1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function's fitness assessed?

Answer: The target variable is the variable whose values are modeled and predicted by other variables. A predictor variable is a variable whose values will be used to predict the value of the target variable.The target variable of a dataset is the feature of a dataset about which you want to gain a deeper understanding. A supervised machine learning algorithm uses historical data to learn patterns and uncover relationships between other features of your dataset and the target. I have explained the basics about Genetic Algorithms. After it was published, I got many requests to discuss more about the Fitness Function and Evaluation Strategies. In this article, we will discuss about fitness functions and how to come up with a fitness function for a given problem.In genetic algorithms, each solution is generally represented as a string of binary numbers, known as a chromosome. We have to test these solutions and come up with the best set of solutions to solve a given problem. Each solution, therefore, needs to be awarded a score, to indicate how close it came to meeting the overall specification of the desired solution. This score is generated by applying the fitness function to the test, or results obtained from the tested solution.Generic Requirements of a Fitness Function The following requirements should be satisfied by any fitness function.

The fitness function should be clearly defined. The reader should be able to clearly understand how the fitness score is calculated. The fitness function should be implemented efficiently. If the fitness function becomes the bottleneck of the algorithm, then the overall efficiency of the genetic algorithm will be reduced. The fitness function should quantitatively measure how fit a given solution is in solving the problem. The fitness function should generate intuitive results. The best/worst candidates should have best/worst score values.

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

Answer:Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes. E.g-->Examples include using neural networks to predict which winery a glass of wine originated from or bagged decision trees for predicting the credit rating of a borrower. Predictive modeling is often performed using curve and surface fitting, time series regression, or machine learning approaches.

Descriptive mining is usually used to provide correlation, cross-tabulation, frequency, etc. The term 'Predictive' means to predict something, so predictive data mining is the analysis done to predict the future event or other data or trends. It is based on the reactive approach. It is based on the proactive approach.Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes

There are many different types of predictive modeling techniques including ANOVA, linear regression (ordinary least squares), logistic regression, ridge regression, time series, decision trees, neural networks, and many more.

Modeling & Simulation Physical: A physical model is a model whose physical characteristics resemble the physical characteristics of the system being modeled. ... Mathematical: A mathematical model is a symbolic model whose properties are expressed in mathematical symbols and relationships. Discrete model: Changes to the system occur at specific times. Continuous model: the state of the system changes continuously over time.

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters.

Answer:There are so many performance evaluation measures when it comes to selecting a classification model that our brain can get tangled just like a thread ball during knitting! In this blog, my intention is to declutter and organize the several jargon used in classification problems from a binary classification point of view. Once the jargons are clear, we can use them in the most appropriate manner and we can knit the perfect classification model.

What are the Performance Evaluation Measures for Classification Models? Confusion Matrix Precision Recall/ Sensitivity Specificity F1-Score AUC & ROC Curve Confusion Matrix: Confusion Matrix usually causes a lot of confusion even in those who are using them regularly. Terms used in defining a confusion matrix are TP, TN, FP, and FN.

Use case: Let's take an example of a patient who has gone to a doctor with certain symptoms. Since it's the season of Covid, let's assume that he went with fever, cough, throat ache, and cold. These are symptoms that can occur during any seasonal changes too. Hence, it is tricky for the doctor to do the right diagnosis.

True Positive (TP): Let's say the patient was actually suffering from Covid and on doing the required assessment, the doctor classified him as a Covid patient. This is called TP or True Positive. This is because the case is positive in real and at the same time the case was classified correctly. Now, the patient can be given appropriate treatment which means, the decision made by the doctor will have a positive effect on the patient and society.

False Positive (FP): Let's say the patient was not suffering from Covid and he was only showing symptoms of seasonal flu but the doctor diagnosed him with Covid. This is called FP or False Positive. This is because the case was actually negative but was falsely classified as positive. Now, the patient will end up getting admitted to the hospital or home and will be given treatment for Covid. This is an unnecessary inconvenience for him and others as he will get unwanted treatment and quarantine. This is called Type I Error.

