DS-UA-112: Introduction to Data Science (Spring 2020)

Final Exam (December 16 6:00-8:00PM)

- The exam has 8 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	9	
2	8	
3	8	
4	7	
5	6	
6	2	
7	4	
8	7	
9	3	
Total:	54	

1. **Conceptual**

   (a) Suppose we have a dataset about emails like Project 2 on spam.

      i. (1 point) __**T**__ **True or False:** The training set has 98 examples of not-spam and 2 examples of spam. We can determine a model with 98% accuracy.

      ii. (1 point) __**F**__ **True or False:** If we chose a model that always predicts not-spam, then the accuracy on the testing set will be at least 50%.

      iii. (1 point) __**T**__ **True or False:** If we want to reduce the number of false negatives, then we should assess the recall of the model.

   (b) Suppose we have a dataset about real estate like Lab 12 on housing prices.

      i. (1 point) __**F**__ **True or False:** If the dataset is large with records of many houses, then batch gradient descent is preferable to stochastic gradient descent for fitting the model?

      ii. (1 point) __**T**__ **True or False:** If the size of the house is reported in square feet and the size of the garage is reported in square meters, then would we need to standardize the data for a model with regularization?

      iii. (1 point) __**F**__ **True or False:** A column in the table storing the training set has data type `int64`. Can we assume that the column represents a numerical attribute?

   (c) Suppose we have a dataset about fuel efficiency like Lecture 18 on cars.

      i. (1 point) __**T**__ **True or False:** We want to include the weight of cars. From among the samples, we weigh several cars twice. Comparing the numbers between different weighing will help us to assess noise in the dataset?

      ii. (1 point) __**T**__ **True or False:** Suppose the dataset reports fuel efficiency in terms of miles per gallon in cities and miles per gallon on highways. Could excluding the numbers about miles per gallon on highways bias the predictions about fuel efficiency?

      iii. (1 point) __**T**__ **True or False:** The continent of the manufacturer is a feature. If we replace continent with country, then have we changed the granularity of the dataset?

2. **Probability and Sampling** We are interested in the relationship between name, gender and age. Rather than study the Social Security Administration's dataset from Lecture 1, we want to collect a sample from the population. Below we have the sampling frame consisting of people from the population eligible for inclusion in the random sample.

Age Range	Male	Female
20-29	10	20
30-39	25	20
40-49	15	30
50-59	10	20

   (a) i. (1 point) __**T**__ **True or False:** In a simple random sample of 30 participants, you expect to have approximately eighteen women.

      ii. (1 point) __**T**__ **True or False:** Suppose you drop the 30-39 age group from the sampling frame. In a stratified sample where each age range is represented equally, you expect to have the same proportion of women as in a simple random sample.

   (b) (1 point) If we use simple random sampling to choose 30 people, then what is the probability that the sample contains more men than women?

      ☐ $\left(1 - \frac{60}{90}\right)^{30}$

      ☐ $\frac{1}{\binom{60}{30}}\left(\binom{60}{0} + \ldots + \binom{60}{15}\right)$

      ☐ $\binom{30-1}{15} \cdot \ldots \cdot \binom{30-15}{15}$

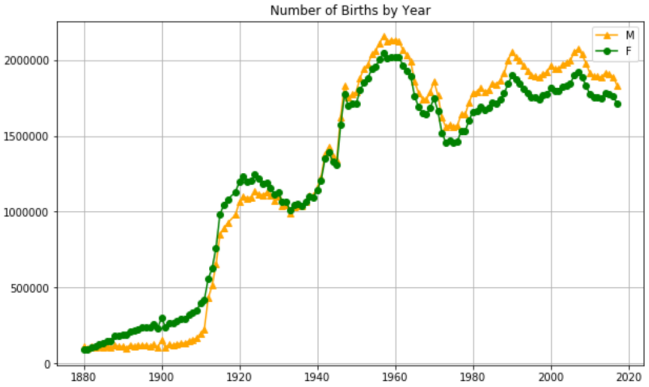
      ■ $\frac{1}{\binom{150}{30}}\left(\binom{60}{16}\binom{90}{14} + \ldots + \binom{60}{30}\binom{90}{0}\right)$

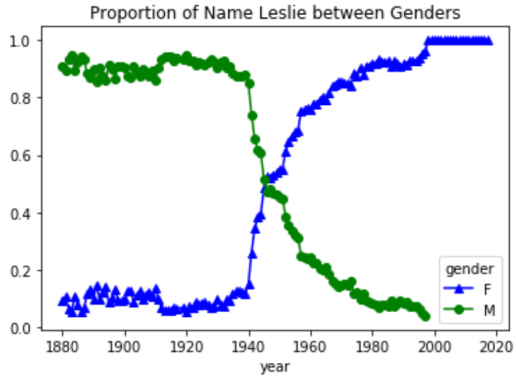
      ☐ $1/\binom{150}{30}$

(c) (1 point) If we use cluster sampling with 2 groups from amongst the 4 age ranges, then what is the probability that the sample contains twice the number of women as men?

