STATISTIC WORKSHEET - 5

1. Expected
2. Frequencies
3. 6
4. Chisquared Distribution
5. F Distribution
6. Hypothesis
7. Null Hypothesis
8. Two-tailed
9. Research hypothesis
10. np

MACHINE LEARNING

1) Re-squared is generally considered a better measure of goodness of fit in regression model compared to RSS. R-squared represents the proportion of variance in the dependent variable that is explained by the independent variables in the model, whereas RSS measures the total squared difference between the observed and predicted values. R-squared provides a standardized measure that ranges from 0 to 1,making it easier to interpret and compare across models, whereas RSS values are not standardized and can vary depending on th scale of the dependent variable.

2) In regression analysis-

- **a)** Total Sum of Squares(TSS) - It measures the total variability in the dependent variable(Y) around its mean. It represents the sum of the squared differences between each observed value of the dependent variable.
- **b)** Explained Sum of Squares(ESS) - It quantifies the variability in the dependent variable(Y) that is explained by the regression model. It represents the sum of the squared differences between the predicted values of the dependent variable(based on the regression model) and the mean of the dependent variable.
- **c)** Residual Sum of Squares(RSS)- It measures the variability in the dependent variable(Y) that is not explained by the regression model.It represents the sum of the squared differences between the observed values of the dependent variable and the predicted values(residual) from the regression model.

  The equation relating these three metrics is
  TSS = ESS + RSS
  Total sum of squares(TSS) is equal to the sum of the explained Sum of Squares(ESS) and the Residual Sum of Squares(RSS).This equation illustrates that the total variability in the dependent variable can be decomposed into the variabilty explained by the regression model(ESS) and the unexplained variability(RSS)

3) Regularization in machine learning is needed to pevent overfitting and improve the generalization of models. Overfitting occurs when a model learns to capture noise or random fluctuations in the training data rather than the underlying true pattern. Regularization techniques introduce a penalty term to the model's objective function, discouraging overly complex models that are prone to overfitting.

There are two main types of regularization:

- **a)** L1 Regularization(Lasso)- Adds a penalty term proportional to the absolute value of the co-efficients, encouraging sparsity of driving some coeffients to zero.This helps in feature selection by automatically selecting the most relevant features and reducing the mode's complexity.
- **b)** L2 Regularization(Ridge) - Adds a penalty term proportional to the square of the coefficients, which discourage large cofficients and helps to smooth out the model parameters. Ridge regularization helps in stabilizing the model by reducing the variance of the estimates, thus improving its generalization performance/

  Regularization technique helps in achieving a balance between bias and variance, leading to more robust and reliable models that perform well on unseen data. They are particularly useful when dealing with high-dimensional data or when the no. of features is much larger than the no. of sample, where overfitting is a common concern.

4) The Gini impurity index is a measurement used in decision tree algorithms for classification tasks. It quantifies the impurity or disorder of a set of data points by calculating the probability of incorrectly classifying a randomly chosen element in the dataset.

5) Yes, unregularized decision trees are prone to overfitting. Overfitting occurs when a model learns to capture noise or random fluctuations in the training data rather than the underlying relationship between features and the target variable. Unregularized decision trees have a high capacity for learning intricate patterns in the training data, which can lead to overly complex trees that memorize the training set instead of generalizing well to unseen data. Without regularization technique such as pruning or limiting the maximum depth of the tree,unregularized decision trees can grow excessively deep and complex, capturing noise and irrelevant details from the training data. This can result in poor performance on new, unseen data because the model has essentially memorized the training set rather than learned meaningful patterns. Regularization techniques help prevent overfitting by imposing constraints on the complexity of the decision tree, encouraging it to focus on the most important features and relationships in the data rather than fitting noise.

6) Ensemble technique in machine learning involve combining multiple models to improve predictive performance. Instead of relying on a single model, ensemble methods leverage the collective wisdom of multiple models to make more accurate predictions.

There are several types of ensemble techniques, including:

- **a)** Bagging(Bootstrap Aggregating): This technique involves training multiple instances of the same model on different subsets of the training data,typically using bootstrapping(sampling with replacement).The final prediction is then made by averaging the predictions of all individual models(for regression tasks) or taking a majority vote(for classsification tasks)
- **b)** Boosting: Boosting is an iterative ensemble technique where models are trained sequentially, with each subsequent model focusing on the mistakes made by the

previos ones. Examples include AdaBoost(Adaptive Boosting) and Gradient Boosting.

c) Random Forest : Random Forest is a popular ensemble method that combines the ideas of bagging and decisions trees. It builds multiple decision trees using random subsets of features and averages their predictions to reduce overfitting and improve generalization.

d) Stacking(Stacked Generalization): Stacking involves training multiple diverse models and then combining their prediction using another model(meta-learner).This meta-learner learns how to best combine the predictions of the base models to make the final prediction. Ensemble technique are powerful because they can often out.

