**FSDS MAY BATCH 2022(ML Assignment -7)**

**Submitted by: Shubham Tiwari**

Q1: 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?

Ans: A target function, also known as a cost or objective function, is a mathematical function that represents the performance of a machine learning model on a given dataset. The target function maps the model's input features to its output predictions, and its goal is to minimize the difference between the predictions and the actual values.

**A real-life example of a target function** could be a linear regression model that predicts the price of a house based on its square footage, number of bedrooms, and location. In this case, the target function would be the mean squared error between the predicted and actual prices of the houses.

**The fitness of the target function is assessed by evaluating the model's performance on a set of data that is separate from the training data, called validation set.** This performance is usually measured by a metric such as accuracy, precision, recall, or mean squared error. The goal is to minimize the value of the target function, so that the predictions of the model will be as close as possible to the true values.

**Another way to assess the fitness of a model is to** use cross-validation, which is a technique that divides the data into several subsets, and use each subset in turn as the validation set and the remaining subsets as the training set. This way the model performance can be evaluated multiple times and an average can be taken to get a better idea of how well the model will perform on unseen data.

Q2: 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.

Ans: Predictive models and descriptive models are two types of models used in machine learning.**Predictive models are used to** make predictions about future events or outcomes based on historical data. These models are trained on a dataset of input-output pairs and are able to generalize to new examples. They are used for tasks such as classification, regression, and forecasting. Examples of predictive models include:

* Linear regression: A model that predicts a continuous outcome based on a linear combination of input features.
* Random Forest: A collection of decision trees that are combined to make predictions.
* Neural Networks: A set of algorithms, modeled loosely after the human brain, that is designed to recognize patterns.

**Descriptive models, on the other hand, are used to** describe or summarize the underlying patterns in the data. These models are used to understand and explain the relationship between different variables and to discover hidden patterns and insights. Examples of descriptive models include:

**Clustering**: A technique that groups similar data points together.

**Principal Component Analysis (PCA):** A technique that reduces the dimensionality of the data by identifying the most important features.

**Factor Analysis:** A technique that identifies latent variables that explain the relationship between multiple observed variables.

In summary, predictive models are used to make predictions about future events, while descriptive models are used to understand and explain the patterns in the data. Predictive models can be used to make predictions on new data while descriptive models are used to understand and explain the patterns in the data.

Q3: Describe the method of assessing a classification model’s efficiency in detail. Describe the various measurement parameters.

Ans: There are several methods for assessing the efficiency of a classification model, but some of the most commonly used methods include:

1. **Accuracy:** This measures the proportion of correct predictions made by the model. It is calculated as the ratio of the number of correct predictions to the total number of predictions. However, accuracy is not always a good metric, especially when the classes are imbalanced.
2. **Confusion Matrix:** A confusion matrix is a table that is used to define the performance of a classification algorithm. It compares the predicted values with the true values and gives the number of correct and incorrect predictions for each class. From the confusion matrix, we can calculate precision, recall, F1 score, and many others.
3. **Precision:** This measures the proportion of true positive predictions out of all positive predictions made by the model. It is calculated as the ratio of true positive predictions to the total number of positive predictions.
4. **Recall:** This measures the proportion of true positive predictions out of all actual positive instances. It is calculated as the ratio of true positive predictions to the total number of actual positive instances.
5. **F1 Score**: The F1 score is the harmonic mean of precision and recall. F1 Score is used to balance precision and recall.
6. **Receiver Operating Characteristic (ROC) curve**: This is a graphical representation of the performance of a classification model. It plots the true positive rate (recall) against the false positive rate at different threshold settings.
7. **Area Under the Curve (AUC):** It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at distinguishing between the given classes.
8. **K-fold cross-validation**: It is a technique for assessing the performance of a machine learning model. It divides the dataset into k-folds, where k-1 of the folds are used as the training set and the remaining fold is used as the test set. This process is repeated k times, with each fold being used as the test set once. The performance metrics are then averaged across all k iterations.
9. These are some of the most commonly used methods for assessing the efficiency of a classification model, but there are many other parameters and methods that can be used depending on the specific problem and requirements.