☐ 0 ☐ $\frac{1}{4}$ ■ $\frac{1}{2}$ ☐ $\frac{1}{8}$ ☐ 1

(d) Note the change in the gender distribution between 1920 and 1980. The probability of being female in 1920 is approximately 0.55 and the probability of being female in 1980 is approximately 0.45.





Note the change in the gender distribution for the name Leslie. For comparison, we want to calculate the relationship between name and gender in both 1920 and 1980. For each year, we calculate the probability of the name Leslie conditional on the gender. Below we have summarized the conditional probabilities in a table

Age Range	Male	Female
1920	0.009	0.001
1980	0.002	0.008

i. (2 points) Compute the probability of being male in 1920 conditional on being named Leslie.

Solution: $P(M=1|L=1) = \frac{P(L=1|M=1)P(M=1)}{P(L=1|M=1)P(M)+P(L=1|F=1)P(F)} = \frac{(0.009)(0.45)}{(0.009)(0.45)+(0.001)(0.55)}$

ii. (2 points) Compute the probability of being male in 1980 conditional on being named Leslie.

Solution: $P(M=1|L=1) = \frac{P(L=1|M=1)P(M=1)}{P(L=1|M=1)P(M)+P(L=1|F=1)P(F)} = \frac{(0.002)(0.55)}{(0.002)(0.55)+(0.008)(0.45)}$

3. Modeling

(a) Suppose we have observations x_i, y_i in the training set. We want to combine both types of regularization

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2 + \alpha\theta^2 + \beta|\theta|$$

Here $\alpha > 0$ and $\beta > 0$ are the extra parameters. We want to choose the parameter $\widehat{\theta}$ that minimizes the function.

i. (1 point) ___F___ **True or False.** If α is large and β is small, then we **expect** θ to be zero.

ii. (1 point) ___F___ **True or False.** If we replace α, β with $-\alpha, -\beta$, then we get the same value of $\widehat{\theta}$

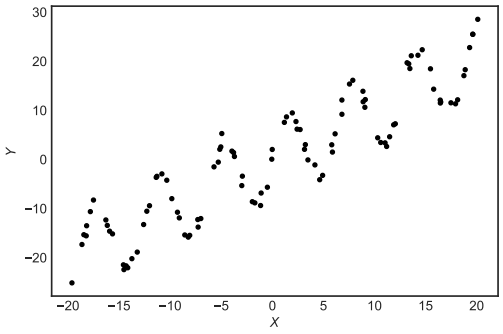
(b) (6 points) Suppose we want to choose weights θ to minimize the following function for a model

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i \cdot \theta)^2 + \lambda\sum_{j=1}^{d}\theta_j^2$$

i. How many observations in the dataset? _____n_____
ii. What is the dimension of the data?
iii. Is this more likely classification or regression? __**regression**__
iv. What type of regularization is being used? _____L_2_____
v. As λ increases what will happen to bias? __**increase**__
vi. As λ decreases what will happen to variance? __**increase**__

4. **Regression**

(a) (1 point) Below we have a scatter-plot of a training set like Homework 5. Which model would be the most appropriate **linear model** to fit the training set?



- ☐ $y = \theta_1 x + \theta_2$
- ☐ $y = \sum_{k=1}^{d} \theta_k x^k$
- ■ $y = \theta_1 x + \theta_2 \sin(x)$
- ☐ $y = \theta_1 x + \theta_2 \sin(\theta_3 x)$
- ☐ Since y is a non-linear function of x, the relationship can't be expressed by a linear model.

(b) (3 points) Suppose we want to predict tips at restaurants like in Lab 11. Recall that the party is smoking or not smoking. Consider the following model for tip y_i:

$$y_i = \theta_S D_{S,i} + \theta_N D_{N,i}$$

Here $D_{S,i}, D_{N,i}$ are dummy variables. If the i^{th} party is smoking, then $D_{S,i} = 1$ and $D_{N,i} = 0$. If the i^{th} party is not smoking, then $D_{S,i} = 0$ and $D_{N,i} = 1$. Suppose we take the square loss function:

$$L(\theta_S, \theta_N) = \sum_{i=1}^{n} \left(y_i - (\theta_S D_{S,i} + \theta_N D_{N,i})\right)^2$$

Show that $\widehat{\theta}_S$ is the average tip amongst smoking parties. You must find $\widehat{\theta}_S$ such that $\frac{\partial}{\partial \theta_S} L\left(\widehat{\theta_S}\right) = 0$.

