**Homework 5** **Name:** Chuhao Feng
**ECE/CS 498 DS Spring 2020** **NetID:** chuhaof2
**Issued: 04/27/20**
**Due: 05/04/20 23:59:59 (No late submission allowed)**

***What to submit:***

- Jupyter notebook for coding part of problem 1
- A PDF report with the results from problem 1 and calculations of problem 2

## Problem 1

Machine Learning has seen wide applications, one of which includes physics and astronomy. You might have heard about the recent achievement of imaging a black hole. The method used in that work involved machine learning. In this homework, you'll explore the use of SVM, Decision Trees and Random Forests classifiers in another problem based in astronomy.

***Data:*** *HTRU_2.csv*

HTRU2 (High Time Resolution Universe Survey) Dataset. It describes a sample of pulsar candidates collected during the survey. Pulsar is a rate type of Neutron star. See the link for details about the study and description of the features in the dataset: https://archive.ics.uci.edu/ml/datasets/HTRU2

The data has 9 columns - 8 of which are features and 'Class' is the label. Class=0: Not a pulsar, Class=1: Is pulsar.

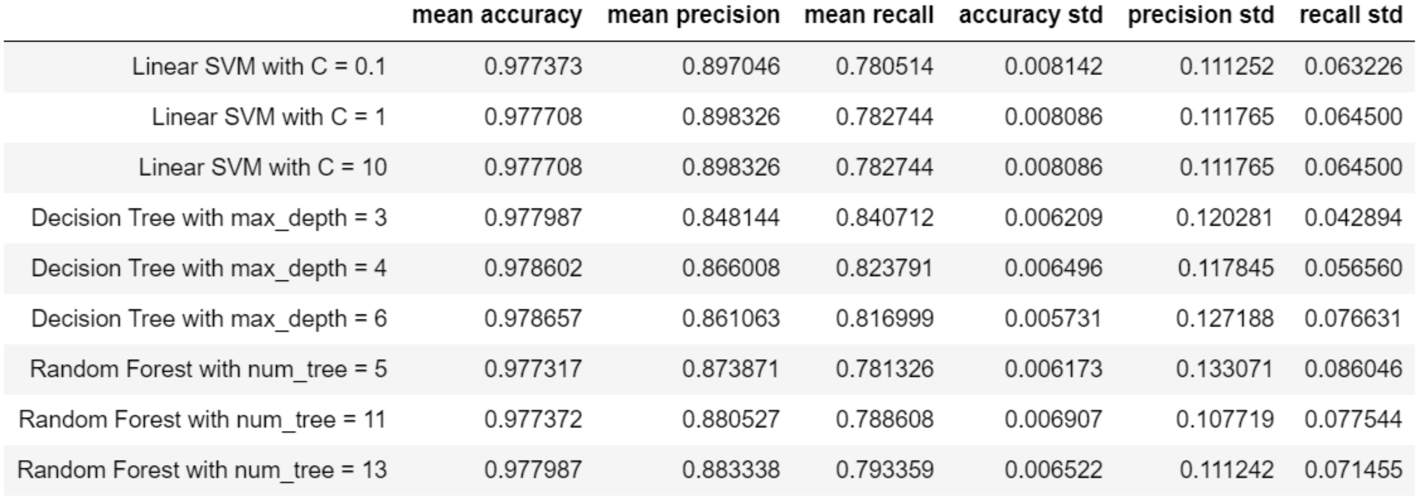
***Problem:***

Use sklearn package for SVM, decision trees, random forests, accuracy, precision and recall. Classify the samples based on the 8 features using the following methods:

- Linear SVM – regularization parameter C in {0.1, 1, 10}
- Decision Trees – maximum depth in {3, 4, 6}
- Random Forests – number of trees in {5, 11, 13}, maximum depth = 5

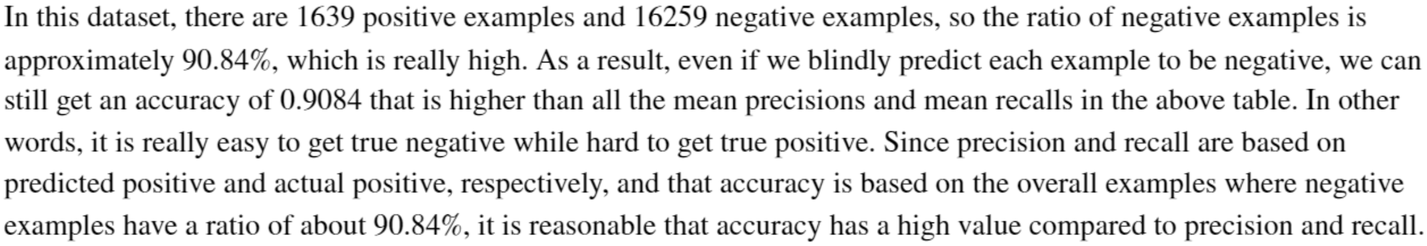
1. For each of the above perform 5-fold cross-validation of the data with 80%-20% train test split. Report the mean accuracy, precision and recall on the test

set. Report precision and recall for the positive class (pulsar). Also report the respective standard deviations. Provide the results in a table.



