INFO 2950 Fall 2022 Final

Do not turn the page until you are instructed to do so.

Instructions

Some students are taking the exam late due to scheduling constraints. Do not discuss the exam unless you are certain that everyone you are talking to has taken it.

You have 120 minutes to complete this exam. Time will be announced and marked on the board. You may use only a writing utensil and paper. If you use any electronic device for any purpose we will immediately confiscate your exam paper. All calculations have been constructed so that you will not need a calculator.

Write answers only in the assigned space on the answer sheet. Make sure that multiple choice answers are clearly circled; if there is any uncertainty as to your selection we will mark the problem as incorrect. When you are done, we will collect your answer sheet ONLY. Graders will see the segment of the answer sheet allocated for each problem and nothing else.

Make sure your name and netid are clearly written on every page of the answer sheet, as we will remove staples to scan it.

The answer sheet is intended to provide more than enough space; don't worry if you don't fill it. Showing your work may allow us to give you partial credit. Do not spend more than 10 minutes on a problem. If you get stuck, move on and come back later.

Raise your hand if you would like to ask a clarifying question.
Good luck!

$$var(X) = \frac{\sum_i (X_i - \bar{X})^2}{N}$$

$$cov(X,Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$ (not the same σ!)

A. Multiple Choice (40 points)

1. What is the covariance of x with itself?
   a. variance
   b. 1
   c. z-score
   d. 0

2. Compare the SQL and Pandas expressions on df1 and df2. Will the output of the Pandas and SQL code be (A) different, or (B) the same?

   Pandas:

```
df1.join(df2.set_index('item'), how = "inner", on = "item")
```

   SQL:

```
SELECT df1.item, df1.wegmans_price, df2.aldi_price
FROM df1 LEFT JOIN df2 ON df1.item = df2.item
```

df1

item	***wegmans_price***
Broccoli	1.00
Tomatoes	2.89
Cucumbers	0.67
Peppers	1.09
Beets	3.49
Yams	0.59

df2

item	***aldi_price***
Broccoli	2.61
Tomatoes	1.75
Cucumbers	0.85
Peppers	3.55
Yams	2.75

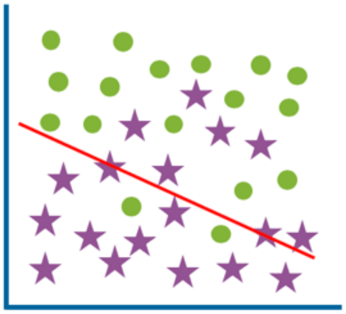
3. Circle all that apply: What are some reasonable things that you could do to choose your input variables to a regression?
   a. check residual plot for randomness
   b. use SVD for feature selection
   c. use a random 10% sample of variables
   d. check for collinearity with correlation matrices
   e. use domain expertise

4. The following code splits a dataframe into train and test sets, and then performs an additional split to create a validation set. How many rows are in the validation set?

```
import pandas as pd
from sklearn.model_selection import train_test_split

df = pd.read_csv('data.csv')
# df.shape ---> (1000, 5)

X = df[['feat-1', 'feat-2', 'feat-3', 'feat-4']]
y = df[['label']]

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.20, random_state=42)
X_train, X_val, y_train, y_val = train_test_split(X_train,
y_train, test_size=0.25, random_state=42)
```

a) 42
b) 200
c) 400
d) 800
e) 1000

5. Circle all that apply: Which of these are likely examples of overfitting on the train set?
   a. You've reached 99.99999% accuracy.
   b. You perform a study on sleep habits of the general population, but only include left-side sleepers in your training set.
   c. Your model fits the training set so well that it can no longer fit to new observations reliably.
   d. The RMSE is much lower for the test set than the training set.
   e. You see this figure:



6. Given the following dataset and three samples from that dataset, were these samples used for (A) a permutation test, or (B) a bootstrap test?