Solution: Since

$$L(\theta_S, \theta_N) = \sum_{i=1, D_{S,i}=1}^{n} (y_i - \theta_S)^2 + \sum_{i=1, D_{S,i}=0}^{n} (y_i - \theta_N)^2$$

we have

$$\frac{\partial}{\partial \theta_S} L(\theta_S) = -\sum_{i=1, D_{S,i}=1}^{n} y_i - \theta_S$$

Therefore $\widehat{\theta}_S = \frac{1}{\#\{i: D_{S,i}=1\}} \sum_{i=1, D_{S,i}=1}^{n} y_i$

(c) (3 points) Suppose we create a new loss function called the high-low loss, defined as follows for a single observation:

$$\ell_{high-low}(\theta, x, y) = \begin{cases} 3(y - f_\theta(x))^2 & f_\theta(x) \geq y \\ 12(y - \frac{1}{2} f_\theta(x))^2 & f_\theta(x) < y \end{cases}$$

You decide to use the constant model $f_\theta(x) = \theta$ with average high-low loss

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_{high-low}(\theta, x_i, y_i)$$

Below we have the training set. Find $\hat{\theta}$ that minimizes the loss.

x	3	1	5	4	7
y	-10	5	0	-20	10

Solution: Note that we receive double the loss for examples where $\theta < y$, and that we can throw away the factor of $\frac{1}{n}$ that averages:

$$L(\theta) = 3 \sum_{\theta \geq y} (y - \theta)^2 + 12 \sum_{\theta < y} (y - \frac{1}{2}\theta)^2$$

$$\frac{dL}{d\theta} = -(3)(2) \sum_{\theta \geq y} (y - \theta) - (12)(2)(\frac{1}{2}) \sum_{\theta < y} (y - \frac{1}{2}\theta)$$

$$0 = \frac{dL}{d\theta} = \sum_{\theta \geq y} (y - \theta) + \sum_{\theta < y} (2y - \theta)$$

Therefore $\widehat{\theta} = \frac{1}{n} \left(\sum_{y \leq \theta} y + 2 \sum_{y > \theta} y \right)$. We can take $\widehat{\theta} = 0$

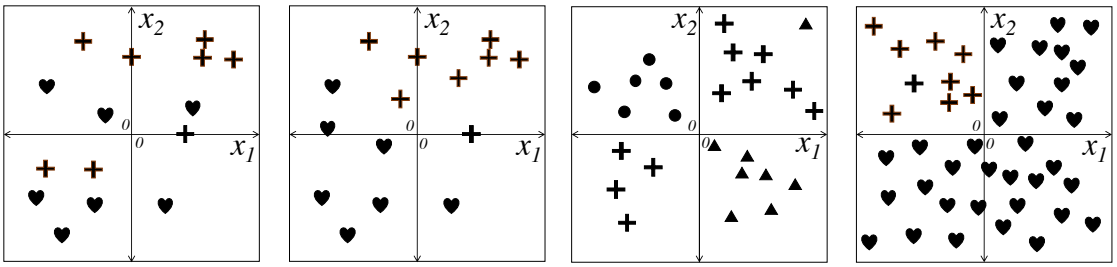
5. **Classification and Metrics**

(a) i. (1 point) **T** **True or False.** AUC is a number between 0 and 1.

ii. (1 point) The points on a ROC plot represent

- ■ **performance trade-offs at different thresholds**
- ☐ a comparison between precision and recall
- ☐ a ranking of categories in classification
- ☐ the average value of the loss function over the training set

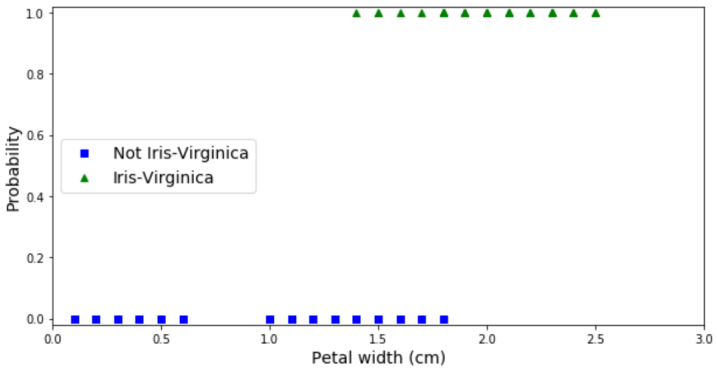
(b) Consider the following figures of different shapes plotted in a two dimensional feature space. Suppose we are interested in classifying the type of shape based on the location.



