1. **Intro to Data Science**

What are the roles of data scientists?

* Data scientists analyze diverse data, understand unstructured data, and find order, value, and meaning in the data with various computer science, mathematics, and statistics-related concepts.

What are the steps of a collaborative project in data science?

1. Planning
2. Data Preparation
3. Modeling
4. Follow up

How would you compare the importance of technical skills vs the domain knowledge

* Domain knowledge is the collection of skills and expertise in a specific field. Meanwhile, technical skills involve the ability to perform specific job-related tasks.

Explain the “velocity” in Big data

* **Velocity:**  Data is being generated very quickly.

Is there some overlap between big data and data science? Explain your reasoning

* Yes, there is an overlap between big data and data science since big data is a more specialized field of data science. Specifically, data science can be done in small data sets; not everything done using Big Data would be considered data science.

Why Python and R are the most used languages for data science

* Both languages

What is the difference between a data scientist and a statistician?

* Data science requires statistics but is not a subset.
* Both fields use data but have different motivations and goals, different backgrounds and context.

1. **Data Science Overview**

What are the differences between supervised, unsupervised, and reinforcement learning? Can you guess what kind of learning might be used for chatGPT ?

* Supervised learning: using data to predict an outcome.
* Unsupervised learning: using data to group items into categories.
* Reinforcement learning: optimizing action based on a response variable.
* ChatGPT uses a combination of supervised learning and reinforcement learning.

What is A/B testing, and provide example

* A/B testing is essentially an experiment where two or more-page variants are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.
* Example: Amazon recommendation page. Use two different recommendation algorithms and see which leads to higher conversion.

What are the differences between linear regression and decision tree

* Decision trees support nonlinearity while linear regression only supports linear solutions.

Explain in brief: Checking for quality issues in data science– completeness, fidelity and consistency

* Completeness:
  + check whether the data is representative of the problem.
  + Check missing observations, attributes
* Fidelity:
  + Do the measurements capture the reality
  + Any issues of bias or variance
* Consistency:
  + Are values follow data types specified
  + Do different attributes agree with each other

1. **Machine Learning**

What are different examples of regression and classification? What are the different algorithms to use for these methods?

* Regression algorithms can be used to solve regression problems such as weather prediction, house price prediction, etc
  + Ex: Linear regression, logistic regression, etc
* Classification problems are used to solve problems such as identification of emails, identification of cancer cells, etc.
  + Decision Tree, K-Nearest Neighbor

Examples of supervised, unsupervised and reinforcement learning. What are different algorithms in each of them

* Supervised learning: Iris classification, image classification, etc
  + Linear Regression, Support Vector Machine, etc
* Unsupervised learning: Clustering, Anomaly detection, etc
  + K-Means Clustering, Mean-shift clustering
* Reinforcement learning: Gaming bots,
  + Q-learning
  + R learding

What is cross-validation, and why is it important

* Cross-validation is a technique used to evaluate the performance of a machine learning model on a dataset.
* Cross-validation help us get a more accurate estimate on how the model perform on new, unseen data through splitting, training, and testing the data in different “folds”.

What is hyperparameter optimization

* Hyperparameter is a parameter whose value is set before the training process begins and is not learned from the data. Thus, different models can perform differently based on the pre-given hyperparameters. Optimizing the hyperparameters also mean optimizing the end model.

Explain the 5 steps of approaching an application in machine learning

1. Define the problem to be solved
2. Collect and label data
3. Choose an algorithm class
4. Choose an optimization metric for learning the model
5. Choose a metric for evaluating the model

What is an objective function in machine learning, and what is its role

* Objective function maximize the posterior probabilities (basically maximizing the end goal performance of a model)

What are the differences between eager and lazy learning algorithms? State with examples

* Lazy: Model is not built during the training phase but instead during prediction phase. This means the algorithm stores the training data in memory and waits until it receives a new input before computing a prediction.
  + Ex: K-Nearest Neighbor
* Eager: Model is built during the training phase and then used to make predictions on new, unseen data. This means that the algorithm generates a model from the training data that can be used to make predictions on new data without having to store the entire training set.
  + Decision Tree, SVM, etc

What are the differences between batch and online learning algorithms? State with examples

* Batch: Learning over groups of patterns
  + SVN, Decision Tree, etc
* Online: Learning based on each pattern as it is observed
  + Reinforcement learning

