HW 2

Casey Ranft
10/21/20

Intro:

I used five different datasets and applied five different classifiers with their default settings then again with their optimized parameters.

The goal was to find which classifier produced the best results for each dataset. Results were determined using Repeated K Fold and Cross validation scores which evaluate an algorithm and returns the mean accuracy score for each classifier.

The results are broken down by each data set. Determining the ranking of the classifiers with default settings and the ranking of the classifiers with optimized settings. The ultimate result was which classifier and parameters produced the best results.

The classifiers being used for each dataset are Gaussian Naïve Bayes, Support Vector Classification, Decision Tree, K Nearest Neighbor, and AdaBoost.

Glass Data Classification

The data has 213 entries total with nine features and seven classes of glass. The classifiers will look at the features to determine what type of glass it is. Each classifier was used on the dataset and produced the accuracy of classifier as the mean and standard deviation. To compare the results with statistical significance, the difference between each classifier was taken.

Default settings:

```
Gaussian Naive Bayes Accuracy: 0.46 (+/- 0.25)
Support Vector Machine Accuracy: 0.34 (+/- 0.19)
Decision Tree Classifier Accuracy: 0.68 (+/- 0.21)
K Nearest Neighbors Accuracy: 0.66 (+/- 0.20)
AdaBoost Accuracy: 0.44 (+/- 0.24)
```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: 0.12 (+/- 0.30) = [-.18, .42]
GNB vs DTC -> Accuracy: -0.22 (+/- 0.30) = [-.52, .08]
GNB vs KNN -> Accuracy: -0.20 (+/- 0.34) = [-.54, .14]
GNB vs ADA -> Accuracy: 0.02 (+/- 0.34) = [-.32, .36]
SVM vs DTC -> Accuracy: -0.34 (+/- 0.30) = [-.64, -.04]
SVM vs KNN -> Accuracy: -0.31 (+/- 0.28) = [-.59, -.03]
SVM vs ADA -> Accuracy: -0.10 (+/- 0.30) = [-.4, .2]
DTC vs KNN -> Accuracy: 0.02 (+/- 0.32) = [-.3, .34]
DTC vs ADA -> Accuracy: 0.24 (+/- 0.35) = [-.11, .59]
KNN vs ADA -> Accuracy: 0.22 (+/- 0.34) = [-.12, .56]
```

The differences above were used to determine the statistical significance of each result and rank the classifiers. The confidence intervals are shown to the right of each difference calculation and are only significant if the interval is all negative or all positive. This is shown in the red confidence interval between SVM, DTC, and KNN being all negative. What this means is the Decision Tree Classifier and the K Nearest

Neighbor Classifier are more accurate than the Support Vector Machine. All other classifiers are equivalent.

Ranking:

1. GNB, DTC, KNN, ADA
2. SVM

Optimized Parameters:

The optimization for each classifier is listed above the new outputted accuracy. The reason these optimizations were chosen and what they do isn't specified but can be looked up on the sklearn website. For this step, only the effects of the parameters will be looked at.

- ```
  gnb = GaussianNB(var_smoothing=1.8)
  Gaussian Naive Bayes Accuracy: 0.49 (+/- 0.20)
  ```

- ```
  clf = svm.SVC(C = 150, kernel='linear')
  Support Vector Machine Accuracy: 0.66 (+/- 0.18)
  ```

- ```
  dtc = DecisionTreeClassifier(max_depth = 10, random_state = 150)
  Decision Tree Classifier Accuracy: 0.68 (+/- 0.19)
  ```

- ```
  knn = KNeighborsClassifier(n_neighbors=3, weights='distance')
  K Nearest Neighbors Accuracy: 0.70 (+/- 0.18)
  ```