i. (1 point) Which figure shows the largest class imbalance?
☐ (A) ☐ (B) ☐ (C) ■ **(D)**

ii. (1 point) Which figure contains linearly separable data.
☐ (A) ■ **(B)** ☐ (C) ☐ (D)

(c) Suppose we have the Iris dataset from Lecture 24 on flowers. We have two categories: variety Iris-Virginica and other variety.



i. (1 point) If we use Logistic Regression to fit the data, then will the weight θ be
☐ Negative ☐ Zero ■ **Positive** ☐ Cannot determine without additional information

(d) (1 point) Draw a sigmoid function that approximates the data. For threshold 0.5 indicate the decision boundary on the horizontal axis.

6. **Regular Expressions**

(a) (2 points) What is the output of the following piece of code?

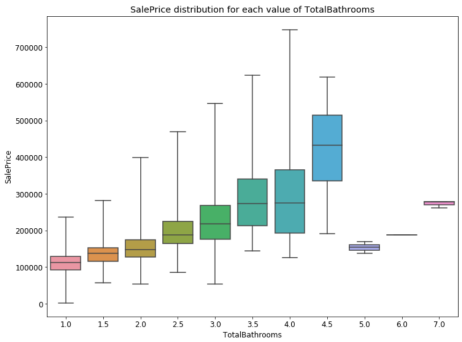
```
import re

def parseString(string):
    regex = r"([a-z0-9]+)"
    match = re.findall(regex, string)
    return len(match)

a = parseString("bB8")
b = parseString("r2d2")
c = parseString("C3PO")
print(a,b,c)
```

Solution: 2 1 1

7. **Features and Visualization**

(a) (1 point) Suppose we have a dataset about real estate like in Homework 6 on housing prices. Below we have boxplots of the distribution of house price divided by number of bathrooms.

__F__ **True or False:** The boxplots suggest dropping the feature from the dataset because houses with 5,6, or 7 bathrooms do not fit the linear trend.



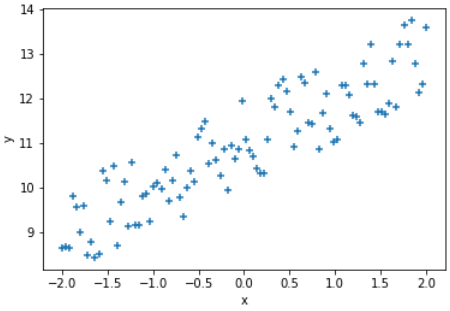
(b) (1 point) You are trying to build a regression model. Suppose we have the following plot which describes the data i.e. relationship between x and y. Which model amongst the options given below would you choose?

☐ Linear model without intercept term

☐ Linear model with quadratic feature

☐ Linear model with log-transformed feature

■ **Linear model with intercept term**



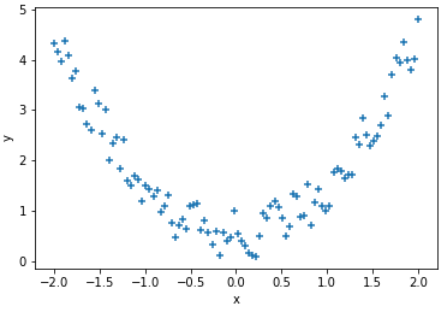
(c) (1 point) You are trying to build a regression model. Suppose we have the following plot which describes the data i.e. relationship between x and y. Which model amongst the options given below would you choose?

☐ Linear model with cubic feature

☐ Linear model with L2 regularization

■ **Linear model with quadratic feature**

☐ Linear model with bias term



(d) (1 point) Suppose you have a feature called `pets` containing categorical data. Below we have a summary of the values. You want to use a one-hot encoding. If the table contains a column for intercept term consisting of the value 1, then how many columns should you add to the table?