Loading Image Learn | Write | Earn Participate and become a part of 800+ data science authors True Negative (TN): Let's say the patient was not suffering from Covid and the doctor also gave him a clean chit. This is called TN or True Negative. This is because the case was actually negative and was also classified as negative which is the right thing to do. Now the patient will get treatment for his actual illness instead of taking Covid treatment.

False Negative (FN): Let's say the patient was suffering from Covid and the doctor did not diagnose him with Covid. This is called FN or False Negative as the case was actually positive but was falsely classified as negative. Now the patient will not get the right treatment and also he will spread the disease to others. This is a highly dangerous situation in this example. This is also called Type II Error.

Summary: In this particular example, both FN and FP are dangerous and the classification model which has the lowest FN and FP values needs to be chosen for implementation. But in case there is a tie between few models which score very similar when it comes to FP and FN, in this scenario the model with the least FN needs to be chosen. This is because we simply cannot afford to have FNs! The goal of the hospital would be to not let even one patient go undiagnosed (no FNs) even if some patients get diagnosed wrongly (FPs) and asked to go under quarantine and special care.

Here is how a confusion matrix looks like:

Classification evaluation measures - confusion metric Accuracy: Accuracy = (TP + TN) / (TP + FP +TN + FN)

This term tells us how many right classifications were made out of all the classifications. In other words, how many TPs and TNs were done out of TP + TN + FP + FNs. It tells the ratio of "True"s to the sum of "True"s and "False"s.

Use case: Out of all the patients who visited the doctor, how many were correctly diagnosed as Covid positive and Covid negative.

Precision: Precision = TP / (TP + FP)

Out of all that were marked as positive, how many are actually truly positive.

Use case: Let's take another example of a classification algorithm that marks emails as spam or not. Here, if emails that are of importance get marked as positive, then useful emails will end up going to the "Spam" folder, which is dangerous. Hence, the classification model which has the least FP value needs to be selected. In other words, a model that has the highest precision needs to be selected among all the models.

Recall or Sensitivity: Recall = TP/ (TN + FN)

Out of all the actual real positive cases, how many were identified as positive.

Use case: Out of all the actual Covid patients who visited the doctor, how many were actually diagnosed as Covid positive. Hence, the classification model which has the least FN value needs to be selected. In other words, a model that has the highest recall value needs to be selected among all the models.

Specificity: Specificity = TN/ (TN + FP)

Out of all the real negative cases, how many were identified as negative.

Use case: Out of all the non-Covid patients who visited the doctor, how many were diagnosed as non-Covid.

F1-Score: F1 score = 2* (Precision * Recall) / (Precision + Recall)

As we saw above, sometimes we need to give weightage to FP and sometimes to FN. F1 score is a weighted average of Precision and Recall, which means there is equal importance given to FP and FN. This is a very useful metric compared to "Accuracy". The problem with using accuracy is that if we have a highly imbalanced dataset for training (for example, a training dataset with 95% positive class and 5% negative class), the model will end up learning how to predict the positive class properly and will not learn how to identify the negative class. But the model will still have very high accuracy in the test dataset too as it will know how to identify the positives really well.

Use case: Let's take an example where we must give equal importance to both the classes – classify an email as Spam and non-Spam. Let's assume that the model was trained only a highly imbalanced training dataset. Here, Spam is "positive" and non-Spam is "negative" and the training dataset was 90% spam emails and 10% non-spam emails. A model with high accuracy will know to correctly identify all the spam emails but will have trouble identifying non-spam emails. Hence, a lot of important emails will end up going to the spam folder. But if we select a model that has a high F1 score, it would perform better in classifying non-spam from spam.

Area Under Curve (AUC) and ROC Curve:

AUC or Area Under Curve is used in conjecture with ROC Curve which is Receiver Operating Characteristics Curve. AUC is the area under the ROC Curve. So let's first understand the ROC Curve.

Classification evaluation measures - area under the curve

A ROC Curve is drawn by plotting TPR or True Positive Rate or Recall or Sensitivity (which we saw above) in the y-axis against FPR or False Positive Rate in the x-axis. FPR = 1- Specificity (which we saw above).

