STATISTICS WORKSHEET-5

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.

ANS d) Expected

2. Chi-square is used to analyses

ANS c) Frequencies

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

ANS c) 6

4. Which of these distributions is used for a goodness of fit testing?

ANS b) Chi-square distribution

5. Which of the following distributions is Continuous

ANS c) F Distribution

6. A statement made about a population for testing purpose is called?

ANS b) Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

ANS a) Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

ANS a) Two tailed

9. Alternative Hypothesis is also called as?

ANS 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

ANS a) np

MACHINE LEARNING

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?

ANS **The residual sum of squares (RSS)** is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term. The smaller the residual sum of squares, the better your model fits your data; the greater the residual sum of squares, the poorer your model fits your data. The RSS is used by financial analysts in order to estimate the validity of their econometric models.

whereas R-squared is the absolute amount of variation as a proportion of total variation

R-squared does not measure goodness of fit. It can be arbitrarily low when the model is completely correct. By making σ2 large, we drive R-squared towards 0, even when every assumption of the simple linear regression model is correct in every particular.

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.

ANS **The Total SS (TSS or SST)** - The total sum of squares is a variation of the values of a dependent variable from the sample mean of the dependent variable. Essentially, the total sum of squares quantifies the total variation in a sample. It can be determined using the following formula:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Where:

- y_i – the value in a sample
- $\bar{y}$ – the mean value of a sample

Regression sum of squares (also known as the sum of squares due to regression or ESS)- The regression sum of squares describes how well a regression model represents the modeled data. A higher regression sum of squares indicates that the model does not fit the data well.

The formula for calculating the regression sum of squares is:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Where:

$\hat{y}_i$ – the value estimated by the regression line

$\bar{y}$ – the mean value of a sample

Residual sum of squares (also known as the sum of squared errors of prediction)- The residual sum of squares essentially measures the variation of modeling errors. In other words, it depicts how the variation in the dependent variable in a regression model cannot be explained by the model. Generally, a lower residual sum of squares indicates that the regression model can better explain the data, while a higher residual sum of squares indicates that the model poorly explains the data. The residual sum of squares can be found using the formula below:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

y_i – the observed value

$\hat{y}_i$ – the value estimated by the regression line

The relationship between the three types of sum of squares can be summarized by the following equation:

$$TSS = SSR + SSE$$

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

ANS Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "In regularization technique, we reduce the magnitude of the features by keeping the same number of features."

4. What is Gini–impurity index?

ANS Gini Impurity is a measurement used to build Decision trees. Trees to determine how the feature of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassifies if it were given a random class label according to the class distribution in the dataset.

For example , say you want to build a classifier that determines if someone will default on their credit card. You have some labeled data with features, such as bins for age, income, credit rating, and whether or not each person is a student. To find the best feature for the first split of the tree- the root node- you could calculate how poorly each feature divided the data into the correct class, default(‘yes’) or didn’t default “no”. This calculation would **measure the impurity** of split, and the feature with the lowest impurity determine the best feature for splitting the current node .

$$Gini(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

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

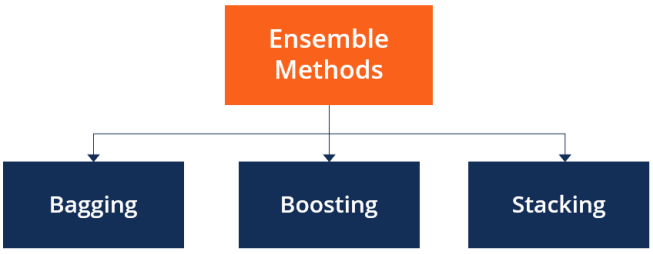
ANS Unregularized Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

An example of this could be predicting if the Boston Celtics will beat the Miami Heat in tonight's basketball game. The first level of the tree could ask if the Celtics are playing home or away. The second level might ask if the Celtics have a higher win percentage than their opponent, in this case the Heat. The third level asks if the Celtic's leading scorer is playing? The fourth level asks if the Celtic's second leading scorer is playing. The fifth level asks if the Celtics are traveling back to the east coast from 3 or more consecutive road games on the west coast. While all of these questions may be relevant, there may only be two previous games where the conditions of tonight game were met.

