# Python

1. **How do you convert a string to a number and vice-versa?**

*# initializing string*

s = "10010"

e = float(s)

print(type(e))

e\_s = str(e)

print(type(e\_s))

<class 'float'>

<class 'str'>

1. **How do you combine two one-dimensional lists into one one-dimensional list?**

a = [1,2,3]

b = [4,5,6]

print(a + b)

[1, 2, 3, 4, 5, 6]

1. **What is a Python module?**

A module is a file consisting of Python code that defines functions, classes and variables.

1. **What is PEP-8?**

It is a style guide providing best practice on how to write Python code.

1. **Write a function that asks the user for a string, then print whether the string is a palindrome. A palindrome is a string that reaads the same forwards and backwards.**

def list\_ends(a\_list):

return [a\_list[0], a\_list[len(a\_list)-1]]

1. **What is the difference between a for loop and a while loop?**

A for loop is used to iterate over elements of a sequence. A while loop instructs the computer to do something so long as a certain condition is met.

1. **What is the output of -12 // 10?**

print(-12//10)

-2

1. **Write a function that asks the user for a number and then prints out a list of all the divisors of that number** num = int(input("Please choose a number to divide: "))

listRange = list(range(1,num+1))

divisorList = []

for number in listRange: if num % number == 0: divisorList.append(number)

print(divisorList)

# SQL

1. **What does the HAVING clause do in SQL?**

Applies criteria to the results of an aggregate, as defined with a GROUP BY clause.

1. **What is the difference between an inner join and a left outer join?**

An inner join returns records with a match in both tables A and B, while a left outer join returns all records found in table A, regardless of a match in table B. If no match is found the resulting join value is NULL.

1. **What does the LIMIT clause do in SQL?**

Limits the results of the SQL query to the first n records.

1. **What is an RDBMS?**

A relational database management system maintains data records and fields in tables. Relationships between data entities are expressed in these tables.

# Statistics/ML/Data Science

1. **What is backpropagation?**

A technique for training neural networks that uses gradient desent to calculate the loss function at output, and distribute it back through the neural network, resulting in adjusted weights for neurons.

1. **What are some different ways to treat missing values?**

There are lots of factors here. If there are fewer than ~5% of missing values, then it is likely safe to proceed with the analysis. If there are more than 80%, then affected variables should be dropped. For a continuous variable, missing values can be imputed with the mean or median of the variable, or interpolated by populating missing values using representative values from similar rows.

1. **What is the difference between artificial intelligence and machine learning?**

Artificial intelligence is a broader concept. The goal of AI is to simulate intelligence to solve complex problems in a way that seems naturally "smart." Machine learning focuses on the ability of a machine to take a set of data and learn for itself, changing algorithms as more is learned.

1. **What are some metrics for evaluating the performance of a linear regression?**

Metrics include r-square, mean absolute error, mean square error, mean absolute percentage error and mean percentage error.

1. **What is underfitting and how do you prevent it?**

Underfit model does not model the training well and does not generalize to new data. More data and more features will help prevent underfitting.

1. **What is a nonparametric model? What are its advantages and disadvantages?**

A non-parametric model is one with no assumptions about a variable's distribution. While this makes it a more flexible model, it provides less powerful insights.

1. **What is the difference between statistics and machine learning?**

The focus of statistics is on establishing causal relations by using hypothesis-driven data analysis, whereas machine learning uses non-explicitly programmed algorithms usually with an emphasis on predictive accuracy.

1. **When might you use a random forest versus a support vector machine and why?**

For a multi-class classification problem, an SVM would be more memory-intensive but would take longer to tune. For a semi-supervised learning problem such as dissimilarity measure, random forest would be choice.

**The Fibonnaci sequence is a sequence of numbers where the next number in the sequence is the sum of the previous two numbers in the sequence. The sequence looks like this: 1, 1, 2, 3, 5, 8, 13, … Write a function that asks the user how many Fibonnaci numbers to generate and then generates them.**

def gen\_fib():

count = int(input("How many fibonacci numbers would you like to generate? "))

i = 1

if count == 0:

fib = []

elif count == 1:

fib = [1]

elif count == 2:

fib = [1,1]

elif count > 2:

fib = [1,1]

while i < (count - 1):

fib.append(fib[i] + fib[i-1])

i += 1

return fib

1. **What is an autoencoder?**

This is a type of neural network with the purpose to reconstruct their original output. It is a form of feature extraction and these features are used in the middle layer of tne neural network.