- ```
  ada = AdaBoostClassifier(n_estimators = 30)
  AdaBoost Accuracy: 0.45 (+/- 0.23)
  ```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: -0.17 (+/- 0.24) = [-.41, .07]
GNB vs DTC -> Accuracy: -0.19 (+/- 0.27) = [-.46, .08]
GNB vs KNN -> Accuracy: -0.22 (+/- 0.28) = [-.5, .06]
GNB vs ADA -> Accuracy: 0.04 (+/- 0.31) = [-.27, .35]
SVM vs DTC -> Accuracy: -0.01 (+/- 0.27) = [-.28, .26]
SVM vs KNN -> Accuracy: -0.04 (+/- 0.28) = [-.32, .24]
SVM vs ADA -> Accuracy: 0.21 (+/- 0.28) = [-.07, .49]
DTC vs KNN -> Accuracy: -0.03 (+/- 0.27) = [-.3, .24]
DTC vs ADA -> Accuracy: 0.22 (+/- 0.30) = [-.08, .52]
KNN vs ADA -> Accuracy: 0.25 (+/- 0.30) = [-.05, .55]
```

From before we know that the confidence interval either needs to be all positive or all negative and since none of the classifiers above have that they are all equivalent. This means the optimized parameters evened the classifiers out a bit more. Comparing the accuracy obtained with the parameters to the defaults show some classifiers improved by about 30% (SVM) while others only improved a little.

Ranking:

1. GNB, DTC, KNN, SVM, ADA
```

Digits Data Classification

The data has 1797 entries total with 17 features and 11 classes of integers. This data is an 8x8 datapoint of a single digit. The classifiers will try and determine which integer it is between 0 - 10.

Each classifier was used on the dataset and produced the accuracy of classifier as the mean and standard deviation. To compare the results with statistical significance, the difference between each classifier was taken.

Default settings:

```
Gaussian Naive Bayes Accuracy: 0.84 (+/- 0.06)
Decision Tree Classifier Accuracy: 0.85 (+/- 0.05)
K Nearest Neighbors Accuracy: 0.99 (+/- 0.02)
AdaBoost Accuracy: 0.25 (+/- 0.07)
Support Vector Machine Accuracy: 0.99 (+/- 0.02)
```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: -0.15 (+/- 0.06) = [-.21, -.09]
GNB vs DTC -> Accuracy: -0.01 (+/- 0.07) = [-.08, .06]
GNB vs KNN -> Accuracy: -0.15 (+/- 0.06) = [-.21, -.09]
GNB vs ADA -> Accuracy: 0.59 (+/- 0.09) = [.5, 68]
SVM vs DTC -> Accuracy: 0.13 (+/- 0.05) = [.08, .18]
SVM vs KNN -> Accuracy: 0.00 (+/- 0.02) = [-.02, .02]
SVM vs ADA -> Accuracy: 0.73 (+/- 0.07) = [.66, .8]
DTC vs KNN -> Accuracy: -0.13 (+/- 0.05) = [-.18, -.08]
DTC vs ADA -> Accuracy: 0.60 (+/- 0.09) = [.51, .69]
KNN vs ADA -> Accuracy: 0.73 (+/- 0.07) = [.66, .8]
```

The differences above were used to determine the statistical significance of each result and rank the classifiers. The confidence intervals are shown to the right of each difference calculation and are only significant if the interval is all negative or all positive. This is shown in the red intervals and if the interval is all positive then the classifier on the left is better and if it's all negative then the right classifier is better. All other classifiers are equivalent.

Ranking:

1. SVM, KNN
2. GNB, DTC
3. ADA

Optimized Parameters:

The optimization for each classifier is listed above the new accuracy. The reason these optimizations were chosen and what they do isn't specified but can be looked up on the sklearn website. For this step, only the effects of the parameters will be looked at.

