My Batch: DS2401
My Name: Mohammed Ikramuddin

MACHINE LEARNING WORKSHEET-5-Answers

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Ans) R-squared (R^2) and Residual Sum of Squares (RSS) are both commonly used measures to assess the goodness of fit of a regression model, but they capture different aspects of model performance, and the choice between them depends on the context and what you want to evaluate. The RSS is absolute amount of explained variation, whereas R^2 is the absolute amount of variation as a proportion of total variation.

R^2, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with a higher value indicating a better fit. R-squared is useful for comparing different models or for determining the proportion of the variability in the dependent variable that is explained by the model.

On the other hand, RSS measures the total amount of unexplained variance in the dependent variable that remains after the model has been fit. It is the sum of the squared differences between the actual and predicted values of the dependent variable. A lower value of RSS indicates a better fit. RSS is useful for evaluating the accuracy of the predictions of the model.

In general, both measures are important and should be considered together when evaluating the goodness of fit of a model. However, ***R^2 is often considered to be a better measure of goodness of fit than RSS because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which is more interpretable and easier to compare across models.***

---

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

Ans) In statistical data analysis the ***total sum of squares (TSS or SST)*** is a quantity that appears as part of a standard way of presenting results of such analyses. For a set of observations, it is defined as the sum over all squared differences between the observations and their overall mean.

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The ***explained sum of squares (ESS),*** alternatively known as the ***model sum of squares*** or ***sum of squares due to regression*** (**SSR** – not to be confused with the residual sum of squares (RSS) or sum of squares of errors), is a quantity used in describing how well a model, often a regression model, represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values.

$$ESS = \sum \hat{y}_i^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

The ***residual sum of squares (RSS)***, also known as the ***sum of squared residuals (SSR)*** or the ***sum of squared estimate of errors (SSE)***, is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model, such as a linear regression. A small RSS indicates a tight fit of the model to the data. It is used as an optimality criterion in parameter selection and model selection.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

TSS = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Squares. The aim of Regression Analysis is to explain the variation of dependent variable Y.

---

3. What is the need of regularization in machine learning?

Ans) While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

L1 and L2 regularization are methods in machine learning that add a penalty term to the loss function. L1 regularization is also known as lasso regression, and L2 regularization is also known as ridge regression.

Regularization techniques are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

---

4. What is Gini–impurity index?

Ans) ***Gini impurity measures how often a randomly chosen element of a set would be incorrectly labelled***, if it was labelled randomly and independently according to the distribution of labels in the set. It reaches its minimum (zero) when all cases in the node fall into a single target category.

Gini Index is a powerful measure of the randomness or the impurity or entropy in the values of a dataset. It aims to decrease the impurities from the root nodes (at the top of decision tree) to the leaf nodes (vertical branches down the decision tree) of a decision tree model.

The Gini Impurity formula is: $1 - (p_1)^2 - (p_2)^2$, where p_1 and p_2 represent the probabilities of the two classes in a binary classification problem.

The Gini index is the measure of the extent to which the distribution of income or consumption among individuals or households within an economy deviates from a perfectly

equal distribution. A Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.

---

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Ans) Decision trees are a popular and powerful method for data mining, as they can handle both numerical and categorical data, and can easily interpret the results. However, ***unregularized decision trees can also suffer from overfitting, which means that they learn too much from the training data and fail to generalize well to new data.***

Overfitting happens when any learning process overly optimizes training set error at the cost test error. While it's possible for training and testing to perform equality well in cross-validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem.

---

6. What is an ensemble technique in machine learning?

Ans) ***Ensemble methods are techniques that aim at improving the accuracy and resilience in forecasting the results in models by combining predictions from multiple models instead of using single model.*** The combined models increase the accuracy of the results significantly.

Ensemble technique aims to mitigate errors or biases that may exist in individual models by leveraging the collective intelligence of the ensemble. Some of the advanced ensemble classifiers are:

- ***Stacking.***
- ***Blending.***
- ***Bagging.***
- ***Boosting***.

Two very famous examples of ensemble methods are gradient-boosted tress and random forests. More generally, ensemble models can be applied to any base learner beyond trees, in averaging methods such as Bagging methods, model stacking, or Voting, or in boosting, as AdaBoost.

