**Question 1. Decision Tree Classifier [10 Points]**

Data: The zip file “*hw2.q1.data.zip*” contains 3 CSV files:

* “*hw2.q1.train.csv*” contains 10,000 rows and 26 columns. The first column ‘*y*’ is the output variable with 2 classes: 0, 1. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.
* “*hw2.q1.test.csv*” contains 5,000 rows and 41 columns. The first column ‘*y*’ is the output variable with 2 classes: 0, 1. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.
* “*hw2.q1.new.csv*” contains 30 rows and 26 columns. The first column ‘*ID*’ is an identifier for 30 unlabeled samples. The remaining 25 columns contain input features: *x\_1*, …, x*\_25*.

**Task 1**. [4 points]

Use 5-fold cross-validation with the 10,000 labeled exampled from “*hw2.q1.train.csv*” to determine the fewest number of rules using which a decision tree classifier can achieve mean cross-validation accuracy of at least 0.96. Report the number of rules needed, the cross-validation accuracy obtained, and all the hyper-parameter values for the *DecisionTreeClassifier*.

*Number of rules needed*: ……………….

*Mean cross-validation accuracy: ……………………….* (*rounded to 4 decimal places*)

*Hyper-parameter values for selected DecisionTreeClassifier model:*

**Task 2**. [2 Points]

Train a *DecisionTreeClassifier* with the hyper-parameter values determined in Task 1 on all 10,000 training samples and use it to predict the output class ‘*y*’ for the 2,000 examples in “*hw2.q1.test.csv*”*.* Report the following:

* *Accuracy on 2,000 test examples*: …………………… (rounded to 4 decimal places)
* *Classification report* for the 2,000 test examples:
* Of the 952 test samples that belong to class *y*=1, how many are correctly predicted (according to your classification report)?

**Task 3.** [2 Points]

Use the model trained in Task 2 to predict the output class ‘*y’* for the 30 examples in “*hw2.q1.new.csv*”. Specify the predicted classes in the table below:

|  |  |
| --- | --- |
| ID | predicted y |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 |  |
| 12 |  |
| 13 |  |
| 14 |  |
| 15 |  |
| 16 |  |
| 17 |  |
| 18 |  |
| 19 |  |
| 20 |  |
| 21 |  |
| 22 |  |
| 23 |  |
| 24 |  |
| 25 |  |
| 26 |  |
| 27 |  |
| 28 |  |
| 29 |  |
| 30 |  |

**Task 4.** [2 Points]

Of the 25 input variables which ones are relevant for this classification task?

The following input variables are relevant for this classification task: …………………

Interpret your trained model and specify the rules that can be used to classify the output based on the inputs.

Rules:

*Rule 1.*

*Rule 2.*

*.*

*.*

*Rule k.*

**Question 2. Supervised machine learning classifiers [10 Points]**

Data: The zip file “*hw2.q2.data.zip*” contains 3 CSV files:

* “*hw2.q2.train.csv*” contains 8,000 rows and 11 columns. The first column ‘*y*’ is the output variable with 4 classes: 0, 1, 2, 3. The remaining 10 columns contain input features: *x1*, …, x*10*.
* “*hw2.q2.test.csv*” contains 2,000 rows and 11 columns. The first column ‘*y*’ is the output variable with 4 classes: 0, 1, 2, 3. The remaining 10 columns contain input features: *x1*, …, x*10*.
* “*hw2.q1.new.csv*” contains 30 rows and 10 columns. The first column ‘*ID*’ is an identifier for 30 unlabeled samples. The remaining 10 columns contain input features: *x1*, …, x*10*.

**Task 1**. [6 points]

Use 4-fold cross-validation with the 8,000 labeled exampled from “*hw2.q2.train.csv*” to identify a classifier that achieves mean cross-validation accuracy of at least 0.96. You should try several *Scikit-Learn* classifiers, including: *GaussianNB, DecisionTreeClassifier, RandomForestClassifier, ExtraTreesClassifier, KNeighborsClassifier, LogisticRegression, SVC, and MLPClassifier*. Try different hyper-parameter values for the better performing classifiers to obtain a good set of hyper-parameter values. Then select the best performing model. Report the following:

***Selected model with hyper-parameter values****:*

*Mean cross-validation accuracy: ……………………….* (*rounded to 4 decimal places*)

**Task 2**. [2 Points]

Train the classifier with the hyper-parameter values determined in Task 1 on all 8,000 training samples and use it to predict the output class ‘*y*’ for the 2,000 examples in “*hw2.q2.test.csv*”*.* Report the following:

* *Accuracy on 2,000 test examples*: …………………… (rounded to 4 decimal places)
* *Classification report* for the 2,000 test examples:
* Of the 500 test samples that belong to class *y*=0, how many are correctly predicted (according to your classification report)?

**Task 3.** [2 Points]

Use the model trained in Task 2 to predict the output class ‘*y’* for the 30 examples in “*hw2.q2.new.csv*”. Specify the predicted classes in the table below:

|  |  |
| --- | --- |
| ID | predicted y |
| ID\_001 |  |
| ID\_002 |  |
| ID\_003 |  |
| ID\_004 |  |
| ID\_005 |  |
| ID\_006 |  |
| ID\_007 |  |
| ID\_008 |  |
| ID\_009 |  |
| ID\_010 |  |
| ID\_011 |  |
| ID\_012 |  |
| ID\_013 |  |
| ID\_014 |  |
| ID\_015 |  |
| ID\_016 |  |
| ID\_017 |  |
| ID\_018 |  |
| ID\_019 |  |
| ID\_020 |  |
| ID\_021 |  |
| ID\_022 |  |
| ID\_023 |  |
| ID\_024 |  |
| ID\_025 |  |
| ID\_026 |  |
| ID\_027 |  |
| ID\_028 |  |
| ID\_029 |  |
| ID\_030 |  |