- `gnb = GaussianNB(var_smoothing=2)`

```
Gaussian Naive Bayes Accuracy: 0.90 (+/- 0.04)
```

- ```
  dtc = DecisionTreeClassifier( criterion = 'entropy', max_depth = 10, ra
  ndom_state = 150)
  Decision Tree Classifier Accuracy: 0.87 (+/- 0.05)
  ```

- ```
  knn = KNeighborsClassifier(n_neighbors=3, weights='distance')
  K Nearest Neighbors Accuracy: 0.99 (+/- 0.02)
  ```

- ```
  ada = AdaBoostClassifier(n_estimators = 30)
  AdaBoost Accuracy: 0.26 (+/- 0.06)
  ```

- ```
  clf = svm.SVC(C = 150, kernel='poly')
  Support Vector Machine Accuracy: 0.99 (+/- 0.01)
  ```

Differences between Classifiers:

```
GNB vs SVM -> Accuracy: -0.09 (+/- 0.04) = [-.13, -.05]
GNB vs DTC -> Accuracy: 0.03 (+/- 0.06) = [-.03, .09]
GNB vs KNN -> Accuracy: -0.09 (+/- 0.04) = [-.13, -.05]
GNB vs ADA -> Accuracy: 0.64 (+/- 0.07) = [.57, .71]
SVM vs DTC -> Accuracy: 0.12 (+/- 0.05) = [.07, .17]
SVM vs KNN -> Accuracy: 0.00 (+/- 0.02) = [-.02, .02]
SVM vs ADA -> Accuracy: 0.73 (+/- 0.06) = [.67, .79]
DTC vs KNN -> Accuracy: -0.11 (+/- 0.05) = [-.16, -.06]
DTC vs ADA -> Accuracy: 0.62 (+/- 0.07) = [.55, .69]
KNN vs ADA -> Accuracy: 0.73 (+/- 0.06) = [.67, .79]
```

From before we know that the confidence interval either needs to be all positive or all negative as shown with the red intervals.

In this case, the optimized parameters didn't really close the gap between classifiers. This is because some of the classifiers already produced a 99% accuracy with a small deviation with the default settings (KNN, SVM) meaning there was virtually no improvement when changing the parameters. In fact, some parameters lowered the result when testing.

Overall, the ranking of classifiers didn't change with optimized parameters meaning some classifiers are simply better fit to handle this dataset.

Ranking:

1. SVM, KNN
2. GNB, DTC
3. ADA

Breast Cancer Data Classification

The data has 569 entries total with 30 features and two class (malignant or benign).

Each classifier was used on the dataset and produced the accuracy of classifier as the mean and standard deviation. To compare the results with statistical significance, the difference between each classifier was taken.

Default settings:

```
Gaussian Naive Bayes Accuracy: 0.94 (+/- 0.07)
Decision Tree Classifier Accuracy: 0.92 (+/- 0.08)
K Nearest Neighbors Accuracy: 0.93 (+/- 0.07)
AdaBoost Accuracy: 0.96 (+/- 0.06)
Support Vector Machine Accuracy: 0.92 (+/- 0.07)
```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: 0.02 (+/- 0.10) = [-.08, .12]
GNB vs DTC -> Accuracy: 0.02 (+/- 0.11) = [-.09, .13]
GNB vs KNN -> Accuracy: 0.01 (+/- 0.10) = [-.09, .11]
GNB vs ADA -> Accuracy: -0.02 (+/- 0.08) = [-.1, .06]
SVM vs DTC -> Accuracy: -0.01 (+/- 0.10) = [-.11, .09]
SVM vs KNN -> Accuracy: -0.01 (+/- 0.11) = [-.12, .1]
SVM vs ADA -> Accuracy: -0.04 (+/- 0.09) = [-.13, .05]
DTC vs KNN -> Accuracy: -0.01 (+/- 0.11) = [-.12, .1]
DTC vs ADA -> Accuracy: -0.04 (+/- 0.10) = [-.14, .06]
KNN vs ADA -> Accuracy: -0.03 (+/- 0.09) = [-.12, .06]
```

The differences above were used to determine the statistical significance of each result and rank the classifiers. The confidence intervals are shown to the right of each difference calculation and are only significant if the interval is all negative or all positive. In this case, none of the intervals are all positive or all negative meaning the classifiers are equivalent.

Ranking:

1. GNB, DTC, SVM, ADA, KNN

Optimized Parameters:

The optimization for each classifier is listed above the new accuracy. The reason these optimizations were chosen and what they do isn't specified but can be looked up on the sklearn website. For this step, only the effects of the parameters will be looked at.

- `gnb = GaussianNB(var_smoothing=.2)`
  `Gaussian Naive Bayes Accuracy: 0.87 (+/- 0.10)`

- `dtc = DecisionTreeClassifier(criterion = 'entropy', max_depth = 11, ran`
  `dom_state = 150)`

