ASSIGNMENT - 5

MACHINE LEARNING

LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?

Answer:

1. R-squared is a more effective measure of how well a regression model fits the data than Residual Sum of Squares (RSS).

2. R-squared explain us how much of the variation in the dependent variable which we are trying to predict is accounted by the independent variables the factors we are using to make predictions. Thud a higher R-squared is the better measure of goodness of fit model in regression

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 ?

Answer:

1. TSS (Total sum of Squares) is the total variance in the dependent variable (Y) before any regression is performed. **TSS=i=1∑n(Yi–Y¯)2**

2. ESS (Explained Sum of Squares) measures the variability in Y that is explained by the independent variables (X) in the regression model. **ESS=i=1∑n(Y^i–Y¯)2**

3. RSS (Residual Sum of Squares) measures the remaining unexplained variability in Y after accounting for the variability explained by the regression model. **RSS=i=1∑n(Yi–Y^i)2**

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

Answer:

The needs of Regularization in machine learning are for prevent overfitting, improving performance over new data, to control the complexity in the model and to handle correlated variables in regression

4. What is Gini–impurity index?

Answer:

The Gini impurity index is used in decision trees which are used to measure how often a randomly chosen item would be incorrectly classified. It ranges from 0 to 1. A lower Gini impurity defines a better classification split.

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

Answer:

Yes, unregularized decision trees can overfit because they are very detailed and complex in nature and during picking up some random patterns in the training data it will not work well with new data.

6. What is an ensemble technique in machine learning?

Answer:

Ensemble technique in machine learning is mixed up with multiple models to improve overall performance. Utilizing the strengths of various models, where ensembles achieves better accuracy too and compared to individual models.

Eg. Bagging, boosting and stacking

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

Answer:

Bagging:

- Reduces variance and helps prevent overfitting
- Trains multiple models independently in parallel
- Uses bootstrap sampling
- Less complexity

Boosting:

- Reduce bias and improve accuracy
- Trains models sequentially and focusing on errors of the previous one
- Adjusts weights of data points based on errors
- More complexity

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

Answer:

Out-of-bag error in random forests is the error rate calculated on the part of the training data that was not used to build each tree in the forest. It's a way to check how well the model predicts without needing a separate validation dataset

9. What is K-fold cross-validation?

Answer:

K-fold cross-validation is a technique used in machine learning to assess how well a model can generalize to new data. It involves Splitting Data, Training, Testing and Evaluation.

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

Answer:

There may occurs some issues if the learning rate in Gradient Descent is too large, the following issues are listed below

- Divergence: The model may fail to converge and instead diverge
- Overshooting: The updates can be too large, which makes the algorithm to overshoot the minimum toggling back and forth without settling.
- Instability: The learning process becomes unstable, with wide loss function fluctuating

13 Differentiate between Adaboost and Gradient Boosting ?

Answer:

AdaBoost:

- It is an boosting method that make weak learners to form a strong learner
- Focuses on the mistakes by giving them more importance in the next series
- Each learner is trained one after the other, focusing on fixing the mistakes made by the previous ones

Gradient Boosting:

- It is also an boosting method which helps to improve performance by combining weak models
- It creates new models that focus on correcting the errors made by the previous models
- Each new model is added one after the other to reduce the overall prediction errors by following the direction that reduces the mistakes the most