CM_HW2_Template

October 8, 2020

```
[ ]: import numpy as np
     import pandas as pd
     import seaborn as sns

     '''
      TIP: 1. Research and import desired feature selection methods from https://
     ↪scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection
          2. For k-fold cross validation, consider https://scikit-learn.org/stable/
     ↪modules/generated/sklearn.model_selection.KFold.html
                                             and/or https://scikit-learn.org/stable/
     ↪modules/generated/sklearn.model_selection.cross_validate.html#sklearn.
     ↪model_selection.cross_validate
          3. Research and import desired classification algorithms from https://
     ↪scikit-learn.org/stable/supervised_learning.html
          4. Research and import desired clustering algorithms from https://
     ↪scikit-learn.org/stable/modules/clustering.html
     '''
```

1 Instructions

For each question, a rough outline has been provided to help you get started under "Part 1.x.x: Work". Feel free to either follow the outline or use your own method for solving the problem. In either case, however, please make sure to include your work in these sections and fill in your answer in the cell titled "Part 1.x.x: Answer".

Embedding Images in the Notebook

To upload an image in a markdown cell in Jupyter Notebook: 1. Go to the menu bar and select Edit -> Insert Image.

2. Select image from your disk and upload.
3. Press Ctrl+Enter or Shift+Enter.

This will make the image as part of the notebook and you don't need to upload it in the directory

Export Jupyter Notebooks
In your local computer, open the notebook you would like to export and navigate to 'File' at the top menu bar. By clickling 'File', you can find 'Download as' in the drop-down menu. Select the

format you want to export the notebook as: either directly as a pdf, or if you download it as an html file, use a website like html2pdf.com to convert it to a pdf file for submission on Gradescope.

Colab does not seem to support exporting their notebooks to other formats, so if you choose to use Colab, you will need to download the notebook as an .ipynb file before following the steps above on your local machine.

2 Question 1

2.0.1 Read Data

[ ]:
```
PATH_TO_Q1_DATA = 'HW2_Q1_DATA.csv' #TODO: Change if your path to data is␣
↪different
df = pd.read_csv(PATH_TO_Q1_DATA)
```

2.1 Part 1.1: Filter-based Feature Selection

2.1.1 Part 1.1.1: Work

[ ]:
```
'''
TODO: Apply a filter-based feature selection method of your choice using␣
↪10-fold cross validation
      and use the results to choose the top 20 features

TIP: Scikit-learn provides implementations of many useful statistical measures.

      Pearson's Correlation Coefficient: f_regression()
      ANOVA: f_classif()
      Chi-Squared: chi2()
      Mutual Information: mutual_info_classif() and mutual_info_regression()

     Also, SciPy provides implementations of many more statistics, such as
     Kendall's tau (kendalltau) and Spearman's rank correlation (spearmanr).
'''
```

2.1.2 Part 1.1.1: Answer

List the top 20 features you found: **YOUR ANSWER HERE**

2.1.3 Part 1.1.2: Work

[ ]:
```
'''
TODO: Create and plot a 20 x 20 correlation heat map using your features from␣
↪part 1.1.1

TIP: 1. Pandas has a correlation functionality for dataframes
        https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.
↪DataFrame.corr.html
```

```
    2. Seaborn has a heatmap functionality
       https://seaborn.pydata.org/generated/seaborn.heatmap.html
'''
```

2.1.4 Part 1.1.2: Answer

Plot the heatmap in one of the above cells or embed it as an image in this cell

What is the average of the values in the heatmap? **YOUR ANSWER HERE**

2.1.5 Part 1.1.3: Work

[ ]:
```
'''
TODO: Train a classifier using your top 1, top 2, ..., top 20 features from␣
 ↪part 1.1.1
     and plot the 10-fold cross-validated accuracy as a function of the number␣
 ↪of features

TIP: 1. scikit-learn has a great collection of classifiers: https://
 ↪scikit-learn.org/stable/auto_examples/classification/
 ↪plot_classifier_comparison.html
    2. scikit-learn also supports different ways of cross-validation: https://
 ↪scikit-learn.org/stable/modules/cross_validation.
 ↪html#cross-validation-iterators
'''
```

2.1.6 Part 1.1.3: Answer

Include the plot as the output of one of the above cells or embed it as an image in this cell

2.1.7 Part 1.1.4: Work

[ ]:
```
'''
TODO: 1. Devise a method for determining a threshold that can be used so␣
 ↪separate
         truly relevant features from those that are spurious
     2. Apply this method to the data to obtain a new set of features
     3. Re-train a classifier with the new features using 10-fold cross␣
 ↪validation
'''
```

2.1.8 Part 1.1.4: Answer

Briefly describe the method you devised for determinining a threshold for truly relevant features: **YOUR EXPLANATION HERE**

List the new set of features obtained from applying your method to the data: **YOUR ANSWER HERE**