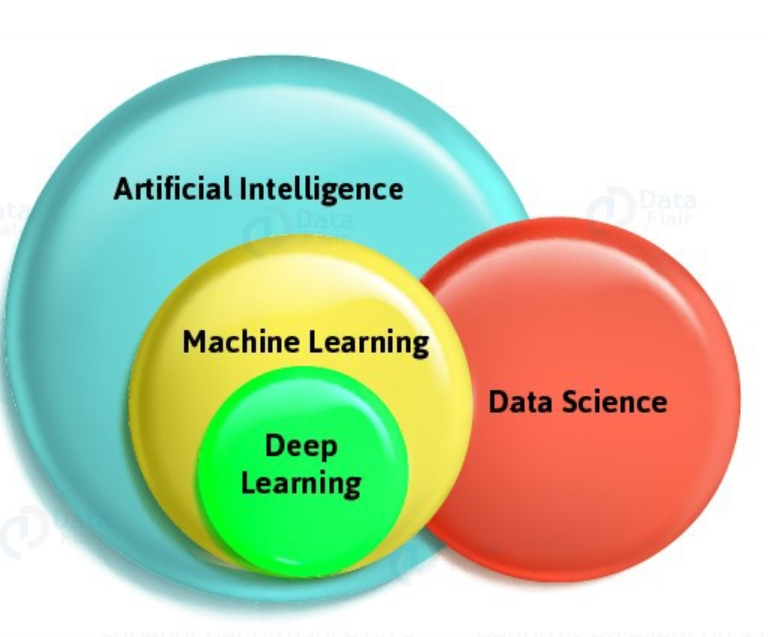
What are the differences between parametric and nonparametric learning algorithms? State with examples

* A non-parametric algorithm uses a flexible number of parameters, and the number of parameters often grows as it learns from more data. A non-parametric algorithm is computationally slower: KNN
* A parametric algorithm has a fixed number of parameters. A parametric algorithm is computationally faster, but can be more flexible with the data: Linear Regression

What are the differences between discriminative and generative learning algorithms? State with examples

* Discriminative algorithms model the boundary between different classes directly, without attempting to model the underlying probability distribution of the data. In other words, they focus on learning a decision boundary that separates different classes in the input space.
  + Ex: SVM
* Generative algorithms, on the other hand, model the underlying probability distribution of the data for each class and use this to make predictions. In other words, they model the distribution of each class in the input space and then use Bayes' rule to calculate the probability of each class given a new input.
  + Naïve Bayes

Draw a ven diagram showing how Artificial intelligence, Machine Learning, Deep Learning and Data science interact



1. **Statistical Learning and Linear Regression**

What are they key differences between a population line and the least squares line ?

* A population line is a theoretical line that represents the true relationship between the two variables in the entire population.
* On the other hand, a least squares line is an estimate of the population line that is calculated based on a sample of data.

How to interpret the 𝑅2 value? What does it mean when 1 and 0 ?

* R2 is a statistical measure that represents the proportion of variance in the dependent variable (y) that can be explained by the independent variables (x) in a regression model.
* The R-squared value ranges from 0 to 1, where 0 indicates that the model does not explain any of the variance in the dependent variable and 1 indicates that the model perfectly explains all the variance in the dependent variable. An R-squared value of 0.7, for example, indicates that 70% of the variance in the dependent variable can be explained by the independent variables in the model.

What is the interpretation of p-value in following table?

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). **A low p-value (< 0.05) indicates that you can reject the null hypothesis**

What is one-way ANOVA test, and what is it used for ?

* One-way ANOVA (analysis of variance) is a statistical test used to determine if there is a significant difference in the means of three or more groups.
* It is used to compare the means of multiple groups and determine whether any of the differences between the groups are statistically significant.

What are the assumptions for linear regression? What is Multicollinearity

* The assumption for linear regression is there is a linear relationship between the independent variable and dependent variable (linearity), No or little multicollinearity, Independence, normality, and homoscedasticity
* Multicollinearity is when two or more independent variables are highly correlated with each other. We can observe multicollinearity using the heat map.

When would you find multiple lines in a linear regression problem ? What is the interpretation when these lines are not parallel?

* Multiple lines in a linear regression problem can be found when there are one or more categorical independent variables that divide the data into distinct groups.
* If the lines are not parallel, it means there are multiple variables that are linearly related with each other, whether it is negative or positive.