Dataset:

5.7	3.9
1.2	2.3
4.8	4.4

Sample 1:

5.7	3.9
1.2	2.3
1.2	2.3

Sample 2:

5.7	3.9
1.2	2.3
5.7	3.9

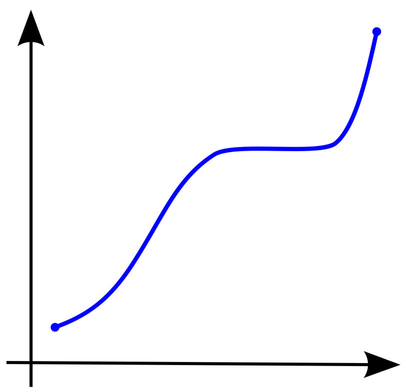
Sample 3:

5.7	3.9
5.7	3.9
4.8	4.4

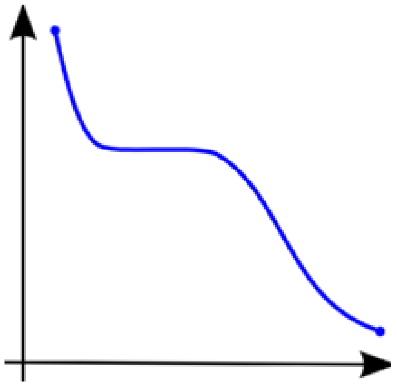
7. Match the distribution type to the situation it would model best:
a. Binomial, b. Negative binomial, c. Geometric, d. Poisson

<u>Situation 1</u>: You ask students if they prefer cats or dogs as pets, until you find a student who prefers cats.
<u>Situation 2:</u> You practice basketball free throws and set a goal of reaching at least 10 successful shots before you're finished practicing.
<u>Situation 3</u>: You track the number of TCAT buses that arrive between 5:00-6:30 p.m.
<u>Situation 4</u>: You flip a coin 20 times and track the number of times the coin lands on heads.

8. Fill in the blanks with one of the following: (a) -1, (b) -0.8, (c) 0.8, (d) 1



A. Pearson correlation: ____________
B. Spearman (rank) correlation: __________



C. Pearson correlation: ____________
D. Spearman (rank) correlation: __________

9. Below, we've provided out-of-order pseudocode for the k-means clustering algorithm. Write the numbers of the 5 steps in the correct order.

```
k: the number of clusters,
D: a data set containing n objects.
—-----------------------------------

(1) end loop if no change;
(2) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(3) arbitrarily choose k objects from D as the initial cluster centers;
(4) start a loop;
(5) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
```

10. You run 1000 experiments and calculate the p-value for each experiment.

    A. If the null hypothesis is true, how many of those p-values would you expect to find to be less than 0.05?

    a. 0
    b. Around 50
    c. Around 500
    d. Around 995

    B. What distribution would best model the p-values across experiments?

    a. Geometric
    b. Normal
    c. Poisson
    d. Negative binomial

    C. What is an example of a way to statistically correct for doing a large number of experimental comparisons? (Not multiple choice; write fewer than 3 words).

