Machine Learning & Statistics Assignment Solution

Name of Student: Mudassar Iqbal Shakil Ahmed Sayyed

Name of assignment: Machine Learning & Statistics

Internship Batch: DS2404

Machine Learning Solution

1. **R-squared** is usually better for checking how well your model fits the data. It's a percentage showing how much your data fits the model, so higher is usually better. RSS is just the total of squared differences, so it's harder to interpret.

2. **TSS** (Total Sum of Squares) shows how much the dependent variable varies. **ESS** (Explained Sum of Squares) shows how much of that variation our model explains.
   RSS (Residual Sum of Squares) shows how much of the variation our model can't explain.
   The equation is like this: **TSS = ESS + RSS.**

3. **Regularization** helps stop your model from fitting too closely to the training data, which can make it perform badly on new data. It's like a penalty for complexity.

4. The **Gini impurity index** is a way of measuring how often a randomly chosen element from the set would be incorrectly labelled. It's used in decision trees.

5. Yes, **decision trees** without regularization can overfit because they might get too complex, trying to fit every detail of the training data, and then do badly on new data.

6. An **ensemble technique** is when you use multiple models together to get better results than any single model could get on its own.

7. **Bagging** and **boosting** are both ensemble techniques. Bagging uses different subsets of the training data with the same algorithm, while boosting trains models one after another, each trying to correct the mistakes of the ones before.

8. **Out-of-bag error** in random forests is a way of testing the model's performance using the training data. It's like a built-in cross-validation.

9. **K-fold cross-validation** is when you split your data into K parts, train your model K times each time leaving out a different part, and then average the results to get a final model.

10. **Hyperparameter tuning** is about choosing the best settings for your model. It's done because different settings can make the model perform better or worse.

11. If the **learning rate** in Gradient Descent is too big, the model might skip over the best solution and never find it, or even get worse and worse.

12. Logistic Regression is a linear method, so it can't handle data that isn't linear unless you transform the data first.

13. **Adaboost** and **Gradient Boosting** are both boosting methods, but they work differently. **Adaboost** changes the weights of the training examples, while Gradient Boosting fits the new model to the residuals of the previous model.

14. The **bias-variance trade-off** is about balancing how much your model pays attention to the training data versus how much it generalizes to new data. Too much of either can make the model perform badly.

15. **Linear, RBF, and Polynomial kernels** are used in SVMs to transform the data. Linear is the simplest and works when the data is linear. RBF can handle nonlinear data, and Polynomial is more complex and can fit more complex data shapes.

Statistics Solution

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.
    a) Mean
    b) Actual
    c) Predicted
    d) Expected
    Answer: d) Expected

2. Chisquare is used to analyse
    a) Score
    b) Rank

c) Frequencies
d) All of these
Answer: c) Frequencies

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?
a) 4
b) 12
c) 6
d) 8
Answer: c) 6

4. Which of these distributions is used for a goodness of fit testing?
a) Normal distribution
b) Chi squared distribution
c) Gamma distribution
d) Poission distribution
Answer: b) Chi squared distribution

5. Which of the following distributions is Continuous
a) Binomial Distribution
b) Hypergeometric Distribution
c) F Distribution
d) Poisson Distribution
Answer: c) F Distribution

6. A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis
c) Level of Significance
d) TestStatistic
Answer: b) Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?
a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis
Answer: a) Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?
a) Two tailed
b) One tailed
c) Three tailed

d) Zero tailed

Answer: a) Two tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis

b) Research Hypothesis

c) Simple Hypothesis

d) Null Hypothesis

Answer: 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

b) n

Answer: a) np