1. **Data Processing and Cleaning**

What is data and attribute?

* Data is a collection of data objects and their attributes
* An attribute is a property or characteristic of an object

What are different types of attributes? What is the difference between ordinal and nominal attributes?

* Nominal: eye color, zip code (string non-related)
* Ordinal: rankings (string related in an order)
* Interval: temperatures
* Ratio: length, time
* Ordinal is based on a specific ranking order while nominal is not based on a ranking order

What is graph data? Provide an example

* Graph data is data represented in the form of a graph.
* Scatter plot, histogram, box plot, etc.

What is noise and outlier in the data ? Provide examples

* Noise: unwanted or irrelevant information or variability of a dataset
* Outliers: data objects with characteristics that are considerably different than most of the other data objects in the data set.

Explain different techniques in handling missing values for different types of attributes

* Using the mean
* Using mode
* Eliminate missing rows

What is the relation between correlation and covariance? **Provide examples**

* Correlation and covariance are both measures of the relationship between two variables. Correlation measures the strength and direction of the linear relationship between two variables, while covariance measures the direction and strength of the linear relationship between two variables in terms of their deviation from their respective means.

Explain curse of dimensionality. What are different techniques to reduce dimensionality. Why is it important to reduce dimensionality

* As the number of dimensions or features in a dataset increases, the amount of data required to maintain the same level of statistical significance increases exponentially, making it more difficult to analyze and model the data effectively
* Methods to reducing dimensionality: Principal Component Analysis (PCA), Feature Selection, etc
* PCA transforms the data into a new coordinate system that maximizes the variance of the data along each axis. This results in a smaller set of principal components that capture most of the variation in the original data.

What are binarization and discretization?

* Transform continuous numerical data into discrete data or binary data through bining.
  + Ex: Age < 27 and Age > 27 => 2 bins

Why normalization is done on the data?

* The goal of normalization is to ensure that all features are treated equally by the machine learning algorithm and to improve the convergence rate and performance of the algorithm.

Explain standardization normalization, and min-max normalization

* Standard normalization uses the Z-scaled with mean 0 and based 1 to normalize data while min-max utilize the min and max of all attributes in the data set.

How would you determine if a categorical or numerical attribute is redundant?

* I can either calculate the chi-squared value or use heatmap

What is entropy-based binning

* Bin continuous data into discrete value or binary value based on concept of entropy

1. **Data Exploration**

How is AAD measured, and when it is used

* AAD (Average Absolute Deviation) is a measure of the average distance between the predicted and actual values of a regression model.
* Often used as evaluation metric for regression model

Why contour plots are needed

useful for visualizing continuous functions and their level sets or contours.

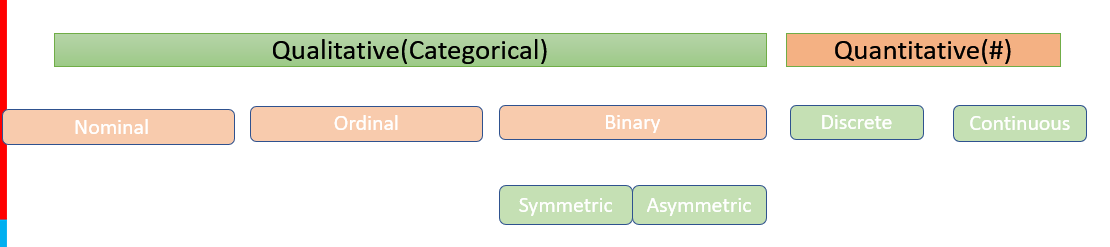
1. Data Similarities and Distances and KNN

What is similarity, dissimilarity and proximity

* Similarity: numerical measure of how alike two data objects are
* Dissimilarity: numerical measure of how different two data objects are
* Proximity: refers to a similarity or dissimilarity

What is qualitative and quantitative data

* Qualitative: Categorical (Ordinal, nominal)
* Quantitative: numerical (ratio, interval)



Why mean absolute deviation is a better approach to handle outliers in a data than standard deviation

* MAD is basing calculation on the median while standard deviation is based on the mean. Thus, it is more influenced by outliers

Why logarithmic transformation is useful

* Compression
* Linearization of relationship
* Reduce skewness