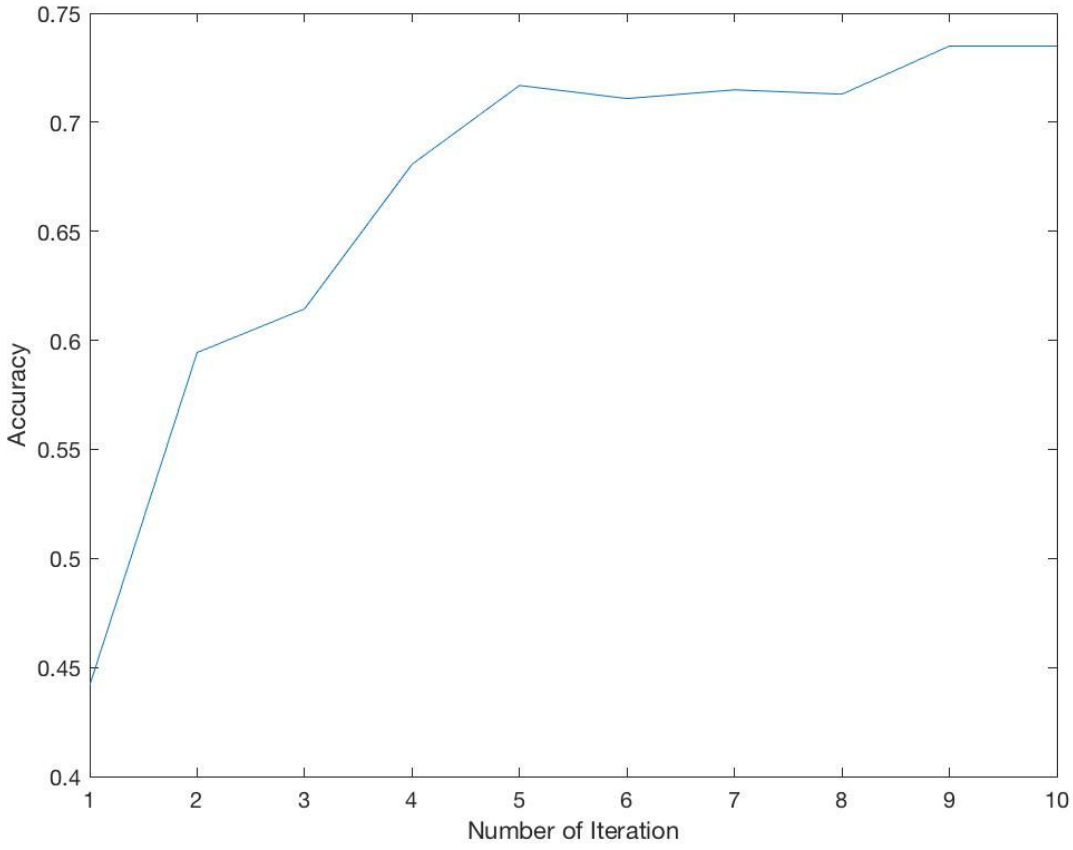
Assignment 5 Report

KNN Classifier:
When choosing the K nearest points, sometimes there will be equal distance between the test point and training points. We solved this problem by picking the first one that the algorithm visited, since it doesn't make that much difference practically. Ideally, when equal distance happened, we can increase K temporally by the number of occurrence of equal distance, but for efficiency and simplicity, we just pick the first one as the estimated class.

Cross Validation:
We ran myCrossValidation() for all features, and the best feature varies each time we run. We have got 9 as the best feature depending on different runs.

Feature selection:
The best set of features we obtained is [9,6,3,2,4,7,1,5,8].
The plot of accuracy vs number of iteration is shown below.



Evaluation:
The accuracy for K = 1 is 0.752, K = 3 is 0.73, and K = 7 is 0.74.
The confusion matrix for K = 1 is:

86	2	0	9	0
7	65	14	12	9
0	12	75	4	5
6	15	5	69	5
1	6	6	6	81

The confusion matrix for K = 3 is:

86	2	1	15	0
6	70	17	18	12
0	7	74	4	6
7	14	3	57	4
1	7	5	6	78

The confusion matrix for K = 7 is:

83	1	0	12	0
9	73	15	17	14
0	3	75	4	3
7	13	3	59	3
1	10	7	8	80

Looking at the confusion matrices, we noticed that classical has the highest value, meaning the classification for this genre is the most accurate. Metal has the second highest value. Among the 5 genres, jazz is most likely confused with other genres. Jazz gets more easily confused with country. This kind of makes sense on a higher level because some of the songs in the two genres share some similar high level features, such as chord progression.

From the confusion matrices and accuracies reported above, we do not notice a huge different when using different K's. K = 1 has the highest accuracy, and K = 2 has the lowest accuracy (although the difference is not large).

K-Means:
We have obtained accuracy of 0.616 using K = 5 with our implementation of K-Means Clustering. Genre may be related to music similarity but not necessarily. The genre is obtained by the ground truth, and K-Means tries to model genre by looking at music similarity. This explains why the result from k-means is better than guessing but still can not achieve accuracy more than 80%.