DS-UA-112: Introduction to Data Science (Fall 2019)

Section Practice Problem

- The exam has 7 pages. Mark your answers on the exam itself; we will not grade answers written on scratch paper. Please mark your answers legibly to help with document scanning.

Name: _____

NYU NetID: _____

NYU Email: _____
(as it appears on Gradescope)

Question	Points	Score
1	0	
2	0	
3	0	
4	0	
5	0	
6	0	
7	0	
8	0	
Total:	0	

1. **Probability and Sampling**

   (a) Assume that the room DS-UA 112 office hours are held in has a maximum capacity of 15 people. Given that there are 100 students in the course, and each student has a 1/200 probability of attending office hours, what is the probability that there is not enough space? Here assume that the probability of each student attending is independent.

   ☐ $\frac{15}{100}$

   ■ $1 - \sum_{i=0}^{15} \binom{100}{i} (\frac{1}{200})^i (\frac{199}{200})^{100-i}$

   ☐ $\sum_{i=16}^{100} \frac{1}{200}$

   ☐ $\sum_{i=16}^{100} (\frac{1}{200})^i (\frac{199}{200})^{100-i}$

   ☐ none of the above

   (b) You and your friend have a favorite neighborhood Japanese restaurant. You know that if your friend orders Ramen, they usually leave full, with probability 0.8. However, if he orders Udon, it doesn't always fill them up, with probability 0.5. As Udon is cheaper, your prior assumption is that your friend orders Udon, with probability 0.7. If your friend left full, what is your estimate of the probability they ordered Ramen?

   Solution: $P(R = 1|F = 1) = \frac{P(F=1|R=1)P(R=1)}{P(F=1|R=1)P(R=1)+P(F=1|R=0)P(R=0)} = \frac{(0.8)(0.3)}{(0.8)(0.3)+(0.5)(0.7)} = \frac{0.24}{0.59}$

2. **Modeling**

   (a) What parameter estimate would minimize the following regularized loss function:

$$\ell(\theta) = \lambda(\theta - 4)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{1}$$

   Solution:
   Taking the derivative of the loss function we get:

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{\partial}{\partial\theta}\lambda(\theta - 4)^2 + \frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{2}$$

$$= 2\lambda(\theta - 4) - \frac{2}{n}\sum_{i=1}^{n}(x_i - \theta) \tag{3}$$

$$\tag{4}$$

Setting the derivative equal to zero and solving for θ:

$$2\lambda(\theta - 4) = \frac{2}{n}\sum_{i=1}^{n}(x_i - \theta) \tag{5}$$

$$\lambda\theta - 4\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \theta \tag{6}$$

$$\lambda\theta + \theta = 4\lambda + \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$\theta = \frac{4\lambda}{\lambda + 1} + \frac{1}{n(\lambda + 1)}\sum_{i=1}^{n} x_i \tag{8}$$

$$\tag{9}$$

How can we prevent overfitting? Select all that apply.

- ■ **Choosing a different loss function**
- ☐ Discard any feature with duplicate values
- ■ **Using cross-validation to choose extra parameters in the model**
- ☐ Making the training set smaller and the testing set larger.
- ☐ Adding features to the model

(b) **TODO: one more sampling question**

3. **Regression**

(a) Recall from lecture that a linear model is defined as a model where our prediction $\hat{y}$ is given by the equation below, where d is the number of parameters in our model:

