Machine Learning Assignment

1. A
2. A
3. B
4. B
5. C
6. B
7. D
8. D
9. A
10. B
11. B
12. B, C

13. Explain the term regularization?

- It is a type of regression that limits or shrinks the coefficient estimates towards zero. It basically discourages learning a more complex or flexible model to avoid risk of overfitting.

14. Which particular algorithms are used for regularization?

- Lasso (L1 Regularization): Acronym for acronym for least absolute shrinkage and selection operator, it adds the absolute value of regression coefficients as a penalty term to the loss function.
- Ridge (L2 Regularization): Ridge regression adds the "squared magnitude" of the coefficient as the penalty term to the loss function.
- Elastic net Regularization: It combines L1-norm Regularization and L2-norm Regularization, and is effective at both reducing overfitting and improving interpretability.
- Decision Tree Pruning: Can be used by pruning branches that do not significantly contribute to reducing impurity

15. Explain the term error present in linear regression equation?

- It also known as 'residual' which refers to the variability that is unexplainable in the data which the model is unable to capture. It represents the difference between the predicted values of the regression model and the actual observed value of the data.
- The formula is Error (i) = Observed Value (i) – Predicted Value (i)
- To reduce the error in order to obtain the best-fit line, this is done by minimizing the sum of squared errors.
- If the sum of squared errors are minimised, then this indicates that the regression model is able to explain well the relationship between the independent variables and dependent variables, while still being able to address the errors.

PYTHON – WORKSHEET 1

1. C
2. B
3. C
4. A
5. D
6. C
7. A
8. C
9. A, C
10. A, B

STATISTICS WORKSHEET-1

1. A
2. A
3. B
4. C
5. C
6. A
7. B
8. A
9. C
10. What do you understand by the term Normal Distribution?

   - Normal distribution is a continuous probability distribution of a random variable in the shape of bell curve, where the values are distributed around the mean.

11. How do you handle missing data? What imputation techniques do you recommend?

   - Using mean, median or mode. So we can replace the missing values with mean, median or mode of the non-missing values for that variable.

12. What is A/B testing?

   - It is split testing. Example: Comparing two website designs, where we can compare user's click-through rate using statistical significance or p-value.

13. Is mean imputation of missing data acceptable practice?

   - Depends on the context and goals of the analysis. It is straightforward and easy to use, but prone to bias as it assumes that the missing values are missing at random.

14. What is linear regression in statistics?

- It measures the relationship between the dependent and independent variable, by predicting future values of the dependent variable based on the values of the independent variables.

15. What are the various branches of statistics

- Descriptive Stats which involves describing data using mean, mode, median, standard deviation.
- Machine Learning where we use regressions to develop algorithms in making a prediction