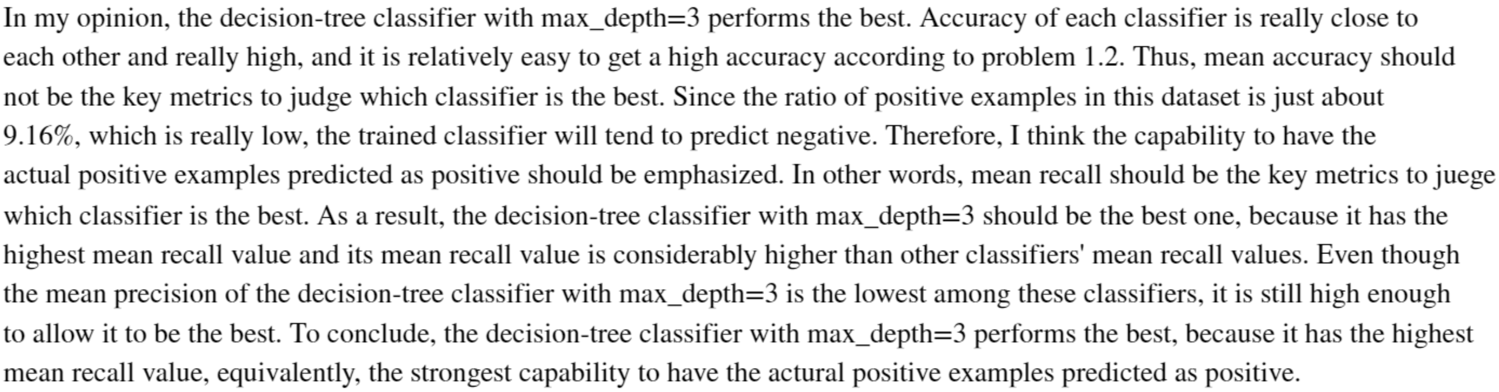
| | mean accuracy | mean precision | mean recall | accuracy std | precision std | recall std |
|---|---|---|---|---|---|---|
| Linear SVM with C = 0.1 | 0.977373 | 0.897046 | 0.780514 | 0.008142 | 0.111252 | 0.063226 |
| Linear SVM with C = 1 | 0.977708 | 0.898326 | 0.782744 | 0.008086 | 0.111765 | 0.064500 |
| Linear SVM with C = 10 | 0.977708 | 0.898326 | 0.782744 | 0.008086 | 0.111765 | 0.064500 |
| Decision Tree with max_depth = 3 | 0.977987 | 0.848144 | 0.840712 | 0.006209 | 0.120281 | 0.042894 |
| Decision Tree with max_depth = 4 | 0.978602 | 0.866008 | 0.823791 | 0.006496 | 0.117845 | 0.056560 |
| Decision Tree with max_depth = 6 | 0.978657 | 0.861063 | 0.816999 | 0.005731 | 0.127188 | 0.076631 |
| Random Forest with num_tree = 5 | 0.977317 | 0.873871 | 0.781326 | 0.006173 | 0.133071 | 0.086046 |
| Random Forest with num_tree = 11 | 0.977372 | 0.880527 | 0.788608 | 0.006907 | 0.107719 | 0.077544 |
| Random Forest with num_tree = 13 | 0.977987 | 0.883338 | 0.793359 | 0.006522 | 0.111242 | 0.071455 |

2. Explain the high value of accuracy compared to precision and recall.



In this dataset, there are 1639 positive examples and 16259 negative examples, so the ratio of negative examples is approximately 90.84%, which is really high. As a result, even if we blindly predict each example to be negative, we can still get an accuracy of 0.9084 that is higher than all the mean precisions and mean recalls in the above table. In other words, it is really easy to get true negative while hard to get true positive. Since precision and recall are based on predicted positive and actual positive, respectively, and that accuracy is based on the overall examples where negative examples have a ratio of about 90.84%, it is reasonable that accuracy has a high value compared to precision and recall.