```
Decision Tree Classifier Accuracy: 0.94 (+/- 0.07)
```

- ```
  knn = KNeighborsClassifier(n_neighbors = 4)
  K Nearest Neighbors Accuracy: 0.93 (+/- 0.06)
  ```

- ```
  ada = AdaBoostClassifier(n_estimators = 100, random_state= 150)
  AdaBoost Accuracy: 0.97 (+/- 0.05)
  ```

- ```
  clf = svm.SVC(C = 150, kernel='poly')
  Support Vector Machine Accuracy: 0.93 (+/- 0.07)
  ```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: -0.05 (+/- 0.12) = [-.17, .07]
GNB vs DTC -> Accuracy: -0.06 (+/- 0.13) = [-.19, .07]
GNB vs KNN -> Accuracy: -0.06 (+/- 0.12) = [-.18, .06]
GNB vs ADA -> Accuracy: -0.10 (+/- 0.11) = [-.21, .01]
SVM vs DTC -> Accuracy: -0.01 (+/- 0.10) = [-.11, .09]
SVM vs KNN -> Accuracy: -0.00 (+/- 0.10) = [-.1, .1]
SVM vs ADA -> Accuracy: -0.04 (+/- 0.08) = [-.12, .04]
DTC vs KNN -> Accuracy: 0.01 (+/- 0.08) = [-.07, .09]
DTC vs ADA -> Accuracy: -0.03 (+/- 0.08) = [-.11, .05]
KNN vs ADA -> Accuracy: -0.04 (+/- 0.08) = [-.12, .04]
```

From before we know that the confidence interval either needs to be all positive or all negative and since none of the classifiers above have that they are all equivalent. This is the same outcome from the default setting meaning the optimization didn't add enough of an improvement to make one classifier better than another.

Ranking:

1. GNB, KNN, DTC, ADA, SVM

Wine Data Classification

The data has 178 entries total with 13 features and three classes for identifying wine.

Each classifier was used on the dataset and produced the accuracy of classifier as the mean and standard deviation. To compare the results with statistical significance, the difference between each classifier was taken.

Default settings:

```
Gaussian Naive Bayes Accuracy: 0.97 (+/- 0.08)
Decision Tree Classifier Accuracy: 0.91 (+/- 0.13)
K Nearest Neighbors Accuracy: 0.69 (+/- 0.22)
Support Vector Machine Accuracy: 0.65 (+/- 0.20)
AdaBoost Accuracy: 0.89 (+/- 0.20)
```
```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: 0.32 (+/- 0.19) = [.13, .51]
GNB vs DTC -> Accuracy: 0.07 (+/- 0.16) = [-.09, .23]
GNB vs KNN -> Accuracy: 0.28 (+/- 0.24) = [.04, .52]
GNB vs ADA -> Accuracy: 0.08 (+/- 0.22) = [-.14, .3]
SVM vs DTC -> Accuracy: -0.25 (+/- 0.23) = [-.48, -.02]
SVM vs KNN -> Accuracy: -0.04 (+/- 0.28) = [-.32, .24]
SVM vs ADA -> Accuracy: -0.24 (+/- 0.26) = [-.5, .02]
DTC vs KNN -> Accuracy: 0.21 (+/- 0.26) = [-.05, .47]
DTC vs ADA -> Accuracy: 0.02 (+/- 0.22) = [-.2, .24]
KNN vs ADA -> Accuracy: -0.20 (+/- 0.33) = [-.53, .13]
```