---

7. What is the difference between Bagging and Boosting techniques?

Ans) Bagging is the simplest way of combining predictions that belong to the same type, while Boosting is a way of combining predictions that belong to the different types. ***Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.***

In Bagging, models are trained independently in parallel on different random subsets of the data. Whereas in boosting, models are trained sequentially, with each model learning from the errors of the previous one.

My Batch: DS2401
My Name: Mohammed Ikramuddin

MACHINE LEARNING WORKSHEET-5-Answers

8. What is out-of-bag error in random forests?

Ans) ***Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models*** utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from.

OOB (out-of-bag) score is a performance metric for a machine learning model, specifically for ensemble models such as random forests. It is calculated using the samples that are not used in the training of the model, which is called out-of-bag samples.

---

9. What is K-fold cross-validation?

Ans) ***K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time***. Performance metrics from each fold are averaged to estimate the model's generalization performance.

One-fold is used for validation and other K-1 folds are used for training the model. To use every fold as a validation set and other left-outs as a training set, this technique is repeated k times until each fold is used once.

K-fold cross-validation has several advantages for predictive analytics, such as reducing the variance of the performance estimate and allowing you to use more data for training. It also helps you avoid overfitting, as it exposes your model to different subsets of the data.

---

10. What is hyper parameter tuning in machine learning and why it is done?

Ans) ***When you're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning.***

Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

These are used to specify the learning capacity and complexity of the model. Some of the hyperparameters are used for the optimization of the models, such as Batch size, learning rate, etc., and some are specific to the models, such as Number of Hidden layers, etc.

My Batch: DS2401
My Name: Mohammed Ikramuddin

MACHINE LEARNING WORKSHEET-5-Answers

11. What issues can occur if we have a large learning rate in Gradient Descent?

Ans) Learning Rate Selection: The choice of learning rate can significantly impact the performance of gradient descent. ***If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge.***

It means that the learning rate remains constant throughout the training process. However, this can lead to suboptimal results or even failure to converge. If the learning rate is too high, the network may overshoot the optimal point and oscillate around it, or diverge and produce large errors.

---

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Ans) Logistic regression has traditionally been used to come up with a hyperplane that separates the feature space into classes. But if we suspect that the decision boundary is non-linear, we may get better results by attempting some non-linear functional forms for the logit function.

Logistic regression is indeed non-linear in terms of Odds and Probability; however, it is linear in terms of Log Odds.

---

13. Differentiate between Ada Boost and Gradient Boosting.

Ans) AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

The main difference between these two algorithms is that Gradient boosting has a fixed base estimator i.e., Decision Trees whereas in AdaBoost we can change the base estimator according to our needs.

Ada Boost builds a new stump based on the errors that the previous stump made. It continues to make stumps in this fashion until it has made the number of stumps we've asked for. In Contrast, Gradient Boost starts by making a single leaf instead of a tree or stump.

---

14. What is bias-variance trade off in machine learning?

Ans) ***In machine learning, as you try to minimize one component of the error (e.g., bias), the other component (e.g., variance) tends to increase, and vice versa. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance trade off.***

In statistics and machine learning, the bias–variance trade off describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

Bias-Variance trade-off is crucial in machine learning because it directly impacts a model's predictive performance. A model with high bias will consistently produce predictions that are far from the actual values, while a model with high variance will produce widely varying predictions for different training datasets.

---

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Ans) ***A linear kernel is a type of kernel function used in machine learning, including in SVMs (Support Vector Machines). It is the simplest and most commonly used kernel function, and it defines the dot product between the input vectors in the original feature space***. Where x and y are the input feature vectors.

Linear kernels are simpler and computationally efficient, suitable for linearly separable data. The application of a support vector machine with a linear kernel is to perform classification or regression. It will perform best when there is a linear decision boundary or a linear fit to the data, thus the linear kernel.

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

RBF Kernel is popular because of its similarity to K-Nearest Neighbourhood Algorithm. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

Decision boundary with a polynomial kernel. The advantage of using the kernelized version is that you can specify the degree to be large, thus increasing the chance that data will become linearly separable in this high-dimensional space, without slowing the model down.