7)

BAGGING	**BOOSTING**
a) Training data subsets are drawn randomly with replacement from the entire	**a)** Each new subset contains the components that were misclassified by training dataset. previous model.
b) Bagging attempts to tackle the over-fitting issue.	**b)** Boosting tries to reduce bias.
c) Every model receives an equal weight.	**c)** Models are weighted by their performance.
d) Objective to decrease variance, not bias.	**d)** Objective to decarese bias not variance.
e) Every model is built independently.	**e)** New models are affected by the performance of the previously developed model.

8) In a random forest algorithm, each decision tree is trained using a bootsrap sample of the data, which means some data points are not included in each trees's training set. The out-of-bag(OOB) error is calculated by evaluating each data point using only the trees for which it was not included in the training set (i.e, out-of-bag data points).This serves as a validation set for each individual tree. The OOB error is then aggregated across all trees to provide an overall estimate of the models's performance without the need for a separate validation set. It's a useful measure to assess the model's performance and generalization ability.

9) K-fold cross-validation is a technique used to evaluate the performance of a machine learning model. It involves partioning the dataset into K equal-sized subsets(or "folds"). The process then consists of the followings steps:

1. Iterating through each fold,using it as the test set while the remaining K-1 folds are used as the training set.
2. Training the model on the training set and evaluating its performance on the corresponding test set.
3. Calculating a performance metric(such as accuracy,precision,recall, or F1-score) for each fold.
4. Aggregating the performance metrics across all folds to obtain an overall performance estimate for the model.

K-fold cross-validation helps in assessing how well the model generalized to unseen data and reduces the variance in performance estimation compared to a single train-test split.Common choices for K include 5-fold and 10-fold cross-validation ,but other values can be chosen depending on the dataset size and computational resource available.

10) Hyperparameter tuning in machine learning refers to the process of selecting the optimal hyperparameters for a given model.Hyperparameters are parametersthat are not learned during the training process but are set before the training begins. They control the behavior of the learning algorithm and influence the performance and complexity of the model. Hyperparameter tuning is done for several reason:

1. Optimizing Model Performance:Selecting the right hyperparameters can significantly improve the performance of a model.For example,adjusting the learning rate or regularization parameter in a neural network can help prevent overfitting and improve generalization.
2. Generalization: Hyperparameter tuning helps ensure that the model generalizes well to unseen data.By finding the optimal hyperparameters,the model is better able to captureunderlying patterns in the data and make accurate predictions on new instances.
3. Avoiding Overfitting: Properly tuned hyperparameters can help prevent overfitting, where the model learns to memorize the training data instead of learning the underlying patterns.By fine-tuning hyperparameters, you can create a more balanced model that performs well on both the training and test datasets.
4. Improving Efficiency: Hyperparameter tuning can lead to more efficient models by optimizing computational resources.By selecting the right hyperparameters, you can often achieve the same level of performance with fewer computational resources, s such as training time or memory.

   Overall, hyperparameter tuning is a critical step in the machine learning pipeline that helps optimize model performance, generalization, and efficiency. It involves systematically searching through the hyperparameter space to find the combination that results in the best model performance.

11) Using a large learning rate in gradient descent can lead to several issues:

1. Divergence: Large steps may overshoot the minimum point, causing oscillations or instability in convergence.
2. Overshooting: Large steps may overshoot the minimum point, causing oscillations or instability in convergence.
3. Instability: Rapid changes in parameter values can lead to instability in the learning process.
4. Unpredictable Behaviour: The algorithm may exhibit erratic behaviour, making it difficult to predict or control.
5. Difficulty in Fine-Tuning: Large learning rates make it challenging to fine-tune the model as it may skip over smaller, more nuanced improvements.
6. Inaccurate Solutions: Large steps may cause the algorithm to settle in a suboptimal solution rather than reaching the global minimum.
7. Difficulty in Diagonising: With large learning rates, diagnosing issues in the optimization process becomes more complex as the cause of divergence or instability may be immediately apparent.

12) Logistic regression is a linear classifier, meaning it assumes a linear relationship between the features and the log-odds of the target variable.Therefore,it may not perform well on nonlinear data becauseit cannot capture complex nonlinear relationships between features and the target variable.

