MACHINE LEARNING

1 What is logistic regression in Data Science?

Logistic regression is **a statistical analysis method used to predict a data value based on prior observations of a data set**. ... A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

2 Name three types of biases that can occur during sampling

1. Sample Bias

We all have to consider sampling bias on our training data as a result of human input. Machine learning models are predictive engines that train on a large mass of data based on the past. They are made to predict based on what they have been trained to predict. These predictions are only as reliable as the human collecting and analyzing the data. The decision makers have to remember that if humans are involved at any part of the process, there is a greater chance of bias in the model.

The sample data used for training has to be as close a representation of the real scenario as possible. There are many factors that can bias a sample from the beginning and those reasons differ from each domain (i.e. business, security, medical, education etc.)

2. Prejudice Bias

This again is a cause of human input. Prejudice occurs as a result of cultural stereotypes in the people involved in the process. Social class, race, nationality, gender can creep into a model that can completely and unjustly skew the results of your model. Unfortunately it is not hard to believe that it may have been the intention or just neglected throughout the whole process.

Involving some of these factors in statistical modelling for research purposes or to understand a situation at a point in time is completely different to predicting who should get a loan when the training data is skewed against people of a certain race, gender and/or nationality.

3. Confirmation Bias

This is a well-known bias that has been studied in the field of psychology and directly applicable to how it can affect a machine learning process. If the people of intended use have a pre-existing hypothesis that they would like to confirm with machine learning (there are probably simple ways to do it depending on the context) the people involved in the modelling process might be inclined to intentionally manipulate the process towards finding that answer. I would personally think it is more common than we think just because heuristically, many of us in industry might be pressured to get a certain answer before even starting the process than just looking to see what the data is actually saying.

4. Group attribution Bias

This type of bias results from when you train a model with data that contains an asymmetric view of a certain group. For example, in a certain sample dataset if the majority of a certain gender would be more successful than the other or if the majority of a certain race makes more than another, your model will be inclined to believe these falsehoods. There is label bias in these cases. In actuality, these sorts of labels should not make it into a model in the first place. The sample used to understand and analyse the current situation cannot just be used as training data without the appropriate pre-processing to account for any potential unjust bias. Machine learning models are becoming more ingrained in society without the ordinary person even knowing which makes group attribution bias just as likely to punish a person unjustly because the necessary steps were not taken to account for the bias in the training data.

3 What is Decision tree Algorithm? What is ASM?

What is ASM in decision tree?

While building a Decision tree, the main thing is to select the best attribute from the total features list of the dataset for the root node as well as for sub-nodes. The selection of best attributes is being achieved with the help of a technique known as the Attribute selection measure (ASM)

What is a decision tree algorithm?

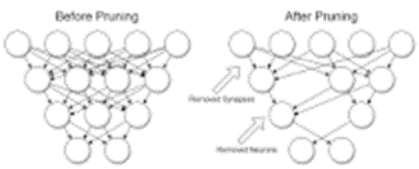
Image result for What is Decision tree Algorithm? What is ASM?

Decision Tree algorithm belongs to the family of supervised learning algorithms. ... The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data)

What is decision tree explain?

A decision tree is a tree-like model that acts as a decision support tool, visually displaying decisions and their potential outcomes, consequences, and costs. From there, the “branches” can easily be evaluated and compared in order to select the best courses of action.

What is tree pruning in data mining?



Pruning is a **data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the** tree that are non-critical and redundant to classify instances. ... A tree that is too large risks overfitting the training data and poorly generalizing to new samples.

4 Name three disadvantages of using a linear model

Main limitation of Linear Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.
Linear Regression Only Looks at the Mean of the Dependent Variable. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables. ...
Linear Regression Is Sensitive to Outliers.

Linear Regression is a great tool to analyze the relationships among the variables but it isn't recommended for most practical applications because it over-simplifies real-world problems by assuming a linear relationship among the variables.

What is linear model in machine learning?

The term linear model implies that the **model is specified as a linear combination of features**. Based on training data, the learning process computes one weight for each feature to form a model that can predict or estimate the target value.

Advantages

Linear Regression is simple to implement and easier to interpret the output coefficients.

When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity to compared to other algorithms.
Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages

On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.

Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.

But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

5 Why do you need to perform resampling?

Generally speaking, a resampling method is a tool consisting in repeatedly drawing samples from a dataset and calculating statistics and metrics on each of those samples in order to obtain further information about something, in the machine learning setting, this something is the performance of a model.

Both data sampling and data resampling are methods that are required in a predictive modeling problem. ... Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.

Informally, resample can mean something a little simpler: repeat any sampling method. For example, if you're conducting a Sequential Probability Ratio Test and don't come to a conclusion, then you resample and rerun the test. For most intents and purposes though, if you read about resampling (as opposed to "resample"), then the author is most likely talking about a specific resampling technique.

6 List out the libraries in Python used for Data Analysis

Scipy, Numpy, Pandas, Matplotlib, Seaborn, Scikit learn

7 What is bias?

The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.

8 What is a Linear Regression?

Linear Regression is **a supervised machine learning algorithm where the predicted output is continuous and has a constant slope**. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

9 What is Ensemble Learning?

Ensemble Methods in Machine Learning: What are They and Why Use Them? | by Evan Lutins | Towards Data Science

url https://bdtechtalks.com/2020/11/12/what-is-ensemble-learning/

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

10 Explain the steps for a Data analytics project

Understanding a business problem
Data exploration
Data preparation for modeling
Running the model and analysis of results
Model validation using new data sets

Understand business problem ...
Obtain and Understand Data. ...
Data Preparation. ...
Data Modelling. ...
Model Evaluation. ...
Deployment and Visualization

11 What is the K-means clustering method?

url(important) K-Means Clustering Algorithm - Javatpoint

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

12 When underfitting occurs in a static model?

Underfitting occurs when an arithmetical model or machine learning algorithm not able to capture the primary trend of the data.

Underfitting occurs when a model is too simple — informed by too few features or regularized too much — which makes it inflexible in learning from the dataset

13 What are the diagonistics of Classification

Classification studies using machine learning generally require four steps: feature extraction, feature selection, dimensionality reduction, and feature-based classification algorithm selection. These procedures require specialized knowledge and multiple stages of optimization, which may be time-consuming

14 What is Precision and recall?

url

url(imp)-definitions

Precision and recall - Wikipedia

Precision vs Recall | Precision and Recall Machine Learning (analyticsvidhya.com)

In the simplest terms, Precision is the ratio between the True Positives and all the Positives. For our problem statement, that would be the measure of patients that we correctly identify having a heart disease out of all the patients actually having it

The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease

Recall is the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search

precision and recall are performance metrics that apply to data retrieved from a collection, corpus or sample space.
Precision=TP/(TP+FP)

Recall = TP/(TP+FN)

15 Explain cluster sampling technique in Data science

url(imp)-concepts

Cluster Sampling: Definition, Method and Examples | QuestionPro
Cluster sampling is a probability sampling technique where researchers divide the population into multiple groups (clusters) for research. Researchers then select random groups with a simple random or systematic random sampling technique for data collection and data analysis

16 While working on a data set, how can you select important variables? Explain

url(imp)-concepts

https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
You can get the feature importance of each feature of your dataset by using the feature importance property of the model. Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable

Chi-square Test. ...
Fisher's Score. ...
Correlation Coefficient. ...
Dispersion ratio. ...
Backward Feature Elimination. ...
Recursive Feature Elimination. ...
Random Forest Importance

17 Treating a categorical variable as a continuous variable would result in a better predictive model?

yes