TPR = TP/ (TP + FN)

FPR = 1 – TN/ (TN+FP) = FP/ (TN + FP)

If we use a random model to classify, it has a 50% probability of classifying the positive and negative classes correctly. Here, the AUC = 0.5. A perfect model has a 100% probability of classifying the positive and negative classes correctly. Here, the AUC = 1. So when we want to select the best model, we want a model that is closest to the perfect model. In other words, a model with AUC close to 1. When we say a model has a high AUC score, it means the model's ability to separate the classes is very high (high separability). This is a very important metric that should be checked while selecting a classification model.

4. i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting? ii. What does it mean to overfit? When is it going to happen? iii. In the sense of model fitting, explain the bias-variance trade-off.

Answer:Your model is underfitting the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples (often called X) and the target values (often called Y).Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen datWhen a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance.

2.Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. Overfitting happens due to several reasons, such as: • The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.Your model is overfitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.

3.If algorithms fit too complex (hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.

Answer:There are many ways to improve the accuracy of your machine learning models. By using methods like feature engineering, adjusting hyperparameters, and trying multiple algorithms, you give yourself a great change to create a really accurate model. improve the efficiency of a model? Method 1: Add more data samples. Data tells a story only if you have enough of it. ... Method 2: Look at the problem differently. ... Method 3: Add some context to your data. ... Method 4: Finetune your hyperparameter. ... Method 5: Train your model using cross-validation. ... Method 6: Experiment with a different algorithm. ... Takeaways

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model?

Answer: Twin sample validation can be used to validate results of unsupervised learning. It should be used in combination with internal validation. It can prove to be highly useful in case of time-series data where we want to ensure that our results remain same across time. The most common unsupervised learning method is cluster analysis, which applies clustering methods to explore data and find hidden patterns or groupings in data. With MATLAB you can apply many popular clustering algorithms: Hierarchical clustering: Builds a multilevel hierarchy of clusters by creating a cluster tree.

The most common unsupervised learning method is cluster analysis, which applies clustering methods to explore data and find hidden patterns or groupings in data. A good resource (with references) for clustering is sklearn's documentation page, Clustering Performance Evaluation. This covers several method, but all but one, the Silhouette Coefficient, assumes ground truth labels are available. This method is also mentioned in the question Evaluation measure of clustering, linked in the comments for this question.

If your unsupervised learning method is probabilistic, another option is to evaluate some probability measure (log-likelihood, perplexity, etc) on held out data. The motivation here is that if your unsupervised learning method assigns high probability to similar data that wasn't used to fit parameters, then it has probably done a good job of capturing the distribution of interest. A domain where this type of evaluation is commonly used is language modeling.

The last option I'll mention is using a supervised learner on a related auxiliary task. If you're unsupervised method produces latent variables, you can think of these latent variables as being a representation of the input. Thus, it is sensible to use these latent variables as input for a supervised classifier performing some task related to the domain the data is from. The performance of the supervised method can then serve as a surrogate for the performance of the unsupervised learner. This is essentially the setup you see in most work on representation learning.

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.

Answer: Classification predictive modeling is the task of approximating a mapping ... A regression can have real valued or discrete input variables.These models then predict outcomes with the best possible accuracy when new data (aka testing datasets) is fed to them. The outcome predicted by a classification algorithm is categorical in nature

Fundamentally, classification is about predicting a label and regression is about predicting a quantity. Why linear regression can't use for classification? The main reason for that is the predicted values are continuous, not probabilistic. So we can't get an exact class to accomplish the classification

Double-click (or enter) to edit

8. Describe the predictive modeling method for numerical values.what distinguishes it from categorical predictive modeling?