What was the 10-fold cross-validated accuracy of the classifier trained with these new features? **YOUR ANSWER HERE**

2.1.9 Part 1.1.5: Work

[ ]:
```
'''
TODO: 1. Use a clustering algorithm of your choice to cluster
         the features you found in part 1.1.4 into 10 clusters
      2. Choose a representative feature from each cluster and
         train a classifier with these features using 10-fold
         cross validation
'''
```

2.1.10 Part 1.1.5: Answer

List the representative features you chose from the 10 clusters: **YOUR ANSWER HERE**

What was the 10-fold cross-validated accuracy of the classifier trained with these representative features? **YOUR ANSWER HERE**

2.2 Part 1.2: Wrapper-based Feature Selection

2.2.1 Part 1.2.1: Work

[ ]:
```
'''
TODO: Apply a wrapper-based feature selection method of your choice to the data

TIP: 1. Scikit learn has an implementation of recursive feature elimination␣
 ↪(RFE)
        https://scikit-learn.org/stable/modules/generated/sklearn.
 ↪feature_selection.RFE.html
     2. The mlxtend library has very thorough documentation and great options␣
 ↪for sequential and exhaustive feature selection
        http://rasbt.github.io/mlxtend/user_guide/feature_selection/
 ↪SequentialFeatureSelector/
        http://rasbt.github.io/mlxtend/user_guide/feature_selection/
 ↪ExhaustiveFeatureSelector/
'''
```

2.2.2 Part 1.2.1: Answer

List the top features selected in at least 8 out of 10 folds: **YOUR ANSWER HERE**

2.2.3 Part 1.2.2: Work

```
[ ]: '''
TODO: Create and plot a 20 x 20 correlation heat map using your features from␣
 ↪part 1.2.1

TIP: 1. Pandas has a correlation functionality for dataframes
        https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.
 ↪DataFrame.corr.html
     2. Seaborn has a heatmap functionality
        https://seaborn.pydata.org/generated/seaborn.heatmap.html
'''
```

2.2.4 Part 1.2.2: Answer

Plot the heatmap in one of the above cells or embed it as an image in this cell

What is the average of the values in the heatmap? **YOUR ANSWER HERE**

2.2.5 Part 1.2.3: Work

```
[ ]: '''
TODO: Train a classifier using your features from part 1.2.1 with 10-fold cross␣
 ↪validation

TIP: 1. scikit-learn has a great collection of classifiers: https://
 ↪scikit-learn.org/stable/auto_examples/classification/
 ↪plot_classifier_comparison.html
     2. scikit-learn also supports different ways of cross-validation: https://
 ↪scikit-learn.org/stable/modules/cross_validation.
 ↪html#cross-validation-iterators
'''
```

2.2.6 Part 1.2.3: Answer

What was the 10-fold cross-validated accuracy of the classifier trained with these features? **YOUR ANSWER HERE**

2.3 Part 1.3: Embedded Feature Selection

2.3.1 Part 1.3.1: Work

```
[ ]: '''
TODO: Apply an embedded feature selection method of your choice using 10-fold␣
 ↪cross validation

TIP: 1. See https://scikit-learn.org/stable/modules/classes.html#module-sklearn.
 ↪tree
```

```
        and/or https://scikit-learn.org/stable/modules/classes.
↪html#module-sklearn.ensemble
        for tree based methods
    2. Check out the SelectFromModel functionality from
        https://scikit-learn.org/stable/modules/generated/sklearn.
↪feature_selection.SelectFromModel.html
'''
```

2.3.2 Part 1.3.1: Answer

List the top features selected in at least 8 out of 10 folds: **YOUR ANSWER HERE**

2.3.3 Part 1.3.2: Work

[ ]:
```
'''
TODO: Create and plot a 20 x 20 correlation heat map using your features from␣
↪part 1.3.1

TIP: 1. Pandas has a correlation functionality for dataframes
        https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.
↪DataFrame.corr.html
    2. Seaborn has a heatmap functionality
        https://seaborn.pydata.org/generated/seaborn.heatmap.html
'''
```

2.3.4 Part 1.3.2: Answer

Plot the heatmap in one of the above cells or embed it as an image in this cell

What is the average of the values in the heatmap? **YOUR ANSWER HERE**

2.3.5 Part 1.3.3: Work

[ ]:
```
'''
TODO: Train a classifier using your features from part 1.3.1 with 10-fold cross␣
↪validation

TIP: 1. scikit-learn has a great collection of classifiers: https://
↪scikit-learn.org/stable/auto_examples/classification/
↪plot_classifier_comparison.html
    2. scikit-learn also supports different ways of cross-validation: https://
↪scikit-learn.org/stable/modules/cross_validation.
↪html#cross-validation-iterators
'''
```

2.3.6 Part 1.3.3: Answer

What was the 10-fold cross-validated accuracy of the classifier trained with these features? **YOUR ANSWER HERE**