$$\hat{y} = f_\theta(x) = \sum_{j=1}^{d} \theta_j \varphi_j(x)$$

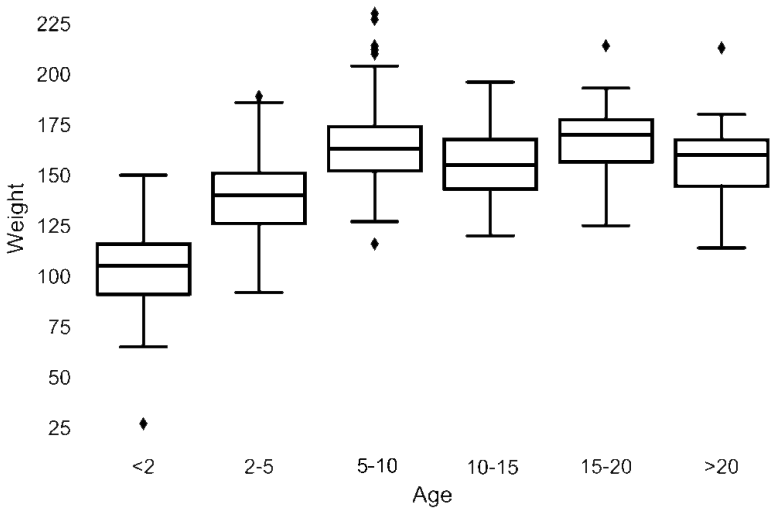
Which of the following models are linear? **Select all that apply.**

- ■ $f_\theta(x) = \theta_1 x + \theta_2 \sin(x)$
- ■ $f_\theta(x) = \theta_1 x + \theta_2 \sin(x^2)$
- ■ $f_\theta(x) = \theta_1$
- ■ $f_\theta(x) = (\theta_1 x + \theta_2)x$
- ☐ $f_\theta(x) = \ln(\theta_1 x + \theta_2) + \theta_3$

(b) i. In developing a model for a donkeys weight, consider the following box plots of weight by age category.

This plot suggests:

- ☐ age is not needed in the model
- ■ **some of the age categories can be combined**
- ☐ age could be treated as a numeric variable
- ☐ none of the above



ii. Suppose that we try to predict a donkey's weight, y_i from its sex alone. (Recall that the sex variable has values: gelding, stallion, and female). Consider the following model consisting of dummy variables:

$$y_i = \theta_F D_{F,i} + \theta_G D_{G,i} + \theta_S D_{S,i}$$

where the dummy variable $D_{F,i} = 1$ if the i^{th} donkey is female and $D_{F,i} = 0$ otherwise. The dummy variables D_G and D_S are dummies for geldings and stallions, respectively.
Check that if we using the following loss function:

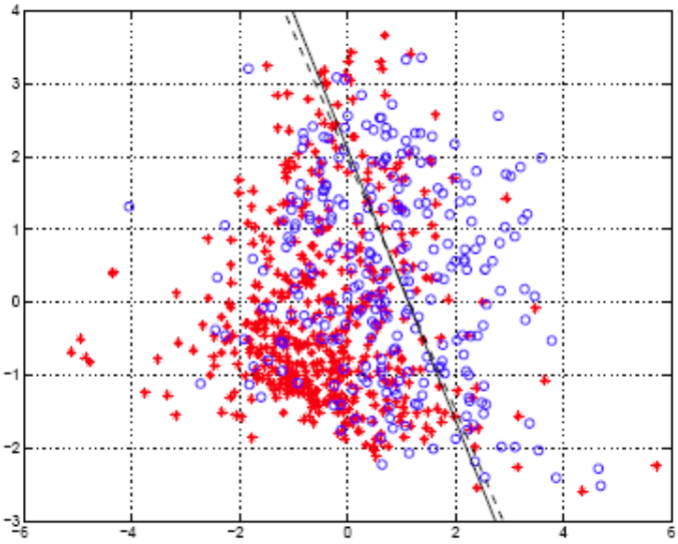
$$L(\theta_F, \theta_G, \theta_S) = \sum_{i=1}^{n} \left(y_i - (\theta_F D_{F,i} + \theta_G D_{G,i} + \theta_S D_{S,i})\right)^2$$

then the loss minimizing value $\hat{\theta}_F = \bar{y}_F$ where $\bar{y}_F$ is the average weight of the female donkeys.

4. **Classification and Metrics**

(a) We want to classify data into two categories. Below is a scatter-plot of the data that indicates the categories with different shapes. We choose to use logistic regression. Having fit the model, we obtain a decision boundary indicated by the line.

___**F**___ **True or False:** We ***expect*** the model to have low bias



(b) Suppose we want to classify NYU students into campus New York, Shanghai, Abu Dhabi by major. Could we use the One-vs-Rest approach with logistic regression to fit...

