Diandra Prioleau
26 September 2018
EEL 5840 – Fundamentals of Machine Learning
HW02

1. When training the probabilistic generative classifier, how does the full covariance compare to diagonal covariance in performance for each of the data sets? Why?

When training the probabilistic generative classifier, the full covariance and the diagonal covariance have close classification accuracies, but the accuracy is lower with the full covariance for the 2D data. Similarly, the classification accuracy is the same for the 7D data. For both the 2D and 7D data set, their accuracy scores resulting when using the full covariance and diagonal are close because the number features and the number of training samples has passed the optimal of features and the number of correlated features is not too high.

In contrast, for the hyperspectral data, the classification accuracy for the full covariance is higher than the diagonal covariance, which indicates that the features in the data are not fully independent of each other. If there are too many correlated features, the covariance matrix can become ill-conditioned, which is what occurred in my case. As a result, I regularized the covariance by adding a small constant to the diagonal of the covariance.

In comparison to the 2D and 7D data sets, the classification accuracy scores for the hyperspectral data set are lower when using the probabilistic generative classifier. This may be due to the dimensionality. As the dimensionality increases, the performance increase until an optimal number of features is reached. Therefore, it may be possible that the number of optimal number of features has been reached for the hyperspectral data set, which means the number of training samples would need to increase to improve performance.

2. When training KNN classifier, what happens as you vary k from small to large? Why?

When training KNN classifier, only odd numbers of k are used to avoid ties. As k varies from small to large, the classification accuracy decreases because the classifier is beginning to fit less to the training data to be able to classify new data. For a smaller k, the classifier is more likely to overfit. Therefore, when conducting the validation test, it possible for the validation data to have been easily classified because it is more similar to the training data, which may be different for the test data. Therefore, bias is decreased (test data may be closer to the training data) and variance is higher for the smaller k. However, as k gets larger, each data point in the test or validation data is being compared to more data points in the training data which decreases the likelihood of choosing a class label arbitrarily but rather dependent on the largest number of training data points in a class that the test data point is close to. This results in higher bias (test data may be farther from training data) and lower variance (less fitted to the training data).

3. Determine which classifier(s) you would use for each data set and give an explanation of your reasoning. Hint: This should incorporate some discussion based on results from cross-validation.

After using cross validation, I decided to use the Probabilistic Generative Classifier for the 2D data set since the average of the classification accuracy across the folds is higher than those for the KNN classifier for most of the k parameters. Since the variances among the accuracy scores for both the full covariance and diagonal covariance are small and can be said equal to 0, the diagonal covariance was chosen to be used with the Probabilistic Generative Classifier since it had a higher classification accuracy. The accuracy scores and the corresponding variances are shown in Figure 1 for the cross validation for the Probabilistic Generative classifier and the KNN classifier for varying parameters of k.

```
Probabilistic Generative Classifier Accuracy Mean:  0.975
PG Variance of Accuracy:  5.90000000000001e-05
Probabilistic Generative Classifier Accuracy Mean with Diagonal Cov:  0.98
PGDiag Variance of Accuracy:  5.60000000000001e-05
Accuracy Mean & Variance of Accuracy for each k of KNN
Accuracy Mean for k =  1:  0.97
Variance of Accuracy Mean for k =  1:  0.00010000000000000018
Accuracy Mean for k =  3:  0.97
Variance of Accuracy Mean for k =  3:  2.0000000000000032e-05
Accuracy Mean for k =  5:  0.963
Variance of Accuracy Mean for k =  5:  0.00012300000000000023
Accuracy Mean for k =  7:  0.966
Variance of Accuracy Mean for k =  7:  6.800000000000012e-05
Accuracy Mean for k =  9:  0.967
Variance of Accuracy Mean for k =  9:  0.000546999999999999
Accuracy Mean for k =  11:  0.965
Variance of Accuracy Mean for k =  11:  0.00025100000000000046
Accuracy Mean for k =  13:  0.965
Variance of Accuracy Mean for k =  13:  0.00013100000000000023
Accuracy Mean for k =  15:  0.964
Variance of Accuracy Mean for k =  15:  4.000000000000008e-05
Accuracy Mean for k =  17:  0.951
Variance of Accuracy Mean for k =  17:  4.3000000000000076e-05
Accuracy Mean for k =  19:  0.96
Variance of Accuracy Mean for k =  19:  0.00045599999999999905
Accuracy Mean for k =  21:  0.961
Variance of Accuracy Mean for k =  21:  0.00013900000000000026
Accuracy Mean for k =  23:  0.956
Variance of Accuracy Mean for k =  23:  2.4000000000000045e-05
Accuracy Mean for k =  25:  0.954
Variance of Accuracy Mean for k =  25:  0.00016399999999999927
Accuracy Mean for k =  27:  0.954
Variance of Accuracy Mean for k =  27:  0.0002280000000000004
Accuracy Mean for k =  29:  0.954
Variance of Accuracy Mean for k =  29:  0.0001080000000000002
Accuracy Mean for k =  31:  0.945
Variance of Accuracy Mean for k =  31:  0.00029899999999999865
Accuracy Mean for k =  33:  0.947
Variance of Accuracy Mean for k =  33:  0.00013099999999999942
Accuracy Mean for k =  35:  0.95
Variance of Accuracy Mean for k =  35:  0.00013999999999999926
Accuracy Mean for k =  37:  0.944
Variance of Accuracy Mean for k =  37:  0.00011999999999999933
Accuracy Mean for k =  39:  0.944
Variance of Accuracy Mean for k =  39:  7.199999999999946e-05
```