3. Which classifier performs the best? Explain the reason why the picked classifier performed the best.

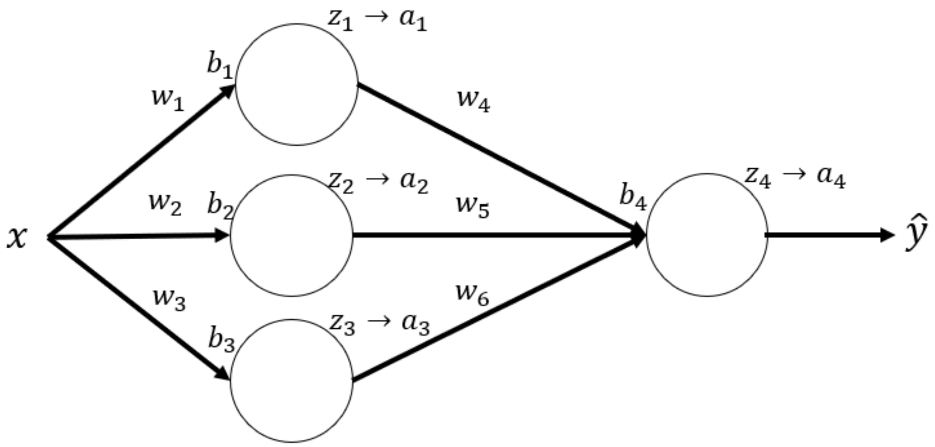


In my opinion, the decision-tree classifier with max_depth=3 performs the best. Accuracy of each classifier is really close to each other and really high, and it is relatively easy to get a high accuracy according to problem 1.2. Thus, mean accuracy should not be the key metrics to judge which classifier is the best. Since the ratio of positive examples in this dataset is just about 9.16%, which is really low, the trained classifier will tend to predict negative. Therefore, I think the capability to have the actual positive examples predicted as positive should be emphasized. In other words, mean recall should be the key metrics to juege which classifier is the best. As a result, the decision-tree classifier with max_depth=3 should be the best one, because it has the highest mean recall value and its mean recall value is considerably higher than other classifiers' mean recall values. Even though the mean precision of the decision-tree classifier with max_depth=3 is the lowest among these classifiers, it is still high enough to allow it to be the best. To conclude, the decision-tree classifier with max_depth=3 performs the best, because it has the highest mean recall value, equivalently, the strongest capability to have the actural positive examples predicted as positive.

## Problem 2

Consider the neural network given below. Assume that all the neurons use the sigmoid activation function.

**Homework 5** **Name: ____________________**
**ECE/CS 498 DS Spring 2020** **NetID: ____________________**
**Issued: 04/27/20**
**Due: 05/04/20 23:59:59 (No late submission allowed)**



1. Write down the expressions for $z_1, z_4, a_1$ and $a_4$. Use them to express the output of the neural network given an input $x_1$, weights $w_1, w_2, w_3, w_4, w_5, w_6$ and biases $b_1, b_2, b_3, b_4$.

$$z_1 = x \cdot w_1 + b_1,\ a_1 = \sigma(z_1)$$
$$z_4 = a_1 \cdot w_4 + a_2 \cdot w_5 + a_3 \cdot w_6 + b_4,\ a_4 = \sigma(z_4)$$
$$\hat{y} = a_4 = \sigma(z_4) = \sigma(a_1 \cdot w_4 + a_2 \cdot w_5 + a_3 \cdot w_6 + b_4)$$
$$= \sigma(\sigma(z_1) \cdot w_4 + \sigma(z_2) \cdot w_5 + \sigma(z_3) \cdot w_6 + b_4)$$
$$= \sigma(\sigma(x_1 \cdot w_1 + b_1) \cdot w_4 + \sigma(x_1 \cdot w_2 + b_2) \cdot w_5 + \sigma(x_1 \cdot w_3 + b_3) \cdot w_6 + b_4)$$