Q4**: i**. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

Ans: **Underfitting in machine learning models refers to** the situation where a model is unable to capture the underlying pattern in the training data, resulting in poor performance on unseen data.

**The most common reason for underfitting is the model is too simple and not able to capture the complexity of the data**. A model that is underfitting has high bias, meaning it is unable to capture the underlying patterns in the data. This can occur when the model is not complex enough, or when the model is not given enough data to learn from.

**Another reason for underfitting is** when the features used in the model are not informative enough and don't have enough correlation with the target variable.

**A simple example of underfitting would be** a linear regression model for a non-linear dataset, where a linear model can't capture the non-linearity, hence would underfit the data.

**To overcome underfitting, we can** try different approaches like using a more complex model, increasing the number of features, or using more data for training.

**ii.** What does it mean to overfit? When is it going to happen?

Ans: **Overfitting refers to** the situation where a model is so complex that it starts to fit the noise or random fluctuations in the training data, rather than the underlying pattern. This results in a model that performs well on the training data but poorly on unseen or new data.

**Overfitting happens** when a model is too complex in relation to the amount and noise level of the training data. This can happen when the model has too many parameters or features, and the model is able to fit the noise in the training data perfectly, but it's not able to generalize well to new data.

The most common symptoms of overfitting are a high accuracy on the training set and a low accuracy on the test or validation set. Another way to detect overfitting is by comparing the training error and the test error, when the training error is low, but the test error is high, it means that the model is overfitting.

**iii**. In the sense of model fitting, explain the bias-variance trade-off.

Ans:**The bias-variance trade-off refers to** the balance between a model's ability to fit the training data well (low bias) and its ability to generalize to new, unseen data (low variance).

Bias refers to the error that is introduced by approximating a real-world problem, which may be extremely complicated, by a much simpler model. High bias models make strong assumptions about the form of the target function, and they will tend to under-fit the data.

Variance, on the other hand, refers to the amount by which the model's predictions will change if we estimate it using different training data. High variance models are more sensitive to the noise in the training data and they will tend to over-fit the data.

In practice, a model that has low bias and low variance is preferred, but it is often impossible to achieve low bias and low variance simultaneously. A model that has high bias and low variance is simple and not flexible enough to fit the data well. A model that has low bias and high variance is too flexible and will fit the noise in the data.

Q5: Is it possible to boost the efficiency of a learning model? If so, please clarify how.

Ans: Yes, it is possible to boost the efficiency of a learning model. Here are a few ways to do so:

1. **Feature selection:** Selecting a subset of the most relevant features can improve the model's efficiency by reducing the dimensionality of the data and eliminating noise.
2. **Feature engineering:** Creating new features from the existing ones can also improve the model's efficiency.
3. **Hyperparameter tuning**: Tuning the model's hyperparameters can also improve its efficiency. This can be done through techniques such as grid search and randomized search.
4. **Regularization:** Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function. This helps to keep the model's weights small and prevent overfitting.
5. **Ensemble methods**: Combining multiple models can also boost the efficiency of a learning model. Ensemble methods such as bagging and boosting can be used to improve the performance of a model.
6. **Using transfer learning:** Using pre-trained models can save time and computational resources and also improve the performance of the model.
7. **Early stopping**: Early stopping is a technique used to stop the training process when the model stops improving on a validation set. This can prevent overfitting and save computational resources.
8. It's important to note that these techniques are not mutually exclusive and often can be combined to achieve a more efficient model.

Q6: How would you rate an unsupervised learning model’s success? What are the most common success indicators for an unsupervised learning model?