However, there are ways to adapt logistic regression for nonlinear data:

1. Feature Engineering: You can engineer nonlinear features from the original features to make the relationship between the features and the target variable more linear.
2. Polynomial Features: You can include polynomial featureof the original features to capture nonlinear relationships. This can be done ny adding squared,cubic, or higher-order terms of the features.
3. Kernel Trick: You can use a kernel trick to implicitly map the original features into a higher-dimensional space where they might become linearly separable.However,this approach is more common in support vector machines(SVMs) rather than logistic regression.

   If the data is highly nonlinaer and cannot be effectively transformed into a linear space,other nonlinear classifiers such as decision tress.random forests, support vector machines with nonlinear kernels, or neural networks may be more suitable than logistic regression.

13)

	ADABOOST	**GRADIENT BOOSTING**
OBJECTIVE FUNCTION	It minimizes the exponential loss function by giving more weight to misclassified data points in each iteration.	It minimizes a predefined differentiable loss function(e.g., squared error for regression, logistic loss for classifiaction) by fitting subsequent models to the residuals or gradient of the loss function.
WEIGHT UPDATING	It updates the weights of incorrectly classified examples at each iteration, focusing more on difficult examples in subsequent rounds.	It fits each subsequent model to the residuals or gradient of the loss function of the previous model, emphasizing the mistakes made bythe previous model.
BASE LEARNERS	It typically uses decision trees with a shallow depth(often called "stumps") as weak learners.	It can use various types of weak learners, often decision tress, but they are typically deeper and more complex compared to Adaboost.

PARALLELISM	It's sequential since each subsequent model depends on the performance of the previous one.	It can be parallelized because the training of each new model depends only on the residuals or gradients of the loss function, which can be computed independently for each data point.
ROBUSTNESS TO NOISE DATA	It's sensitiveto noisy data and outliers, as it tries to fit them correctly by assigning them higher weights.	It's less sensitive to noisy data and outliers due to its focus on minimizing a differentiable loss function, which tends to smooth out noise

14) The bias-variance tradeoff is a fundamental concept in machine learning that refers to the balance between the bias of the model and its variance.
Bias- Bias measures how closely the average prediction of a model matches the true value. A high bias indicates that the model is too simplistic and fails to capture the underlying patterns in the data.Models with high bias are often too simple and mayunderfit the data.

Variance: Variance measures the variability of the model's predictions across different datasets.A high variance indicates that the model is highly sensitive to small fluctuations in the training data and captures noise along with the underlying patterns.

Models with high variance are often too complex and may overfit the training data.
The tardeoff arises because reducing bias often increases variance and vice-versa.Finding the right balance is crucial for building models that generalize well to unseen data.

- **a)** Underfitting: Occurs when the model is too simple(high bias) to capture the underlying structure of the data, resulting in poor performance on both the training and test datasets.
- **b)** Overfitting:Occurs when the model is too complex(high variance) and captures noise along with the underlying patterns in the traing data, resulting in good performance on the training dataset but poor performance on the test dataset.

  The goal in machine learning is to find a model that achieves low bias and low variance, stiking the right balance between ssimplicity and complexity.Techniques such as cross-validation,regularization and model selection help manage the bias-variance tradeoff and build models that generalized well to new data.

15) Here's a brief description of each kernel used in Support Vector Machines(SVM):

- **a)** Linear kernel:
  Description: The linear kernel computes the dot product of the feature vectors in the original space.It works well for linearly separable data.
  Use case: Suitable for linearly separable or nearly linearly separable data where classes can be separated by a straight line or hyperplane.

b) RBF(Radial Basis Function) kernel:
Description: The RBF kernel computes the similarity between vectors in a higher-dimensional space by mapping them into an infinite-dimensional space.It uses a Gaussian function to measure similarity.
Use case: Suitable for data that is not linearly separable and where decision boundaries are complex or nonlinear.RBF kernel can capture complex relationships between features and target classes.

c) Polynomial kernel:
Description: The polynomial kernel computes the similarity between feature vectors in a higher-dimensional space using polynomial functions.It maps the input space into a higher-dimesional space,allowing thr SVM to find nonlinear decision boundaries.
Use case: Suitable for data that is not linearly separable and where decision boundaries are polynomial in nature.Polynomial kernel can capture nonlinear relationship between features and target classes.

Each kernel function has its own characteristics and is suitable for diffferent types od data and problem scenarios.The choice of kernel depends on the underlying data distribution and the complexity of the decision boundaries needed to separate the classes effectively.