Using only two games as the basis for our classification would not be adequate for an informed decision. One way to combat this issue is by setting a max depth. This will limit our risk of overfitting; but as always, this will be at the expense of error due to bias. Thus, if we set a max depth of three, we would only ask if the game is home or away, do the Celtics have a higher winning percentage than their opponent, and is their leading scorer playing. This is a simpler model with less variance sample to sample but ultimately will not be a strong predictive model.

6. What is an ensemble technique in machine learning?

ANS Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.



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

ANS **Bagging-** Bagging, the short form for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models. Bagging is classified into two types, i.e., bootstrapping and aggregation

Bagging is advantageous since weak base learners are combined to form a single strong learner that is more stable than single learners. It also eliminates any variance, thereby reducing the overfitting of models. One limitation of bagging is that it is computationally expensive. Thus, it can lead to more bias in models when the proper procedure of bagging is ignored.

Boosting -Boosting is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future. The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models. Boosting works by arranging weak learners in a sequence, such that weak learners learn from the next learner in the sequence to create better predictive models

Why is bagging better than boosting

From the dataset, bagging creates extra data for training. Random sampling and substitution from the original dataset is used to achieve this. In each new training data set, sampling with replacement may repeat certain observations. Every Bagging element has the same chance of emerging in a fresh dataset. Multiple models are trained in parallel using these multi datasets. It is the average of all the forecasts from several ensemble models. When determining classification, the majority vote obtained through the voting process is taken into account. Bagging reduces variation and fine-tunes the prediction to a desired result.

How are the main differences bagging and boosting?

Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification. It attempts to increase the weight of an observation if it was erroneously categorized. Boosting creates good predictive models in general.

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

ANS **Out-of-bag (OOB)** error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from. OOB error is the mean prediction error on each training sample xi, using only the trees that did not have xi in their bootstrap sample.

Bootstrap aggregating allows one to define an out-of-bag estimate of the prediction performance improvement by evaluating predictions on those observations that were not used in the building of the next base learner.

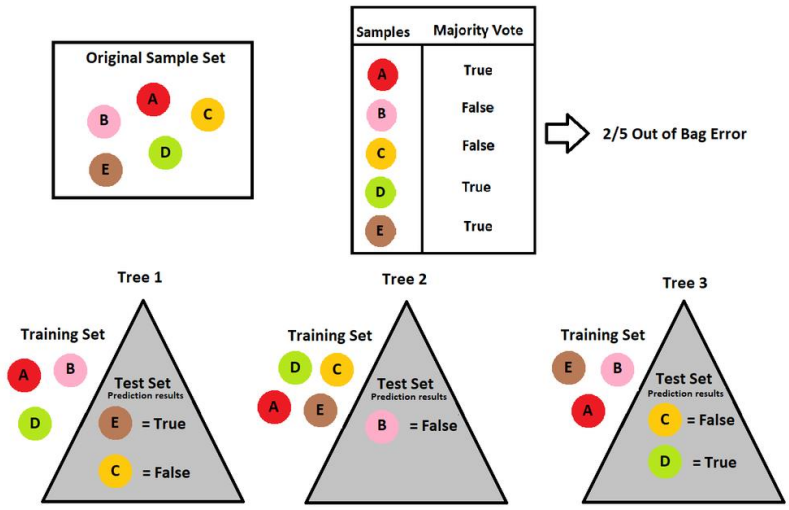
Calculating out of bag error

Since each out-of-bag set is not used to train the model, it is a good test for the performance of the model. The specific calculation of OOB error depends on the implementation of the model, but a general calculation is as follows.

1 Find all models (or trees, in the case of a random forest) that are not trained by the OOB instance.

2 Take the majority vote of these models' result for the OOB instance, compared to the true value of the OOB instance.

3 Compile the OOB error for all instances in the OOB dataset.



Shown in the example to the right, the OOB error can be found using the method above once the forest is set up.

