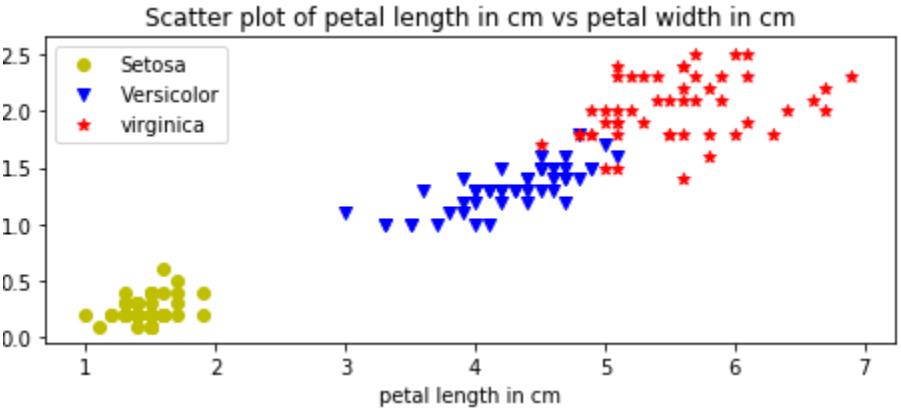
Solutions and Screenshots

Q1.1 What is the role of the fit and predict methods?
A1.1 Fit is used to fit the model based on the training data X on target y and predict is for classifying the test data based on training data.

Ans 1.2

```
Accuracy of predicted_iris 0.9933333333333333
Accuracy of predicted_sepal 0.82
Accuracy of predicted_petal 0.9666666666666667
Precision score of predicted_iris 0.9934640522875816
Precision score of predicted_sepal 0.8205539943797672
Precision score of predicted_petal 0.9667867146858743
Recall score of predicted_iris 0.9933333333333333
Recall score  of predicted_sepal 0.82
Recall score  of predicted_petal 0.9666666666666667
F1 score of predicted_iris 0.9933326665999933
F1 score of predicted_sepal 0.8198378540686618
F1 score of predicted_petal 0.9666633329999667
Confusion matrix of predicted_iris [[50  0  0]
 [ 0 49  1]
 [ 0  0 50]]
Confusion matrix of predicted_sepal [[50  0  0]
 [ 0 38 12]
 [ 0 15 35]]
Confusion matrix of predicted_petal [[50  0  0]
 [ 0 47  3]
 [ 0  2 48]]
```

Because our training and texting data is same **predicted_iris** has close to 100 percent accuracy. Next best model is **predicted_petal** as feature set used is petal_length and petal_width and this feature set clearly distinguish all 3 classes of iris data as seen below -



Ans 1.3

```
10 fold cross validation mean 0.9733333333333334
10 fold cross validation standard deviation 0.04422166387140532
5 fold cross validation mean 0.9800000000000001
5 fold cross validation standard deviation 0.016329931618554516
```

To prevent overfitting, we use K fold cross validation where we divide data into k-1 folds and train our model using them. The remaining set of data is used to validate the model.

As our data set is small 5-fold cross validation performs better compared to 10 fold cross validation

Ans 1.4

Implemented function performs nearly as good as sklearn's cross validation.

```
Custom fold mean 0.9600000000000002
Custom fold std 0.038873012632301994
```

If instead of 5-fold cross validation we do a 4-fold cross validation, then implemented function works better compared to sklearn's cross validation.

```
4 fold cross validation mean 0.9797008547008548
4 fold cross validation standard deviation 0.011752136752136757
Custom fold mean 0.9866642958748222
Custom fold std 0.013338074706322667
```

Solution 2

Ans 2.1 categories specify which category to load if the list of categories is provided as parameter, load all categories if None is provided.

Ans 2.2

CountVectorizer - It converts text data into token counts matrix.

TfidfTransformer - The count matrix is transformed to a normalized tf or tf-idf

TfidfVectorizer - It converts text data TF-IDF feature matrix.

Ans 2.3

```
Accuracy of Random Forest 0.7336696760488582
Precision score of Random Forest 0.7415305712607776
Recall score of Random Forest 0.7336696760488582
f1 score score of Random Forest 0.7268115977604259
Confusion matrix of Random Forest [[204   3   1 ...,   4   3  11]
 [  3 244  27 ...,   1   1   2]
 [  1  24 300 ...,   0   0   1]
 ...,
 [ 16   2   0 ..., 299   7   1]
 [  3   4   1 ...,   5 141   0]
 [ 42   4   3 ...,   2   5  70]]
```

Ans 2.4

```
Accuracy after pipleine 0.7254381306425917
Precision after pipleine 0.7362047314735485
Recall score after pipleine 0.7254381306425917
f1 score after pipleine 0.7198975111843331
Confusion matrix after pipleine[[192   2   4 ...,   5   0  14]
 [  2 256  26 ...,   0   1   1]
 [  4  32 282 ...,   0   1   0]
 ...,
 [ 20   4   1 ..., 289   5   1]
 [  1   4   1 ...,   6 143   0]
 [ 36   2   1 ...,   3   4  75]]
```

Ans 2.5

```
Accuracy after MLPClassifier 0.6477695167286245
Precision after MLPClassifier 0.7105842635164872
Recall score after MLPClassifier 0.6477695167286245
f1 score after MLPClassifier 0.6592959429797546
Confusion matrix after MLPClassifier [[248   0   0 ...,   0   8  29]
 [ 11 183  60 ...,   0   1   3]
 [  6  14 240 ...,   1   0   6]
 ...,
 [  8   0   0 ..., 213   5  25]
 [ 56   0   0 ...,   0 145  18]
 [ 92   0   0 ...,   0  16 107]]
```

Solution 3

Ans 3

File name is **wine_final_classification.csv** that contains classification. Best result is produced by using polynomial kernel in support vector machines as the data points are not linearly separable. Also 2nd dimension of wine data that is 'Ash' is ignored as they are overlapping for different categories.

References

[1] "3.3. Model evaluation: quantifying the quality of predictions — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed: 21- Sep- 2017].

[2] "3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/cross_validation.html. [Accessed: 21- Sep- 2017].

[3] "5.6.2. The 20 newsgroups text dataset — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/datasets/twenty_newsgroups.html. [Accessed: 21- Sep- 2017].

[4] "Sample pipeline for text feature extraction and evaluation — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/auto_examples/model_selection/grid_search_text_feature_extraction.html. [Accessed: 21- Sep- 2017].

[5] "Scikit Learn - Feature Extraction", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text. [Accessed: 21- Sep- 2017].

[6] "1.17. Neural network models (supervised) — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/neural_networks_supervised.html. [Accessed: 21- Sep- 2017].