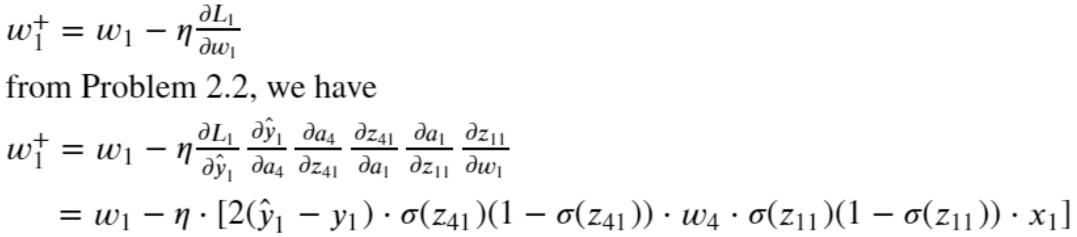
2. In the following questions, we use mean squared error as the loss function $L$. Consider a single sample $x_i$ , with predicted label $\hat{y}_i$ and true label $y_i$. Write down the equation for $L_i$ and for the gradients $\frac{\partial L_i}{\partial w_1}, \frac{\partial L_i}{\partial w_4}, \frac{\partial L_i}{\partial b_1},$ and $\frac{\partial L_i}{\partial b_4}$. (You may use $\hat{y}_i, y_i, z_{1i,} z_{2i,} z_{3i,} z_{4i,}\ w_1, w_2, w_3, w_4, w_5, w_6,\ b_1, b_2, b_3, b_4$ and $x_i$ in your expressions.)

$$L_i = (\hat{y}_i - y_i)^2$$
$$\frac{\partial L_i}{\partial w_1} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial a_4} \frac{\partial a_4}{\partial z_{4i}} \frac{\partial z_{4i}}{\partial a_1} \frac{\partial a_1}{\partial z_{1i}} \frac{\partial z_{1i}}{\partial w_1}$$
$$\frac{\partial L_i}{\partial w_4} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial a_4} \frac{\partial a_4}{\partial z_{4i}} \frac{\partial z_{4i}}{\partial w_4}$$
$$\frac{\partial L_i}{\partial b_1} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial a_4} \frac{\partial a_4}{\partial z_{4i}} \frac{\partial z_{4i}}{\partial a_1} \frac{\partial a_1}{\partial z_{1i}} \frac{\partial z_{1i}}{\partial b_1}$$
$$\frac{\partial L_i}{\partial b_4} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial a_4} \frac{\partial a_4}{\partial z_{4i}} \frac{\partial z_{4i}}{\partial b_4}$$

$$\because \frac{\partial L_i}{\partial \hat{y}_i} = 2(\hat{y}_i - y_i),\ \frac{\partial \hat{y}_i}{\partial a_4} = 1$$
$$\because \frac{\partial a_4}{\partial z_{4i}} = \sigma(z_{4i})(1 - \sigma(z_{4i})),\ \frac{\partial z_{4i}}{\partial a_1} = w_4$$
$$\because \frac{\partial a_1}{\partial z_{1i}} = \sigma(z_{1i})(1 - \sigma(z_{1i})),\ \frac{\partial z_{1i}}{\partial w_1} = x_i$$
$$\because \frac{\partial z_{4i}}{\partial w_4} = a_1 = \sigma(z_{1i}),\ \frac{\partial z_{1i}}{\partial b_1} = 1,\ \frac{\partial z_{4i}}{\partial b_4} = 1$$

$$\therefore \frac{\partial L_i}{\partial w_1} = 2(\hat{y}_i - y_i) \cdot \sigma(z_{4i})(1 - \sigma(z_{4i})) \cdot w_4 \cdot \sigma(z_{1i})(1 - \sigma(z_{1i})) \cdot x_i$$
$$\therefore \frac{\partial L_i}{\partial w_4} = 2(\hat{y}_i - y_i) \cdot \sigma(z_{4i})(1 - \sigma(z_{4i})) \cdot \sigma(z_{1i})$$
$$\therefore \frac{\partial L_i}{\partial b_1} = 2(\hat{y}_i - y_i) \cdot \sigma(z_{4i})(1 - \sigma(z_{4i})) \cdot w_4 \cdot \sigma(z_{1i})(1 - \sigma(z_{1i}))$$
$$\therefore \frac{\partial L_i}{\partial b_4} = 2(\hat{y}_i - y_i) \cdot \sigma(z_{4i})(1 - \sigma(z_{4i}))$$

3. Given a single training sample $x_1$, write the gradient descent equation for updating weight $w_1$. (The learning rate, which determines the step size at each iteration, is denoted as $\eta$.)



