**DSCI 35600 – HW 03 (24 points)**  Name\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Problem 1 (14 pts).** For each of the sentences below, fill in each blank with one of the terms provided at the top of the page. You may not need to use every term, and some terms might be used more than once.

|  |  |  |  |
| --- | --- | --- | --- |
| **overfit** | **underfit** | **classification** | **regression** |
| **loss** | **testing** | **training** | **validation** |
| **minimize** | **maximize** | **r-squared** | **accuracy** |
| **fit** | **predict** | **SSE** | **negative log-likelihood** |

1. The goal of a learning algorithm in a supervised learning task (whether classification or regression) is to

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ the value of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ function, as calculated on the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ set.

1. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ set is used to compare the performance of different models, and to select the final model.
2. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ set is used to assess how well your final model will generalize to new, unseen examples.
3. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ loss function is used for classification algorithms.
4. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ loss function is used for regression algorithms.
5. The score method of a Scikit-Learn regression model returns the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ metric for the model, as calculated on the dataset provided.
6. The score method of a Scikit-Learn classification model returns the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ metric for the model, as calculated on the dataset provided.
7. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ task is a supervised learning problem in which the target values are categorical.
8. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ task is a supervised learning problem in which the target values are real numbers.
9. An \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ model will perform very well on the training data, but will not likely generalize well.
10. An \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ model performs poorly on the training data, and will also not likely generalize well.
11. The more rigid a model is, the more likely it is to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
12. The more flexible a model is, the more likely it is to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Problem 2 (10 pts).** Assume that you are provided with two 2D arrays, X\_num and X\_cat, whose contents are as shown below. The array X\_num contains the values for a single numerical (quantitative) feature, while X\_cat contains the values for two categorical (qualitative) features. The feature arrays are preprocessed and combined into a single array by running the code below. Provide the contents of the array X\_preprocessed by completing the table on the right. You may not need all of the columns provided. If not, leave any extra columns blank.



**X\_num X\_cat X\_preprocessed**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 |  | 2 | S |  |  |  |  |  |  |  |  |  |  |  |
| 1 |  | 1 | F |  |  |  |  |  |  |  |  |  |  |  |
| 0 |  | 1 | S |  |  |  |  |  |  |  |  |  |  |  |
| 2 |  | 1 | D |  |  |  |  |  |  |  |  |  |  |  |
| 4 |  | 2 | D |  |  |  |  |  |  |  |  |  |  |  |
| 2 |  | 1 | F |  |  |  |  |  |  |  |  |  |  |  |