The differences above were used to determine the statistical significance of each result and rank the classifiers. The confidence intervals are shown to the right of each difference calculation and are only significant if the interval is all negative or all positive. This is shown in the red intervals and if the interval is all positive then the classifier on the left is better and if it's all negative then the right classifier is better. All other classifiers are equivalent. In this case, SVM is better than GNB and DTC. KNN is also better than GNB. The ADA classifier is equivalent to all of them though so just put it at both ranking

Ranking:

1. SVM, KNN, ADA
2. GNB, DTC, ADA

Optimized Parameters:

The optimization for each classifier is listed above the new accuracy. The reason these optimizations were chosen and what they do isn't specified but can be looked up on the sklearn website. For this step, only the effects of the parameters will be looked at.

- `gnb = GaussianNB(var_smoothing=1e-009)`
  `Gaussian Naive Bayes Accuracy: 0.98 (+/- 0.07)`

- `dtc = DecisionTreeClassifier(criterion = 'entropy', random_state = 100)`
  `Decision Tree Classifier Accuracy: 0.95 (+/- 0.12)`

- `knn = KNeighborsClassifier(n_neighbors=2, weights='distance')`
  `K Nearest Neighbors Accuracy: 0.77 (+/- 0.19)`

- `clf = svm.SVC(C = 150, kernel='linear')`
  `Support Vector Machine Accuracy: 0.96 (+/- 0.08)`

- `ada = AdaBoostClassifier(n_estimators = 100, random_state= 50)`
  `AdaBoost Accuracy: 0.91 (+/- 0.13)`

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: 0.01 (+/- 0.11) = [-.1, .12]
GNB vs DTC -> Accuracy: 0.03 (+/- 0.14) = [-.11, .17]
GNB vs KNN -> Accuracy: 0.21 (+/- 0.21) = [0, .42]
GNB vs ADA -> Accuracy: 0.07 (+/- 0.16) = [-.09, .23]
SVM vs DTC -> Accuracy: 0.01 (+/- 0.16) = [-.15, .17]
SVM vs KNN -> Accuracy: 0.20 (+/- 0.22) = [-.02, .42]
SVM vs ADA -> Accuracy: 0.06 (+/- 0.14) = [-.08, .2]
DTC vs KNN -> Accuracy: 0.18 (+/- 0.23) = [-.05, .41]
DTC vs ADA -> Accuracy: 0.04 (+/- 0.18) = [-.14, .22]
KNN vs ADA -> Accuracy: -0.14 (+/- 0.23) = [-.37, .09]
```

From before we know that the confidence interval either needs to be all positive or all negative, shown by the red interval. Comparing these to the default settings shows that most classifiers had improvement and were made equivalent each other. Some even had a 30% improvement in accuracy when changing the parameters to better fit the data (SVM). Almost all showed some level of improvement though, about 5%.

Ranking:

1. GNB, SVM, DTC, ADA
2. KNN

Pima Indian Diabetes Data Classification

The data has 6903 entries total with eight features and two classes for identifying whether Pima Indians females have diabetes.

Each classifier was used on the dataset and produced the accuracy of classifier as the mean and standard deviation. To compare the results with statistical significance, the difference between each classifier was taken.

Default settings:

```
Gaussian Naive Bayes Accuracy: 0.75 (+/- 0.10)
Decision Tree Classifier Accuracy: 0.70 (+/- 0.11)
K Nearest Neighbors Accuracy: 0.72 (+/- 0.10)
Support Vector Machine Accuracy: 0.76 (+/- 0.08)
AdaBoost Accuracy: 0.75 (+/- 0.09)
```

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: -0.01 (+/- 0.13) = [-.14, .12]
GNB vs DTC -> Accuracy: 0.05 (+/- 0.16) = [-.11, .21]
GNB vs KNN -> Accuracy: 0.04 (+/- 0.14) = [-.1, .18]
GNB vs ADA -> Accuracy: 0.00 (+/- 0.13) = [-.13, .13]
SVM vs DTC -> Accuracy: 0.06 (+/- 0.14) = [-.08, .2]
SVM vs KNN -> Accuracy: 0.04 (+/- 0.13) = [-.09, .17]
```

```
SVM vs ADA -> Accuracy: 0.01 (+/- 0.11) = [-.1, .12]
DTC vs KNN -> Accuracy: -0.01 (+/- 0.15) = [-.16, .14]
DTC vs ADA -> Accuracy: -0.05 (+/- 0.13) = [-.18, .08]
KNN vs ADA -> Accuracy: -0.03 (+/- 0.14) = [-.17, .11]
```