B. Short Answer (36 points)

1. You have decided to run a log-linear model.
   (a) Does your data look like figure A or B? ______

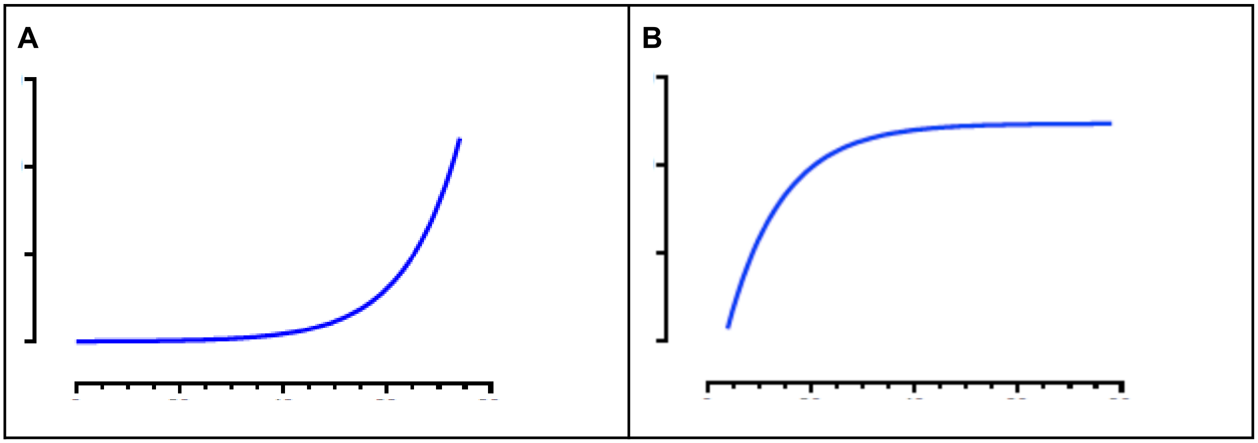


Image source: https://www.graphpad.com/guides/prism/latest/user-guide/when_to_use_a_logarithmic_axis.htm

   (b) Fill in the blanks to derive how an increase in input *x* by 1 would affect the output *y*.
   Step 1: log(y) = a+b*x
   Step 2: Define new variable x_{new} = x+1
   Step 3: Define new variable y_{new} so that $\log(y_{new})$ = a+b*x_{new}
   Step 4: Plug in step 2 to step 3 so the right-hand side is in terms of x:
   $\log(y_{new})$ = _____________
   Step 5: Plug in step 1 to step 4 so the right-hand side is in terms of y:
   $\log(y_{new})$ = _____________
   Step 6: Use logarithmic rules to exponentiate and simplify from step 5:
   y_{new} = ___________
   Step 7: Describe in a sentence how to interpret the change in *y* if you increase *x* by 1.

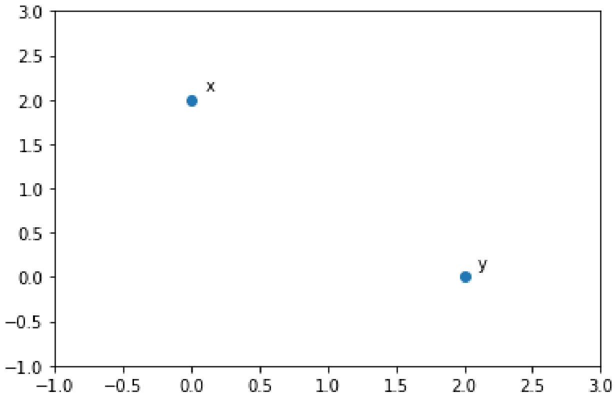
2. When playing basketball, Grace Hopper makes 1/4 of the free throws she attempts. As a warm up, Grace Hopper likes to shoot until she makes a successful free throw. Let **K** be the number of throws it takes Grace Hopper to successfully make her first free throw. Assume that the results of each throw are independent.

   Find the probability that it takes Grace Hopper strictly fewer than 3 attempts (i.e., it takes at most 2 attempts) to make her first free throw. Express your answer as a fraction.

3. The US Social Security Administration records the frequency of US baby names. In 2021 they recorded 6.6M births. The name "Galaxy" appears 66 times.
   a. What is the probability that a randomly selected 2021 US baby would be named Galaxy?
   b. What is the (natural) log probability?

4. It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that, given a spam email, it has a 90% probability of detecting it as spam. And, their probability for a false positive (given a non-spam email, it is detected as spam) is 10%.
   (A) What probabilities are stated in the problem statement above? Write at least three statements in the format P(___) = ___ or P(__ | __) = ___.
   (B) Write the following problem statement in P(___ | ____) form: Given that an email is detected as spam, then what is the probability that it is in fact a non-spam email?
   (C) Calculate the answer to the above question. Show your work.

