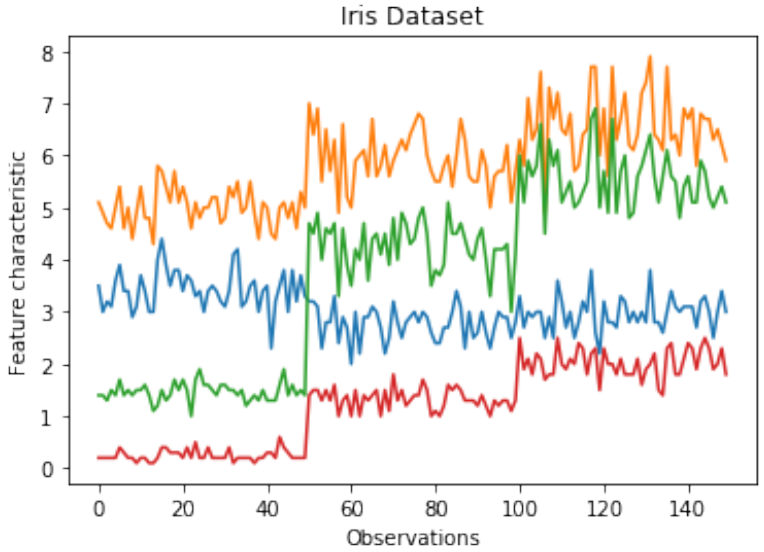
```
Classification report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        50
           1       0.94      0.94      0.94        50
           2       0.94      0.94      0.94        50

    accuracy                           0.96       150
   macro avg       0.96      0.96      0.96       150
weighted avg       0.96      0.96      0.96       150
```

- Confusion matrix using cross-validation

		Predicted class	
True class	50	0	0
	0	47	3
	0	3	47

Iris dataset - discussion

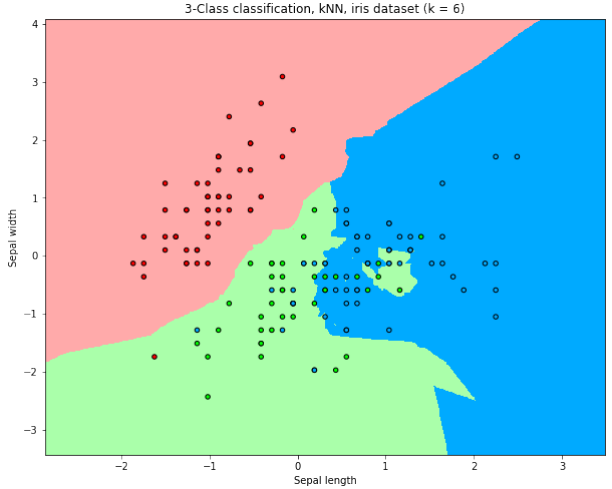
- kNN works better in lower dimension (here p=4), might be due to curse of dimensionality.
- Logistic has a slightly lower accuracy and slightly lower results in the classification report and confusion matrix.
- The accuracies are however pretty similar, might be due to the classes having some linearly separable tendencies. However, as seen in the feature plot, some features might overlap, which complicates classification with linear decision boundaries.

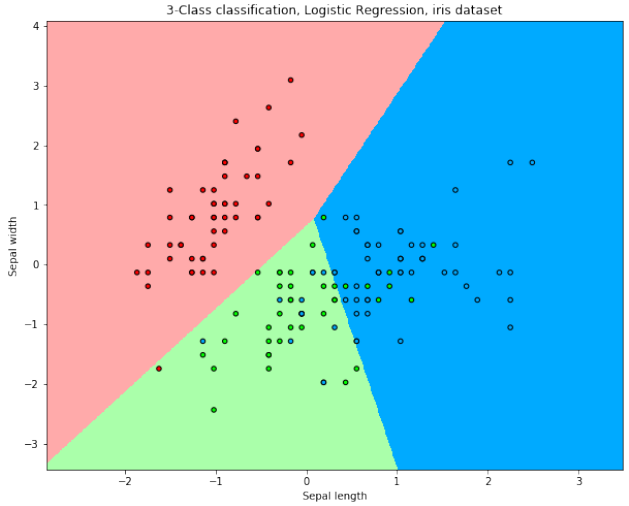


Decision boundaries

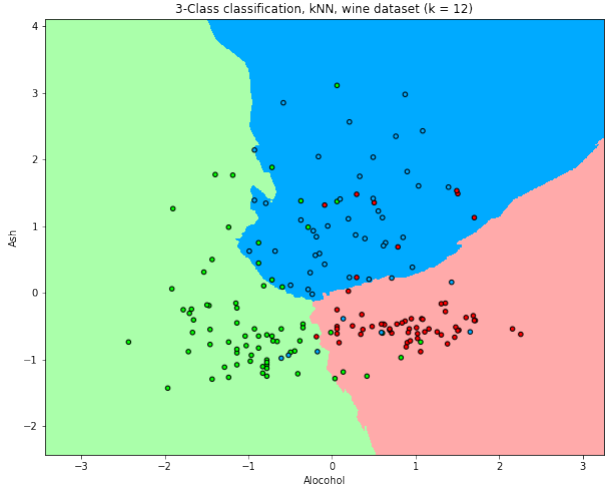
Another method to analyse the classifier strength is to train and predict the model on two features at the time instead of including all.

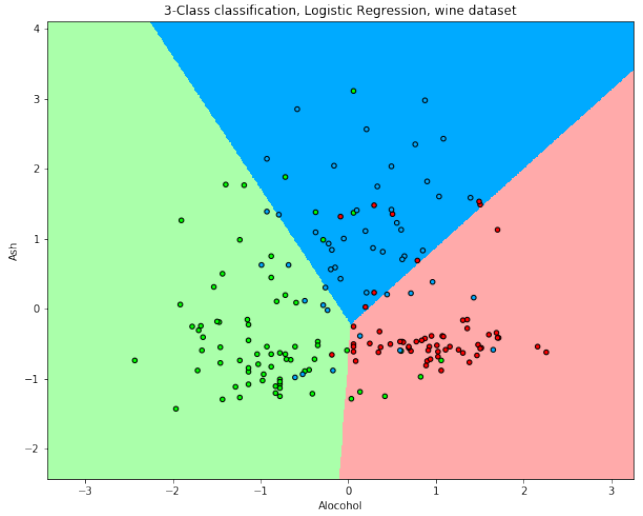
- Iris dataset





- Wine dataset





Conclusion

- kNN is a better classifier for the iris dataset than the wine dataset since it works better in lower dimensions, due to the curse of dimensionality. kNN also tend to be outperformed in unbalanced datasets.

- Logistic regression is a better classifier for the wine dataset than for the iris dataset due to the high dimensionality and due to the linearly separability of the classes in the data set, since logistic regression uses linear decision boundaries.

- Cross-validation and holdout method gives different result for smaller datasets. Cross-validation is preferred for smaller datasets since the test-set in holdout method becomes too small for testing.