☐ 1 ■ **2** ☐ 3 ☐ 4

	pets
count	35
unique	3
top	dog
freq	20

8. **Regularization and Validation**

(a) Suppose you are fitting a model parameterized by θ using a regularized loss with regularization parameter λ. Indicate which error you should use to complete each of the following tasks.

i. (1 point) To optimize θ you should use the:

■ **Training Error** ☐ Cross-Validation Error ☐ Test Error

ii. (1 point) To determine the best value for λ you should use the:
☐ Training Error ■ **Cross-Validation Error** ☐ Test Error

iii. (1 point) To evaluate the degree of polynomial features you should use the:
☐ Training Error ■ **Cross-Validation Error** ☐ Test Error

iv. (1 point) To evaluate the quality of your final model you should use the:
☐ Training Error ☐ Cross-Validation Error ■ **Test Error**

(b) Suppose we have the following chart comparing error on the folds of the cross validation sets to the error on the training set. Note that we have applied a log transformation to the errors.

i. (1 point) **__T__ True or False:** We **expect** the curve for cross validation to lie above the curve for training.

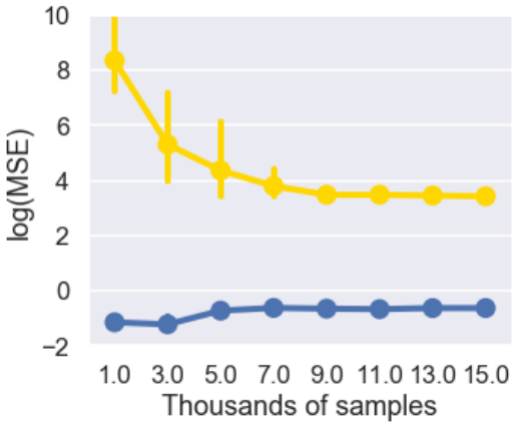
ii. (1 point) Would you suggest any of the following approaches to decrease the difference between error on the validation set and error on the training set? Select all that apply

■ **Increase the amount of samples**

■ **Reduce the number of features**

■ **Incorporate regularization**

☐ Use stochastic gradient descent instead of batch gradient descent



(c) (1 point) How can we prevent against underfitting the training set? Select all that apply.

■ **Choose a loss function sensitive to outliers**

☐ Discard any feature from the training set containing duplicate values

■ **Reduce the extra parameter λ that multiplies the regularization term**

☐ Change the split to make the training set larger and the testing set smaller.

■ **Incorporate additional parameters into the model through inclusion of more features**

9. **Loss Functions and Gradient Descent**

(a) (1 point) Assume you are implementing batch gradient descent. At each iteration, you calculate the gradient using the entire dataset. Below we have Figure A and Figure B. Which one has large learning rate and which has small learning rate?

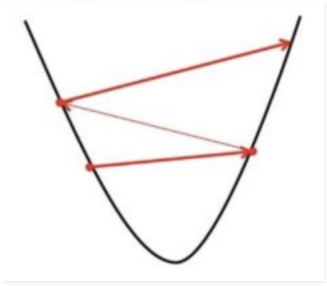


Figure A

Figure B

■ **Figure A has large rate and Figure B has small rate**

☐ Figure B has large rate and Figure A has small rate

☐ Both have equal learning rates

☐ Not enough information

(b) (2 points) Momentum is a variation of gradient descent where we include the gradient at a previous iteration in the current iteration. The update rule is

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\left(\theta^{(t)}\right) - \gamma \frac{\partial L}{\partial \theta}\left(\theta^{(t-1)}\right)$$

Here $\gamma > 0$ is the learning rate for the additional term. Assume for iteration $t = 0$ and $t = -1$, we set $\theta^{(t)} = \texttt{t0}$ the initial guess. Fill in the twelve blanks in the code with the following variables to implement gradient descent with momentum. Note that the same variable can be used multiple times. Some variables may not be used at all. Only use one variable per blank.

theta	phi	y	theta_prev	num_iter	t0
temp	alpha	gamma	range	len	t

```
def grad(phi, y, theta):
    """Returns dL/dtheta. Assume correct implementation."""

def grad_desc_momentum(phi, y, num_iter, alpha, gamma, t0):
    """ Returns theta computed after num_iter iterations.
    phi: matrix, design matrix
    y: vector, response vector
    num_iter: scalar, number of iterations to run
    alpha: scalar, learning rate
    gamma: scalar, weight of momentum
    t0: theta for t=0"""

   theta, theta_prev = ____________, ____________
    for __________ in __________(____________):
        g = grad(phi, y, theta)
        m = grad(phi, y, ____________)
        _______, _________ = _________ - _________ * g - ________ * m, __________
    return theta
```

END OF EXAM – PRESENT YOUR NYU ID AT SUBMISSION