(c) Suppose we have a classifier to determine whether an image contains a picture of a bobcat – the NYU mascot. We test it on 23 images.

- There were 12 images that contain bobcats. The classifier predicted 9 of them to be bobcats and 3 to be not bobcats.
- There were 11 images that did not contain bobcats. The classifier predicted 3 or them to be bobcats and 8 to be not bobcats.

i. Fill in the following confusion matrix. Please compute the total in each row and column.

		Observed: True	Observed: False	Total
Predicted	True	**9**	**3**	**12**
	False	**3**	**8**	**11**
	Total	**12**	**11**	

ii. Compute the following metrics

1. **Accuracy** **17/23**
2. **Recall** **9/12**
3. **Precision** **9/12**

5. **Regular Expressions**

ccccc

Id	Date	Product	Company
0	99/99/99	Debt collection	California Accounts Service
1	06/15/10	Credit reporting	EXPERIAN INFORMATION SOLUTIONS INC
3	10/21/14	MORTGAGE	OCWEN LOAN SERVICING LLC
5	03/30/15		The CBE Group Inc
6	02/03/16	Debt collection	The CBE Group, Inc.
7	01/07/17	Credit reporting	Experian Information Solutions Inc.
8	03/15/17	Credit card	FIRST NATIONAL BANK OF OMAHA

i. Select all the true statements from the following list.

- ■ **Some of the product values appear to be missing.**
- ■ **Some of the date values appear to be missing.**
- ☐ The file is comma delimited

- ■ **The file is fixed width formatted.**
- ■ **To analyze the companies we will need to correct for variation in capitalization and punctuation.**
- ☐ None of the above statements are true.

ii. Select all of the following regular expressions that properly match the dates.

- ☐ `\d?/\d?/\d?`
- ■ `\d+/\d+/\d+`
- ■ `\d*/\d*/\d*`
- ■ `\d\d/\d\d/\d\d`
- ☐ None of the above regular expressions match.

iii. which of the following regular expressions exactly matches the entry `FIRST NATIONAL BANK OF OMAHA`? Select all that matches.

- ☐ `[A-Z]*`
- ■ `FIR[A-Z,\s]* OMAHA`
- ■ `F[A-Z, \s]+A`
- ☐ `F[A-Z]*`
- ☐ None of the above regular expressions match.

6. **Features**

(a) Suppose we have

$$\text{training set: } [4, 2, 6, 4]$$

$$\text{test:} [8, 6]$$

We want to normalize the data by applying a transformation. For example, we can subtract the mean and divide by the standard deviation like in homework 6. Should we

- ☐ Subtract 4 from all original values.
- ☐ Subtract 5 from all original values.
- ■ **Subtract 4 from all train values and 7 from all test values.**

(b) Suppose we have a column of a table containing age. Note that some of the rows have blank entries and no row contains an age greater than 100. We decide to encode ages in particular ranges with numbers

0 to 20	0
20 to 60	1
60 to 100	2
missing	9

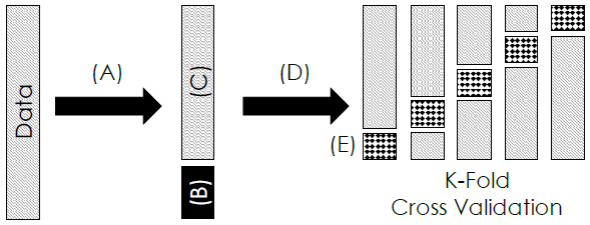
i. **__F__ True or False:** Through the encoding, we have incorporated polynomial features into the model?

ii. **__F__ True or False:** Without impacting the model, we could reduce from four values (0,1,2,9) to three values (0,1,2) by dropping missing entries?

iii. **__T__ True or False:** We should replace 9 with `NaN` to indicate that the data was missing.

7. **Validation**

(a) In this question we complete the following figure describing the train-test split and k-fold cross validation. Note the data table with many records and a few columns is depicted on the left as a tall rectangle.



