Internship assignment- 4

Name: Priti Chauhan

Batch: 1845

Internship batch: 32

Machine learning worksheet:

1. **C**
2. **D**
3. **C**
4. **B**
5. **D**
6. **C**
7. **D**
8. **A**
9. **D**
10. **C**
11. Some observations within a set of data may fall outside the general scope of the other observations. Such observations are called **outliers**.

    We can use the **IQR method** of identifying outliers to set up a "fence" outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. This is the method that Minitab uses to identify outliers by default.

12. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

13. Adjusted R-squared value can be calculated **based on value of r-squared, number of independent variables (predictors), total sample size**. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.
    Adjusted R squared is calculated by **dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2

14. Difference between standardisation and normalisation

Normalisation	**Standardisation**
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

15. Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

Advantages of Cross Validation

1. Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1. Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.

For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

Statistics worksheet:

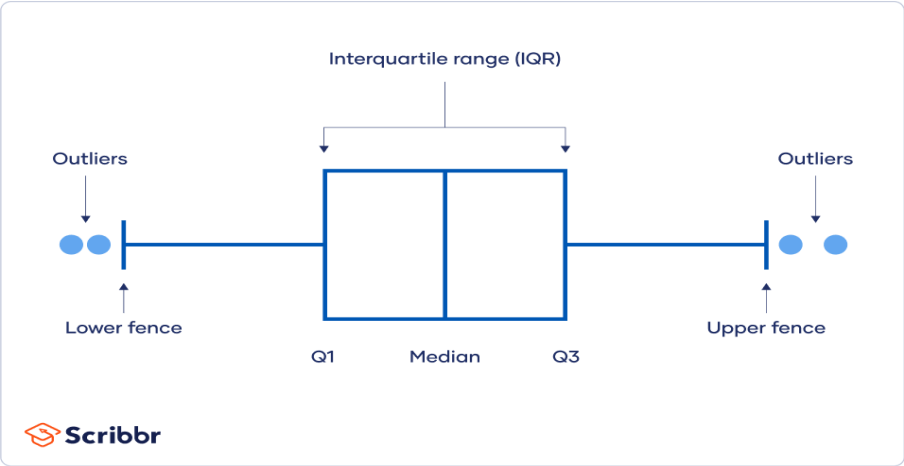
1. The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.
2. There are **two types of sampling methods**: Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.
3. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.
4. A normal distribution is **an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme**.
5. **Covariance is an indicator of the extent to which 2 random variables are dependent on each other**. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.
6. **Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables**.
7. The sensitivity is calculated by **dividing the percentage change in output by the percentage change in input**.
8. A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population.

   Hypothesis testing is formulated in terms of two hypotheses: • H0: the null hypothesis; • H1: the alternate hypothesis.

The hypothesis we want to test is if H1 is “likely” true. So, there are two possible outcomes: • Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1; • Do not reject H0 because of insufficient evidence to support H1.

9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
10. The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), **first find the median (middle value) of the lower and upper half of the data**. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
11. A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.
12. **There are four ways to identify outliers:**
    - **I.** Sorting method.
    - **II.** Data visualization method.
    - **III.** Statistical tests (z scores)
    - **IV.** Interquartile range method.

The **interquartile range** (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.



This method is helpful if you have a few values on the extreme ends of your dataset, but you aren't sure whether any of them might count as outliers.

Interquartile range method

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

13. The p value is **a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true**. P values are used in hypothesis testing to help decide whether to reject the null hypothesis
14. The binomial distribution is given by the formula: **$P(X= x) = {}^{n}C_{x}p^{x}q^{n-x}$**, where = 0, 1, 2, 3, … P(X = 6) = 105/512. Hence, the probability of getting exactly 6 heads is 105/512.
15. Analysis of variance, or ANOVA, is **a statistical method that separates observed variance data into different components to use for additional tests**. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.