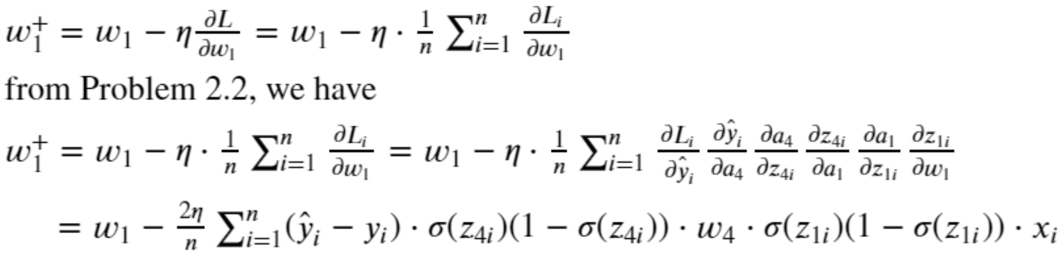
$$w_1^+ = w_1 - \eta \frac{\partial L_1}{\partial w_1}$$

from Problem 2.2, we have

$$w_1^+ = w_1 - \eta \frac{\partial L_1}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial a_4} \frac{\partial a_4}{\partial z_{41}} \frac{\partial z_{41}}{\partial a_1} \frac{\partial a_1}{\partial z_{11}} \frac{\partial z_{11}}{\partial w_1}$$

$$= w_1 - \eta \cdot [2(\hat{y}_1 - y_1) \cdot \sigma(z_{41})(1 - \sigma(z_{41})) \cdot w_4 \cdot \sigma(z_{11})(1 - \sigma(z_{11})) \cdot x_1]$$

4. Given $n$ training samples $x_i$ for $i = 1,2,3,\ldots,n$, write the gradient descent equation for updating weight $w_1$. (The learning rate, which determines the step size at each iteration, is denoted as $\eta$.)



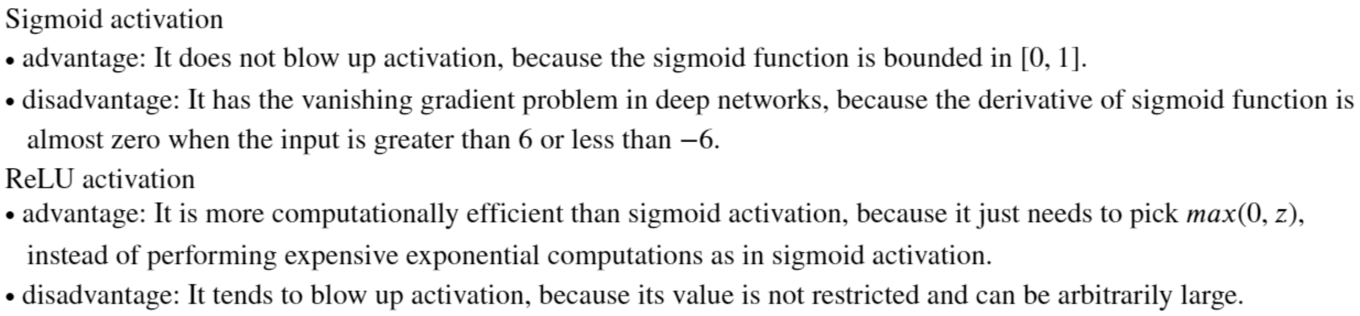
$$w_1^+ = w_1 - \eta \frac{\partial L}{\partial w_1} = w_1 - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L_i}{\partial w_1}$$

from Problem 2.2, we have

$$w_1^+ = w_1 - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L_i}{\partial w_1} = w_1 - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial a_4} \frac{\partial a_4}{\partial z_{4i}} \frac{\partial z_{4i}}{\partial a_1} \frac{\partial a_1}{\partial z_{1i}} \frac{\partial z_{1i}}{\partial w_1}$$

$$= w_1 - \frac{2\eta}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot \sigma(z_{4i})(1 - \sigma(z_{4i})) \cdot w_4 \cdot \sigma(z_{1i})(1 - \sigma(z_{1i})) \cdot x_i$$

5. In the lecture we briefly mentioned two activation functions: 1) Sigmoid activation, and 2) ReLU activation. **List one advantage and one disadvantage for each of these two activation functions and briefly explain.** You may want to search online for this question. You can also conclude the advantages, disadvantages by explaining the properties of the functions itself or their first order derivatives.



Sigmoid activation

- advantage: It does not blow up activation, because the sigmoid function is bounded in [0, 1].
- disadvantage: It has the vanishing gradient problem in deep networks, because the derivative of sigmoid function is almost zero when the input is greater than 6 or less than $-6$.

ReLU activation

- advantage: It is more computationally efficient than sigmoid activation, because it just needs to pick $max(0, z)$, instead of performing expensive exponential computations as in sigmoid activation.
- disadvantage: It tends to blow up activation, because its value is not restricted and can be arbitrarily large.