1. **What is bagging and boosting in an ensemble model?**

Bagging and boosting produce new training data sets by random sampling with replacement from the original data set. In bagging, any element has the same probability to appear in a new dataset, while for boosting obnservations are weighted.

1. **What is a loss function in a neural network?**

The loss function describes the magnitude of error a the network made its prediction. The error can then be backpropagated through the model, adjusting weights to get a more predictive answer during the next run.

1. **What is regularization and why is it useful?**

Regularization is used to prevent overfitting and improve the generalization of a model. It introduces a regularization term to the general loss function. Ridge regression and LASSO are the common regularization methods.

1. **Explain a false positive and false negative. When might one be more important than another?**

A false positive means improperly reporting the presence of a condition when it's not actually there. A false negative is improprrly reporting the absence of a condition when it actually is there.

A false positive would be more important than a false negative in the case of a non-contagious disease where treatment delay doesn't have any long-term consequences but the treatment itself is very difficult.

A false negative is more important than a false positive when early treatment is essential for good outcomes, and not to catch it could have serious implications at a later point.

1. **What is the binomial distribution? What are some examples of phenomena that follow a binomial distribution?**

A binomial distribution represent the distribution of the number of successes when an event with two or more distinct possible outcomes is repeated many times.

Examples of the binomial distribution in real life include coin or die tosses or the results of a sports game.

1. **Write a function that takes a list and returns a new list containing all unique elements of the first list (i.e., no duplicates).**

*# Exercise 13:*

*# Write a function that takes a list and returns a new list that contains*

*# all the elements of the first list minus duplicates.*

*# this one uses a for loop*

def dedupe\_v1(x):

y = []

for i in x:

if i not in y:

y.append(i)

return y

*#this one uses sets*

def dedupe\_v2(x):

return list(set(x))

a = [1,2,3,4,3,2,1]

print a

print dedupe\_v1(a)

print dedupe\_v2(a)

1. **Why is area under the curve a better evaluation metric than raw accuracy?**

Raw accuracy is based on a specific cutpoint, whereas the ROC area under the curve looks at the trade-offs between sensitivity and specifity at all cutpoints.

1. **What is the bias/variance tradeoff?**

If a model is too simple and has too few parameters then it may have high bias and low variance. If a model has large number of parameters, it is likely to have high variance and low bias. The goal is to manage these trade-offs without overfitting and underfitting the data.

**Make a two-player Rock-Paper-Scissors game. Ask the players to input plays, compare them, then print a message of congratulations to a winner and ask if the players want to play again.**

import sys

user1 = input("What's your name?")

user2 = input("And your name?")

user1\_answer = input("%s, do yo want to choose rock, paper or scissors?" % user1)

user2\_answer = input("%s, do you want to choose rock, paper or scissors?" % user2)

def compare(u1, u2):

if u1 == u2:

return("It's a tie!")

elif u1 == 'rock':

if u2 == 'scissors':

return("Rock wins!")

else:

return("Paper wins!")

elif u1 == 'scissors':

if u2 == 'paper':

return("Scissors win!")

else:

return("Rock wins!")

elif u1 == 'paper':

if u2 == 'rock':

return("Paper wins!")

else:

return("Scissors win!")

else:

return("Invalid input! You have not entered rock, paper or scissors, try again.")

sys.exit()

print(compare(user1\_answer, user2\_answer))

1. **What is selection bias?**

Selection bias indicates that a sample is not representative of the intended population, due to sampling, non-response, or other factors.

1. **What are precision and recall?**

Precision measures the portion of positive identifications in a classification set that were actually correct, whereas recall measures the proportion of actual positives that were identified correctly.

1. **What is a perceptron?**

A perceptron is an algorithm for supervised classification of the input into one of several possible non-binary outputs.

1. **Write a function that takes a list of numbers and returns a list of only the first and last elements of that list.**

def list\_ends(a\_list):

return [a\_list[0], a\_list[len(a\_list)-1]]