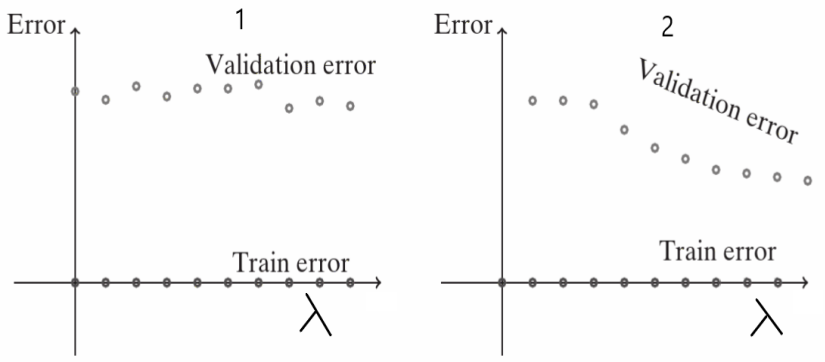
i. This part of the figure refers to the validation data.
☐ (A) ☐ (B) ☐ (C) ☐ (D) ■ **(E)**

ii. This part of the figure refers to the testing data
☐ (A) ■ **(B)** ☐ (C) ☐ (D) ☐ (E)

iii. This part of the figure refers to the process of constructing the train-test split.
■ **(A)** ☐ (B) ☐ (C) ☐ (D) ☐ (E)

iv. Select all the following statements that apply to the above figure.
- ■ **This figure illustrates 5-fold cross validation.**
- ☐ This figure illustrates 6-fold cross-validation.
- ■ **Assuming all the data points are distinct each of the validation data sets are also distinct.**
- ☐ The test data should be used during cross-validation to fully evaluate the model.

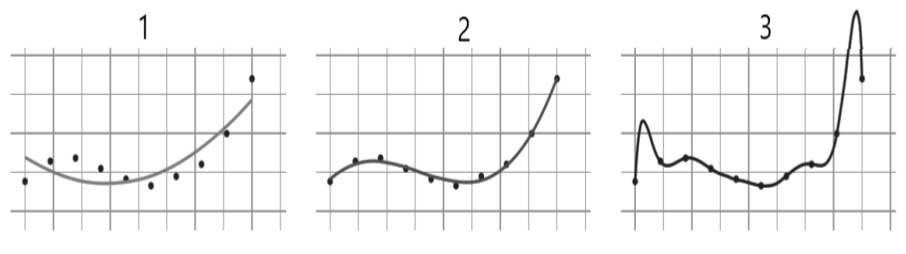
(b) Suppose we are fitting a small dataset and a large dataset. Take λ to be the extra parameter that affects regularization. Here λ reanges from small values on the left to large values on the right. Which of the following charts would likely correspond to the small dataset and large dataset?



☐ 1. Large 2. Small
■ **1. Small 2. Large**

8. **Regularization**

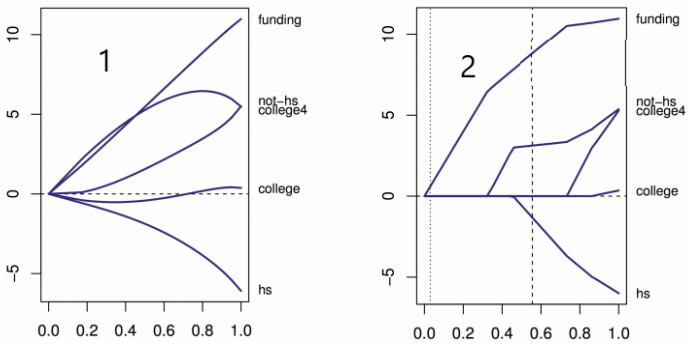
(a) Based on the sample, how would you ***expect*** to label the following fits:



☐ 1. Overfitting 2. Neither 3. Underfitting

☐ 1. Neither 2. Overfitting 3. Underfitting
☐ 1. Underfitting 2. Overfitting 3. Neither
■ **1. Underfitting 2. Neither 3. Overfitting**

(b) Label the regularization paths as Ridge Regression or Lasso Regression



■ **1. Ridge 2. Lasso**
☐ 1. Lasso 2. Ridge

END OF EXAM – PRESENT YOUR NYU ID AT SUBMISSION