Figure 1 Accuracy Scores for Probabilistic Generative Classifier and KNN Classifier using 4-fold Cross Validation for 2D Data Set

I decided to use the Probabilistic Generative Classifier for the 7D data set since the classification accuracy is 100%, which is higher than those for the KNN classifier for all k parameters. It seemed skeptical having a 100% classification accuracy, which is shown in Figure 2, when running a 4-fold cross validation on the 7D using Probabilistic Generative Classifier; therefore, it was re-ran using different numbers of fold for the cross validation; however, the classification accuracy was 100% for each of these cross validations. As a result, the Probabilistic Generative Classifier was chosen. When using the full covariance and diagonal covariance, the classification accuracy was 100% for both of them. As a result, the diagonal covariance was chosen for the test data set since the diagonal covariance indicates that features are independent.

```
Probabilistic Generative Classifier Accuracy Mean:  1.0
PG Variance of Accuracy:  0.0
Probabilistic Generative Classifier Accuracy Mean with Diagonal Cov:  1.0
PGDiag Variance of Accuracy:  0.0
Accuracy Mean & Variance of Accuracy for each k of KNN
Accuracy Mean for k =  1:  0.996
Variance of Accuracy Mean for k =  1:  0.0
Accuracy Mean for k =  3:  0.995
Variance of Accuracy Mean for k =  3:  1.100000000000002e-05
Accuracy Mean for k =  5:  0.991
Variance of Accuracy Mean for k =  5:  3.5000000000000065e-05
Accuracy Mean for k =  7:  0.991
Variance of Accuracy Mean for k =  7:  1.900000000000003e-05
Accuracy Mean for k =  9:  0.992
Variance of Accuracy Mean for k =  9:  5.60000000000001e-05
Accuracy Mean for k =  11:  0.99
Variance of Accuracy Mean for k =  11:  3.600000000000007e-05
Accuracy Mean for k =  13:  0.988
Variance of Accuracy Mean for k =  13:  4.000000000000008e-05
Accuracy Mean for k =  15:  0.989
Variance of Accuracy Mean for k =  15:  2.7000000000000046e-05
Accuracy Mean for k =  17:  0.987
Variance of Accuracy Mean for k =  17:  3.5000000000000065e-05
Accuracy Mean for k =  19:  0.988
Variance of Accuracy Mean for k =  19:  5.60000000000001e-05
Accuracy Mean for k =  21:  0.984
Variance of Accuracy Mean for k =  21:  1.600000000000003e-05
Accuracy Mean for k =  23:  0.983
Variance of Accuracy Mean for k =  23:  1.100000000000002e-05
Accuracy Mean for k =  25:  0.984
Variance of Accuracy Mean for k =  25:  2.4000000000000048e-05
Accuracy Mean for k =  27:  0.984
Variance of Accuracy Mean for k =  27:  1.600000000000003e-05
Accuracy Mean for k =  29:  0.98
Variance of Accuracy Mean for k =  29:  0.00024800000000000044
Accuracy Mean for k =  31:  0.983
Variance of Accuracy Mean for k =  31:  8.300000000000015e-05
Accuracy Mean for k =  33:  0.982
Variance of Accuracy Mean for k =  33:  5.20000000000001e-05
Accuracy Mean for k =  35:  0.983
Variance of Accuracy Mean for k =  35:  0.0001070000000000002
Accuracy Mean for k =  37:  0.981
Variance of Accuracy Mean for k =  37:  0.00013100000000000023
Accuracy Mean for k =  39:  0.982
Variance of Accuracy Mean for k =  39:  0.00014800000000000026
```

