MSDM 5054

Assignment 4

Problem 1

1 and 2 Please check the writing paper PDF (P1Q1Q2.pdf)

3. The parameter $\gamma>0$ in the radial basis function (RBF) kernel affects the bias and variance of the SVM model as follows:

Bias: A small value of γ results in a smoother decision boundary, leading to higher bias. This means the model may underfit the data.

Variance: A large value of γ creates a more complex decision boundary that can fit the training data very closely, which increases variance and may lead to overfitting.

In summary, Low γ: High bias, low variance. High γ: Low bias, high variance.

4. PCA typically requires complete datasets.

a. Imputation: Fill in missing values using mean imputation, median imputation, or more complex models like K-nearest neighbors.

b. Complete Case Analysis: Remove any observations with missing data before applying PCA.

c. Expectation-Maximization (EM): Use the EM algorithm to estimate missing values iteratively.

Best Practices for Preprocessing Before PCA:

a. Standardize the data (mean = 0, variance = 1).

b. Handle missing data appropriately.

c. Remove outliers that may distort the PCA results.

d. Consider dimensionality reduction if the data is high-dimensional.

5. K-means++ is an improved version of K-means. It addresses the issue of poor initial centroid placement.

Advantages over K-means:

a. Better Initialization: K-means++ selects the initial centroids with better clustering results.

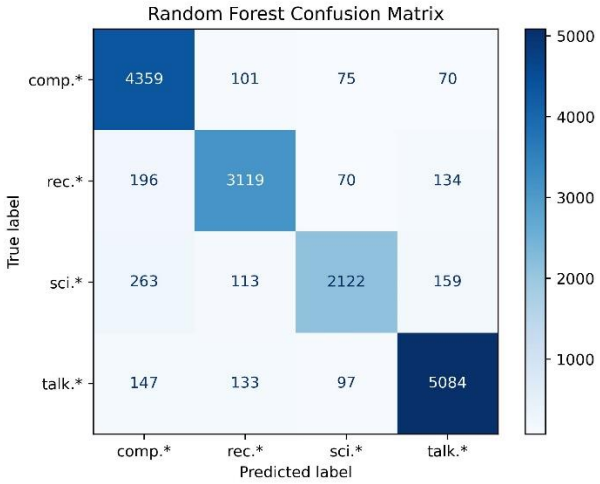
b. Faster Convergence: the initial centroids are selected based on distance from existing centroids, K-means++ often converges faster than K-means, thus saving the cost of calculation and leading to a more stable result.

c. Less Sensitivity to Initial Conditions: K-means++ reduces the variance in clustering. In K-means, however, due to random initialization it often occurs.

Problem 2

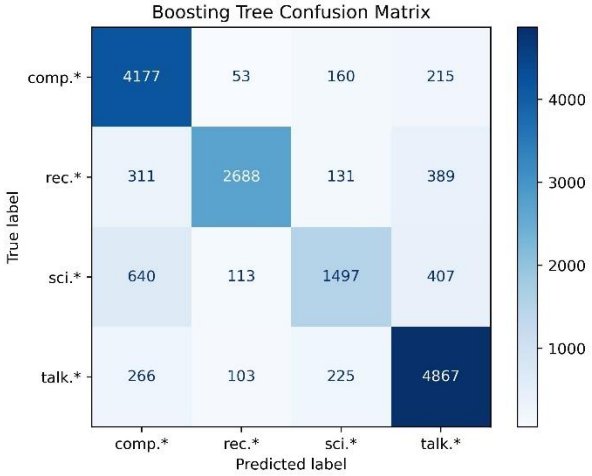
1.

```
----------Random Forest---------
Tuning parameters:

{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None,
'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'n_estimators': 150, 'n_jobs': -1,
'oob_score': False, 'random_state': 14, 'verbose': 0, 'warm_start': False}
Best CV Error: 0.24627096325205033
Top 10 Important Keywords: ['windows', 'god', 'christian', 'car', 'government', 'team', 'jews', 'graphics', 'space', 'religion']
```



2.

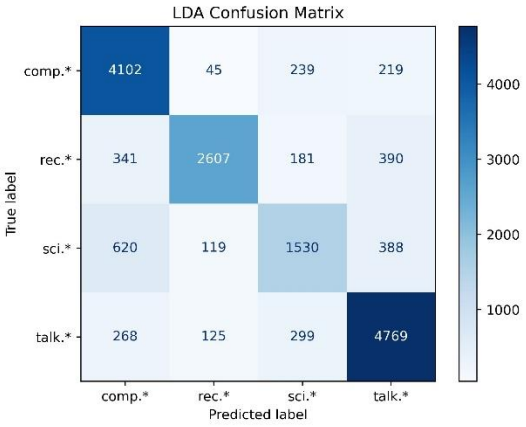
```
---------boosting tree----------
Tuning parameters:

{'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'log_loss', 'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': None,
'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_iter_no_change': None,
'random_state': 14, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}
Best CV Error: 0.2218877312761638
```



3. The selected Random Forest and Boosting Trees models exhibit similar cross-validation error performance. However, the latter employs more tree nodes, making the model relatively more complex compared to the Random Forest.

