Neha Tomar **22 July 2024**

Assignment Worksheet 2

ML 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?**

- R-squared is a better measure of goodness of fit because it provides a proportion of the variance explained by the model, making it easier to interpret. RSS is the total squared error and doesn't provide a relative measure.

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.**

- TSS measures the total variance in the dependent variable. ESS measures the variance explained by the model. RSS measures the variance not explained by the model. The relationship is \(TSS = ESS + RSS \).

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

- Regularization is needed to prevent overfitting by penalizing complex models, thus improving the model's generalization to new data.

4. **What is Gini-impurity index?**

- The Gini-impurity index measures the probability of incorrectly classifying a randomly chosen element if it was randomly labeled according to the distribution of labels in the dataset.

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

- Yes, because they can create overly complex trees that fit the noise in the training data, leading to poor generalization on unseen data.

6. **What is an ensemble technique in machine learning?**

- An ensemble technique combines multiple models to produce a better performance than any single model by reducing variance, bias, or improving predictions.

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

- Bagging involves training multiple models independently on different subsets of data and averaging their predictions, while Boosting trains models sequentially, with each model correcting the errors of the previous ones.

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

- Out-of-bag error is an estimate of the prediction error for a random forest model, calculated using the data not included in the bootstrap samples for each tree.

9. **What is K-fold cross-validation?**

- K-fold cross-validation involves dividing the data into K equally sized folds, training the model K times, each time using a different fold as the validation set and the remaining folds as the training set, and averaging the results.

10. **What is hyperparameter tuning in machine learning and why is it done?**

- Hyperparameter tuning involves selecting the best hyperparameters for a model to optimize its performance, as these parameters significantly influence the model’s ability to learn from data.

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

- A large learning rate can cause the model to converge too quickly to a suboptimal solution or even diverge, missing the optimal point entirely.

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

- Logistic Regression is not suitable for non-linear data unless transformed features or non-linear basis functions are used, as it assumes a linear relationship between the independent variables and the log-odds of the dependent variable.

13. **Differentiate between Adaboost and Gradient Boosting.**

- Adaboost focuses on misclassified instances by increasing their weights for the next model, while Gradient Boosting optimizes the loss function by adding models that correct residual errors of previous models.

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

- The bias-variance trade-off is the balance between the error introduced by the model's assumptions (bias) and the error introduced by sensitivity to fluctuations in the training set (variance). Reducing one often increases the other.

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

- **Linear Kernel:** Suitable for linearly separable data, computes a linear decision boundary.

- **RBF (Radial Basis Function) Kernel:** Maps data into a higher-dimensional space, suitable for non-linear data, sensitive to the distance between data points.

- **Polynomial Kernel:** Computes non-linear decision boundaries by considering polynomial combinations of input features, allows control over complexity via the degree of the polynomial.

Statistics Answers

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.

d) Expected

2. Chisquare is used to analyse

c) Frequencies

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

c) 6

4. Which of these distributions is used for a goodness of fit testing?

b) Chisquared distribution

5. Which of the following distributions is Continuous?

c) F Distribution

6. A statement made about a population for testing purpose is called?

b) Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

a) Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

a) Two tailed

9. Alternative Hypothesis is also called as?

b) Research Hypothesis

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

a) np