DSCI 35600 – HW 03 (24 points) Due: Monday, 2/11 Name Key

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**

a) The goal of a learning algorithm in a supervised learning task (whether classification or regression) is to minimize the value of the loss function, as calculated on the training set.

b) The validation set is used to compare the performance of different models, and to select the final model.

c) The testing set is used to assess how well your final model will generalize to new, unseen examples.

d) The negative log likelihood loss function is used for classification algorithms.

e) The SSE loss function is used for regression algorithms.

f) The score method of a Scikit-Learn regression model returns the r-squared metric for the model, as calculated on the dataset provided.

g) The score method of a Scikit-Learn classification model returns the accuracy metric for the model, as calculated on the dataset provided.

h) A classification task is a supervised learning problem in which the target values are categorical.

i) A regression task is a supervised learning problem in which the target values are real numbers.

j) An overfit model will perform very well on the training data, but will not likely generalize well.

k) An underfit model performs poorly on the training data, and will also not likely generalize well.

l) The more rigid a model is, the more likely it is to underfit.

m) The more flexible a model is, the more likely it is to overfit.

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.

```
from sklearn.preprocessing import PolynomialFeatures, OneHotEncoder

poly = PolynomialFeatures(2)
Xp = poly.fit_transform(X_num)

enc = OneHotEncoder(sparse=False)
Xe = enc.fit_transform(X_cat)

X_preprocessed = np.hstack((Xp, Xe))
```

X_num

3
1
0
2
4
2

X_cat

2	S
1	F
1	S
1	D
2	D
1	F

X_preprocessed

1	3	9	0	1	0	0	1		
1	1	1	1	0	0	1	0		
1	0	0	1	0	0	0	1		
1	2	4	1	0	1	0	0		
1	4	16	0	1	1	0	0		
1	2	4	1	0	0	1	0		