5. Calculate the Manhattan, Euclidean, and Cosine distances between points x and y in the following plot. Recall that cosine distance is defined as 1 minus cosine similarity. Cosine similarity is defined as $\cos(\theta) = A \cdot B / (||A||\ ||B||)$. You are allowed to leave your solution in mathematical terms.



6. The matrix **A** $\in \mathbb{N}_0^{100 \times 20}$ represent data on 100 Instacart shoppers (rows) and how many times they buy the 20 most popular grocery items (columns). **A** can be decomposed into three matrices U, S and V in a way such that $A = USV^T$. We find a rank 10 SVD.
   a. What do U, S and V^T represent? (at an abstract level, write at most one sentence per matrix).
   b. If we only wanted to compute the rank 2 approximation, we could truncate U, S, and V^T to create new matrices U', S', and V'^T and multiply them to get the new approximation A'. What are the shapes of each: U', S', and V'^T?

7. Using Naive Bayes, what is the numerator of the probability that "meow" was said by a dog? (Ignore the denominator used for normalizing.) Use the Laplace correction. Express your answer as a fraction.

Text	Category
"bark woof"	dog
"woof bark woof woof"	dog
"meow meow"	cat

C. Long Answer (24 points)

1. The following is a logistic regression output table where the binary outcome variable is whether a patient is expected to develop lung disease in the next five years (1: develops lung disease, 0: does not develop lung disease). The input features for this logistic regression model are: whether the patient is a smoker or not (*Smoker*, binary), the Air Quality Index (*AQI*, a numeric metric that runs from 0 to 500, where 0 is good and 300+ is hazardous) in the area the patient lived in for the last 5 years, and whether the patient exercises five times a week as opposed to not exercising five times a week (*Exercise*, binary). Note: these results were run on synthetic data.

Term	**Coefficient**	**Standard Error**	**p-value**
Constant	-3.22	0.17	0.00
Smoker	2.91	0.076	0.00
AQI	0.008	0.0005	0.09
Exercise	-0.11	0.067	0.00

   a. Interpret the regression (use our usual framework, and specify units and reference variables; an exponentiated lookup table for different numbers is provided below).

   b. Which **β** coefficients are significant at the 5% significance level?

x	***e^x***	***$e^x/(1+e^x)$***
-3.22	*0.04*	*0.038*
-0.11	*0.90*	*0.47*
0.00	*1*	*0.5*
0.008	*1.008*	*0.502*
0.17	*1.19*	*0.54*
2.91	*18.36*	*0.95*

2. Prof. Mimno has noticed that Sparky seems quite good at doing backflips. Sparky successfully landed 60 backflips out of 192 backflip attempts. Across all other cats in Ithaca, 800 backflips were attempted and 200 were successful.
    a) Define a null and alternative hypothesis based on this scenario that you could answer.
    b) Assuming a binomial distribution, how many standard deviations is Sparky away from the average Ithaca cat in backflip successes?
    c) What percentage of Ithaca cats do you estimate are within that many standard deviations of the mean?
    d) Make a one-sentence argument for whether or not you would reject the null hypothesis.

1. You work at a social media website and are considering switching from using an endless scroll to implementing a “next” button to move page to page. You decide to perform A/B testing. In this experiment, write one sentence to answer each of the following questions:
    a. What is A?
    b. What is B?
    c. What metric would you use?
    d. Name one pitfall you would try to avoid.

D. Extra Credit (5 points)

1. Which of the following strings would match this regular expression? '[rgd]i[a-z]en'
    a. "Riven"
    b. "risen"
    c. "given"
    d. "dozen"
    e. "giver"

2. Which of the following are reasons why singular value decomposition (SVD) is the best matrix decomposition?
    a) It helps with interpreting important concepts.
    b) You can approximate any matrix.
    c) It can be used to fill in random missing elements in a dataframe.
    d) It allows you to reduce high-dimensional data.