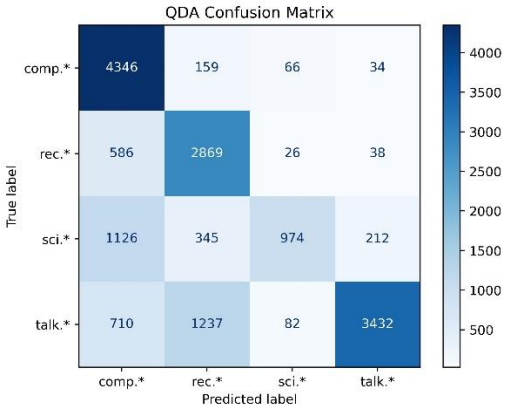
4.

```
---------LDA----------
Best CV Error: 0.24183731409588705
```



5.

```
---------QDA----------
Best CV Error: 0.33363810691277496
```



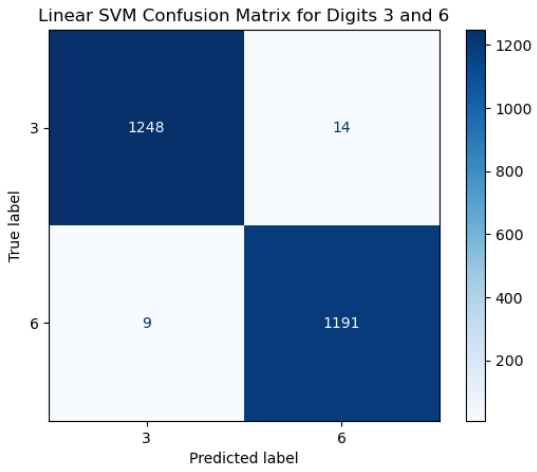
6.

```
-------------------------------------
Random Forest CV Error: 0.24627096325205033
Boosting Tree CV Error: 0.2218877312761638
LDA CV Error: 0.24183731409588705
QDA CV Error: 0.33363810691277496
```

Problem 3

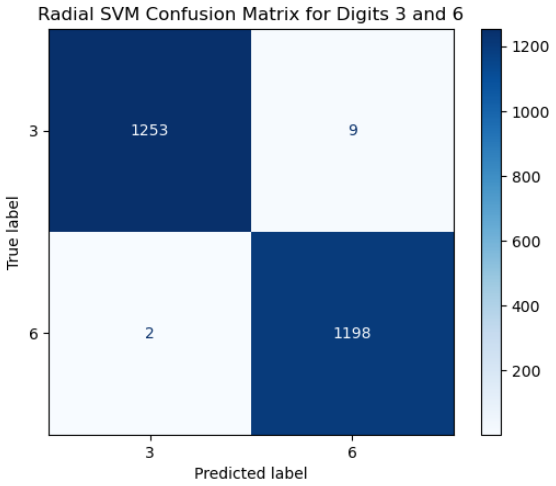
1.

```
-------------Linear SVM for Digits 3 and 6------------
Best Cost Parameter (C): 11
Misclassification Error: 0.009341998375304583
Confusion Matrix:
 [[1248   14]
 [   9 1191]]
Time Cost of Training (seconds): 2.0600426197052
```



2.

```
---------Radial SVM for Digits 3 and 6----------
Best Cost Parameter (C): 10
Best Cost Parameter (gamma): 0.01
Misclassification Error: 0.004467912266450047
Confusion Matrix:
 [[1253    9]
 [   2 1198]]
Time Cost of Training (seconds): 29.77127504348755
```

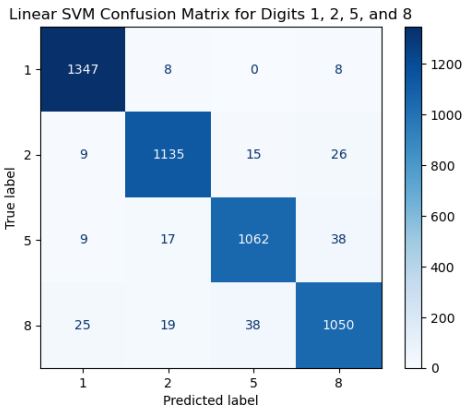


3.SVM with linear kernel has smaller classification error rate and its time cost of training is significantly less than the SVM model using radial SVM. In the binary classification problem here, it would be smarter to adopt the linear kernel due to its convenience for computing.

```
--------Comparison for Digits 3 and 6--------
Linear SVM Error: 0.009341998375304583
Radial SVM Error: 0.004467912266450047
```

4.

```
-------Linear SVM for Digits 1, 2, 5, and 8-------
Best CV Parameters: {'C': 1}

Misclassification Error: 0.044111527259759467
Confusion Matrix:
 [[1347    8    0    8]
 [   9 1135   15   26]
 [   9   17 1062   38]
 [  25   19   38 1050]]
Time Cost of Training (seconds): 954.9446861743927
```



5.

```
-------Multi-class SVM for All Digits----------
Best CV Parameters: {'C': 10, 'gamma': 0.01}
Misclassification Error: 0.036
Confusion Matrix:
 [[1124    0    6    1    0    3    5    0    1    0]
 [   0 1342   12    1    1    0    1    2    3    1]
 [   3    1 1155    6    5    1    0    8    5    1]
 [   0    2   38 1197    0    7    0    6    6    6]
 [   2    2   18    0 1134    2    3    3    0   11]
 [   3    2    8   19    1 1068   10    2   11    2]
 [   2    0   17    0    4    7 1167    0    3    0]
 [   1    2   24    1    6    1    0 1208    0   21]
 [   4    3    9   11    6    7    1    5 1077    9]
 [   4    2   19    3   12    2    0   12    3 1096]]
Time Cost of Training (seconds): 9527.876990795135
```