Ans: Evaluating the success of an unsupervised learning model can be challenging because there is no clear definition of what constitutes a "good" outcome. However, there are several common indicators that can be used to gauge the performance of an unsupervised learning model:

1. **Clustering accuracy:** This measures how well the model has grouped similar data points together in clusters. The higher the accuracy, the better the model is at identifying patterns in the data.
2. **Silhouette score**: This measures the similarity of the data points within a cluster to the other points in the same cluster, and how different they are from the points in other clusters. A high silhouette score indicates that the clusters are well-defined and the data points within them are similar.
3. **Inertia:** This measures the sum of the squared distances between the data points and the cluster centroid. The lower the inertia, the better the model is at identifying clusters in the data.
4. **Reconstructed Error:** This measure the reconstruction error of the data based on the learned representation. Lower error score means better representation of data.
5. **Visualization:** Visualizing the data in two or three dimensions can help identify patterns and evaluate the performance of the model.
6. Keep in mind that these indicators can be used in combination to evaluate the performance of an unsupervised learning model. It's also important to consider the specific context and problem you're working on, and to use domain-specific knowledge to interpret the results of our model.

Q7: 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.

Ans: It is possible to use a classification model for numerical data or a regression model for categorical data, but it may not be the best approach depending on the specific problem and data at hand.A classification model is typically used to predict a categorical outcome, such as a label or class, based on input features. It can be used for numerical data if the outcome is also numerical, but has discrete values.

A regression model is typically used to predict a numerical outcome, such as a continuous value, based on input features. It can be used for categorical data if the outcome is also categorical, but has numerical values.

However, it is generally recommended to use the appropriate model for the specific type of data and outcome. Using a classification model for numerical data or a regression model for categorical data may lead to suboptimal performance and inaccurate predictions.

Q8: Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling?

Ans: Predictive modeling for numerical values, also known as regression analysis, is a statistical method used to predict a continuous outcome variable based on one or more predictor variables. The goal is to find the relationship between the predictor variables and the outcome variable, and use this relationship to make predictions about new data.

**The main distinguishing characteristic between regression analysis and categorical predictive modeling is** the type of outcome variable. In regression analysis, the outcome variable is numerical and continuous, while in categorical predictive modeling, the outcome variable is categorical and discrete.

For example, in a regression problem, the outcome variable could be the price of a house, and the predictor variables could be the square footage, number of bedrooms, and location. In a categorical predictive modeling problem, the outcome variable could be the type of iris flower (setosa, versicolor, or virginica), and the predictor variables could be the sepal length, sepal width, petal length, and petal width.

Different types of regression models include linear regression, polynomial regression, decision tree regression, Random Forest Regression and others.

Q9: The following data were collected when using a classification model to predict the malignancy of a group of patient’s; 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.

Ans: **Error rate**: (3 + 7) / (15 + 75 + 3 + 7) = 0.1 or 10%

**Kappa value:** (Accurate estimates - Expected Accurate estimates) / (Total - Expected Accurate estimates)

Where Expected Accurate estimates = (15 + 75) \* (15 + 3) / (15 + 75 + 3 + 7) + (3 + 7) \* (75 + 7) / (15 + 75 + 3 + 7) = 0.849

**Sensitivity** (also known as true positive rate or recall): 15 / (15 + 3) = 0.833

**Precision** (also known as positive predictive value): 15 / (15 + 7) = 0.682

**F-measure**: 2 \* (Precision \* Sensitivity) / (Precision + Sensitivity) = 0.747

Note: The above formula for Kappa is the Cohen's Kappa (usually used for binary classification).

Q10: Make quick notes on:

1. The process of holding out.

Ans: Holdout is a method of evaluating the performance of a machine learning model by dividing the available data into two parts: a training set and a test set. The model is trained on the training set and its performance is then evaluated using the test set. The idea is that the test set is representative of unseen data, so the model's performance on the test set provides an estimate of how well the model will perform on new, unseen data. The holdout method is a form of cross-validation, which is a technique used to mitigate the risk of overfitting, a common problem in machine learning where a model performs well on the training data but poorly on new, unseen data.

2. Cross-validation by tenfold.

Ans: K-fold cross-validation is a technique used to mitigate the risk of overfitting in machine learning by dividing the available data into k subsets (or "folds"), and then training and evaluating the model k times, each time using a different fold as the test set and the remaining k-1 folds as the training set. The performance of the model is then averaged across the k iterations.