Answer:Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends. Simple linear regression: A statistical method to mention the relationship between two variables which are continuous. 2. Multiple linear regression: A statistical method to mention the relationship between more than two variables which are continuous.Predictive Analytics is used to make predictions about unknown future events. Whereas statistics is the science and it's mainly used in 'Research'. Statistics helps in making a conclusion from the data by collecting, analyzing, and presenting. In the case of prediction, the accuracy relies on how well you guess the value for new data. As noted, predictive analytics uses advanced mathematics to examine patterns in current and past data in order to predict the future. Machine learning is a tool that automates predictive modeling by generating training algorithms to look for patterns and behaviors in data without explicitly being told what to look for The classification technique is used to categorise the data, depending on its similarities and to identify the class.In the case of classification, the accuracy relies on encountering the class label accurately. As noted, predictive analytics uses advanced mathematics to examine patterns in current and past data in order to predict the future. Machine learning is a tool that automates predictive modeling by generating training algorithms to look for patterns and behaviors in data without explicitly being told what to look for The classification technique is used to categorise the data, depending on its similarities and to identify the class.In the case of classification, the accuracy relies on encountering the class label accurately.It is also created from a training set.

9. The following data were collected when using a classification model to predict the malignancy of a group of patients' tumors:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model's error rate, Kappa value, sensitivity, precision, and F-measure.

Answer:

10. Make quick notes on:

1. The process of holding out

2. Cross-validation by tenfold

3. Adjusting the parameters

Answer:1.The Process of holding out-->Holdout Method is the simplest sort of method to evaluate a classifier. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set. A classifier performs function of assigning data items in a given collection to a target category or class.Holdout data refers to a portion of historical, labeled data that is held out of the data sets used for training and validating supervised machine learning models. It can also be called test data.Sometimes referred to as “testing” data, a holdout subset provides a final estimate of the machine learning model's performance after it has been trained and validated. Holdout sets should never be used to make decisions about which algorithms to use or for improving or tuning algorithms.

2..Cross Validation by tenfold---->With this method we have one data set which we divide randomly into 10 parts. We use 9 of those parts for training and reserve one tenth for testing. We repeat this procedure 10 times each time reserving a different tenth for testing.Cross-validation is usually used in machine learning for improving model prediction when we don't have enough data to apply other more efficient methods like the 3-way split (train, validation and test) or using a holdout dataset Introduction. When performing cross-validation, it is common to use 10 folds

3...Adjusting the parametrer--->Adjustment Method Parameters. To define an adjustment calculation method, you must specify one or more parameters, depending on the method.

11. Define the following terms:

1. Purity vs. Silhouette width

2. Boosting vs. Bagging

3. The eager learner vs. the lazy learner

Answer:---> 1.Purity vs. Silhouette width....In classification, purity measures the extent to which a group of records share the same class. It is also termed class purity or homogeneity, and sometimes impurity is measured instead.Within the context of cluster analysis, Purity is an external evaluation criterion of cluster quality. It is the percent of the total number of objects(data points) that were classified correctly, in the unit range [0..1].

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.Silhouette also is any outline or sharp shadow of an object. The word was satirically derived from the name of the parsimonious mid-18th-century French finance minister Étienne de Silhouette, whose hobby was the cutting of paper shadow portraits (the phrase à la Silhouette grew to mean “on the cheap”)

2.Boosting vs Bagging....Bagging and boosting are both ensemble learning methods in machine learning. Bagging and boosting are similar in that they are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one. Ensemble learning helps to improve machine learning model performance by combining several models. This approach allows the production of better predictive performance compared to a single model. The basic idea behind ensemble learning is to learn a set of classifiers (experts) and to allow them to vote. This diversification in Machine Learning is achieved by a technique called ensemble learning. The idea here is to train multiple models, each with the objective to predict or classify a set of results.Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.In the bagging method all the individual models will take the bootstrap samples and create the models in parallel. Whereas in the boosting each model will build sequentially. The output of the first model (the erros information) will be pass along with the bootstrap samples data.Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees.

3. The Eager learner vs the lazy learner----->In machine learninh Eager learning methods construct general, explicit description of the target function based on the provided training examples. Lazy learning methods simply store the data and generalizing beyond these data is postponed until an explicit request is made.Is decision tree lazy or eager. An advantage of Eager Learning algorithms is that it takes less time to classify the test data when it receives it, however, it takes a longer time to learn from the data compared to Lazy Learning Algorithms. An example of this is the Decision Tree algorithm.In machine learning, lazy learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system, as opposed to eager learning, where the system tries to generalize the training data before receiving queries.