Figure 2 Accuracy Scores for Probabilistic Generative Classifier and KNN Classifier using 4-fold Cross Validation for 7D Data Set

After using the cross validation, I have decided to use the KNN Classifier with k=13 for the hyperspectral data since the performance when k=13 was higher among the varying parameters of k and higher than the classification accuracy from probabilistic generative classifier, as shown in Figure 3 from the outputs. In addition, the variance of the classification accuracy scores for the 4-fold cross validation when using the KNN classifier with a k of 13 was low and can be said to be 0.

```
Probabilistic Generative Classifier Accuracy Mean:  0.81
PG Variance of Accuracy:  0.000179999999999999
Probabilistic Generative Classifier Accuracy Mean with Diagonal Cov:  0.7290000000000001
PGDiag Variance of Accuracy:  0.0004430000000000008
Accuracy Mean & Variance of Accuracy for each k of KNN
Accuracy Mean for k =  1:  0.786
Variance of Accuracy Mean for k =  1:  0.0001160000000000002
Accuracy Mean for k =  3:  0.8009999999999999
Variance of Accuracy Mean for k =  3:  0.00010699999999999936
Accuracy Mean for k =  5:  0.813
Variance of Accuracy Mean for k =  5:  0.00024299999999999875
Accuracy Mean for k =  7:  0.829
Variance of Accuracy Mean for k =  7:  0.0002829999999999989
Accuracy Mean for k =  9:  0.82
Variance of Accuracy Mean for k =  9:  0.0003599999999999989
Accuracy Mean for k =  11:  0.823
Variance of Accuracy Mean for k =  11:  0.00019499999999999907
Accuracy Mean for k =  13:  0.833
Variance of Accuracy Mean for k =  13:  0.00025100000000000046
Accuracy Mean for k =  15:  0.828
Variance of Accuracy Mean for k =  15:  0.0003039999999999994
Accuracy Mean for k =  17:  0.831
Variance of Accuracy Mean for k =  17:  1.900000000000003e-05
Accuracy Mean for k =  19:  0.822
Variance of Accuracy Mean for k =  19:  0.00039599999999999884
Accuracy Mean for k =  21:  0.819
Variance of Accuracy Mean for k =  21:  0.0006349999999999985
Accuracy Mean for k =  23:  0.8180000000000001
Variance of Accuracy Mean for k =  23:  0.0005239999999999985
Accuracy Mean for k =  25:  0.8280000000000001
Variance of Accuracy Mean for k =  25:  0.00020799999999999926
Accuracy Mean for k =  27:  0.819
Variance of Accuracy Mean for k =  27:  0.00014699999999999905
Accuracy Mean for k =  29:  0.826
Variance of Accuracy Mean for k =  29:  0.0006119999999999989
Accuracy Mean for k =  31:  0.83
Variance of Accuracy Mean for k =  31:  0.001027999999999999
Accuracy Mean for k =  33:  0.8260000000000001
Variance of Accuracy Mean for k =  33:  0.0008039999999999983
Accuracy Mean for k =  35:  0.816
Variance of Accuracy Mean for k =  35:  0.00019999999999999882
Accuracy Mean for k =  37:  0.8280000000000001
Variance of Accuracy Mean for k =  37:  0.001063999999999999
Accuracy Mean for k =  39:  0.826
Variance of Accuracy Mean for k =  39:  0.0006119999999999987
```

Figure 3 Accuracy Scores for Probabilistic Generative Classifier and KNN Classifier using 4-fold Cross Validation for Hyperspectral Data Set