**Ten-fold cross-validation is a specific type of k-fold cross-validation where k is set to 10. This means that the available data is divided into 10 equal-sized subsets, and the model is trained and evaluated 10 times, each time using a different subset as the test set. The performance of the model is then averaged across the 10 iterations. This is a widely used method for getting a robust estimate of the model's performance.**

**In brief, Ten-fold cross-validation is a technique used to e**valuate the performance of a machine learning model by dividing the available data into 10 subsets (or "folds"), and then training and evaluating the model 10 times, each time using a different fold as the test set and the remaining 9 folds as the training set. The performance of the model is then averaged across the 10 iterations.

3. Adjusting the parameters.

Ans: Adjusting the parameters, also known as hyperparameter tuning or hyperparameter optimization, is the process of finding the best set of parameters for a machine learning model. The process of adjusting the parameters is usually performed after the model has been trained on the training set and its performance has been evaluated using a validation set. The goal is to find the parameters that result in the best performance on the validation set, which is expected to generalize well to new, unseen data.

There are several methods for adjusting the parameters of a machine learning model, including:

* Grid search: A simple method that involves specifying a set of possible parameter values, and then training and evaluating the model for each combination of parameter values. Grid search is easy to implement, but it can be computationally expensive and may not always find the optimal set of parameters.
* Random search: A method that involves sampling random combinations of parameter values and training and evaluating the model for each set of parameter values. Random search can be less computationally expensive than grid search and may find the optimal set of parameters faster.
* Bayesian optimization: A more sophisticated method that uses Bayesian inference to model the distribution of the performance of the model with respect to the parameter values. Bayesian optimization can be more efficient than grid search or random search, but it can also be more complex to implement.
* The choice of method depends on the complexity of the model, the size of the data set, and the computational resources available. In most cases, a combination of methods and techniques is applied in order to find the optimal set of parameters.

Q11: Define the following terms:

1. Purity vs. Silhouette width.

Ans: **Purity and silhouette width are two different measures used to evaluate the performance of clustering algorithms.** Purity is a measure of the accuracy of a clustering algorithm, which compares the number of points in a cluster that are correctly classified with the total number of points in the cluster. A higher purity value indicates that the points in the cluster are more similar to each other. **Silhouette width is a measure of the similarity of points within a cluster to other points in the same cluster, relative to points in other clusters**. A higher silhouette width indicates that the points in the cluster are more similar to each other than to points in other clusters.

2. Boosting vs. Bagging.

Ans: Boosting and bagging are both ensemble learning techniques used to improve the performance of machine learning models.

**Bagging stands for Bootstrap Aggregating**, it is a method that involves training multiple instances of a model on different subsets of the training data, and then averaging the predictions of all of the instances to make a final prediction. Bagging can reduce the variance of a model by averaging the predictions of multiple instances, resulting in a more stable and robust model.

**Boosting is a method that involves** training multiple instances of a model, but unlike bagging, each instance is trained on a different subset of the training data, with more emphasis placed on the instances that are misclassified by the previous instances. Boosting can reduce the bias of a model by iteratively training instances on the errors made by previous instances, resulting in a more accurate model.

**Boosting is more powerful than Bagging** but also more prone to overfitting and it's more computationally expensive.

3. The eager learner vs. the lazy learner.

Ans: **Eager learners are algorithms that actively build a model during the training phase,** and then use that model to make predictions on new data. Examples of eager learners include decision trees, k-nearest neighbors (k-NN), and support vector machines (SVMs). These algorithms require a significant amount of computation during training, but once the model is built, predictions can be made quickly and efficiently.

**Lazy learners, on the other hand, do not build a model during the training phase. Instead, they simply store the training data and wait until a new data point is encountered before making a prediction**. Examples of lazy learners include k-NN and the k-means clustering algorithm. Lazy learners are less computationally expensive during training, but they may take longer to make predictions on new data.

It's worth noting that the division is not absolute, some algorithm can have both characteristics, being considered as a semi-supervised learner.