The differences above were used to determine the statistical significance of each result and rank the classifiers. The confidence intervals are shown to the right of each difference calculation and are only significant if the interval is all negative or all positive. In this case, none of the intervals are all positive or all negative meaning the classifiers are equivalent.

Ranking:

1. GNB, SVM, DTC, KNN, ADA

Optimized Parameters:

The optimization for each classifier is listed above the new accuracy. The reason these optimizations were chosen and what they do isn't specified but can be looked up on the sklearn website. For this step, only the effects of the parameters will be looked at.

- `gnb = GaussianNB(var_smoothing=1e-009)`
  `Gaussian Naive Bayes Accuracy: 0.76 (+/- 0.10)`

- `dtc = DecisionTreeClassifier(max_depth = 3, random_state = 30)`
  `Decision Tree Classifier Accuracy: 0.74 (+/- 0.10)`

- `knn = KNeighborsClassifier(n_neighbors = 7)`
  `K Nearest Neighbors Accuracy: 0.73 (+/- 0.10)`

- `clf = svm.SVC(kernel = 'poly')`
  `Support Vector Machine Accuracy: 0.76 (+/- 0.10)`

- `ada = AdaBoostClassifier(n_estimators = 30, random_state= 150)`
  `AdaBoost Accuracy: 0.75 (+/- 0.09)`

Difference between Classifiers:

```
GNB vs SVM -> Accuracy: -0.00 (+/- 0.13) = [-.13, .13]
GNB vs DTC -> Accuracy: 0.02 (+/- 0.14) = [-.12, .16]
GNB vs KNN -> Accuracy: 0.03 (+/- 0.15) = [-.12, .18]
GNB vs ADA -> Accuracy: 0.00 (+/- 0.15) = [-.15, .15]
SVM vs DTC -> Accuracy: 0.02 (+/- 0.14) = [-.12, .16]
SVM vs KNN -> Accuracy: 0.03 (+/- 0.14) = [-.11, .17]
SVM vs ADA -> Accuracy: 0.01 (+/- 0.14) = [-.13, .15]
DTC vs KNN -> Accuracy: 0.01 (+/- 0.14) = [-.13, .15]
DTC vs ADA -> Accuracy: -0.02 (+/- 0.14) = [-.16, .12]
KNN vs ADA -> Accuracy: -0.03 (+/- 0.14) = [-.17, .11]
```

From before we know that the confidence interval either needs to be all positive or all negative and since none of the classifiers above have that they are all equivalent. This is the same outcome from the default setting meaning the optimization didn't add enough of an improvement to make one classifier better than another.

I was unable to truly test the SVM parameters because it was taking so long to load that it would crash, or I would cancel the program. Almost all modification attempted failed to produce a result under five minutes. This means the SVM classifier could have been better, but I had no way of testing it.

Ranking:

GNB, DTC, KNN, SVM, ADA

Conclusion:

Overall, determining which is the best classifier depended on the type, size, and quality of the data and less on the classifier itself. It was more about which classifier fit the shape of the data best.

The optimizations didn't have a huge impact unless the default classifier was poorly match to the data to begin with. When the default settings already matched the data to the best of its' ability, the optimizations only improved the algorithm a small amount or not at all.

The biggest impact in parameter optimization was almost always when a different algorithm was applied to fit the data better. For example, SVM has multiple algorithms to fit the shape of the data like linear, ploy, and sigmoid, which were found to improve the accuracy about 30%. The other classifiers didn't really have this option though so they could only improve a little.