9. What is K-fold cross-validation?

ANS **K-Fold Cross-Validation**

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

10. What is hyper parameter tuning in machine learning and why it is done?

ANS **Hyperparameter tuning** consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

We consider these essential hyperparameters for tuning SVMs:

C: A tradeoff between a smooth decision boundary (more generic) and a neat decision boundary (more accurate for the training data). A low value may cause the model to incorrectly classify some training data, while a high value may cause the model to incur overfitting. Overfitting creates an analysis too specific for the current data set and possibly unfit for future data and unreliable for future observations.

Gamma: The inverse of the influence radius of data samples we selected as support vectors. High values indicate the small radius of influence and small decision boundaries that do not consider relatively close data samples. These high values cause overfitting. Low values indicate the significant effect of distant data samples, so the model can’t capture the correct decision boundaries from the data set.

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

ANS I've implemented my own gradient descent algorithm for an OLS, code below. It works, however, when the learning rate is too large (i.e., learn rate >= .3), my approach is unstable. The coefficients explode and I get an overflow error. I understand that if my learning rate is too large, I get bad results. The algorithm will take too big of steps and continuously miss the optima.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

ANS Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.

Logistic regression is neither linear nor is it a classifier. The idea of a "decision boundary" has little to do with logistic regression, which is instead a direct probability estimation method that separates predictions from decision.

13. Differentiate between AdaBoost and Gradient Boosting.

ANS **AdaBoost** - AdaBoost or Adaptive Boosting is the first Boosting ensemble model. The method automatically adjusts its parameters to the data based on the actual performance in the current iteration. Meaning, both the weights for re-weighting the data and the weights for the final aggregation are re-computed iteratively.

In practice, this boosting technique is used with simple classification trees or stumps as base-learners, which resulted in improved performance compared to the classification by one tree or another single base-learner.

Gradient Boosting - Gradient Boost is a robust machine learning algorithm made up of Gradient descent and Boosting. The word ‘gradient’ implies that you can have two or more derivatives of the same function. Gradient Boosting has three main components: additive model, loss function and a weak learner.

The technique yields a direct interpretation of boosting methods from the perspective of numerical optimization in a function space and generalizes them by allowing optimization of an arbitrary loss function.

The Comparison:

1.Loss Function-The technique of Boosting uses various loss functions. In case of Adaptive Boosting or AdaBoost, it minimizes the exponential loss function that can make the algorithm sensitive to the outliers. With Gradient Boosting, any

differentiable loss function can be utilized. Gradient Boosting algorithm is more robust to outliers than AdaBoost.

2.Flexibility-AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

3.Benefits- AdaBoost minimizes loss function related to any classification error and is best used with weak learners. The method was mainly designed for binary classification problems and can be utilized to boost the performance of decision trees. Gradient Boosting is used to solve the differentiable loss function problem. The technique can be used for both classification and regression problems.

Shortcomings

In the case of Gradient Boosting, the shortcomings of the existing weak learners can be identified by gradients and with AdaBoost, it can be identified by high-weight data points.

14. What is bias-variance trade off in machine learning?

ANS In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters. The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.

The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting)

15. Give short description each of Linear, RBF, Polynom

ANS **Linear Regression** - is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable,

criterion variable, endogenous variable. The independent variables can be called exogenous variables, predictor variables, or regressors. Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable

Radial basis function (RBF) networks are a commonly used type of artificial neural network for function approximation problems. Radial basis function networks are distinguished from other neural networks due to their universal approximation and faster learning speed. An RBF network is a type of feed forward neural network composed of three layers, namely the input layer, the hidden layer and the output layer. An RBF network with a specific number of nodes (i.e. 10) in its hidden layer is chosen.

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

y= b0+b1x1+ b2x12+ b2x13+..... bnx1n

It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

It is a linear model with some modification in order to increase the accuracy. The dataset used in Polynomial regression for training is of non-linear nature. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.