UNIT-III

Name: K.Mahesh
Regdno: 18B91A0595
Sec: CSE(B)

1) Identify the statistics of central tendency, statistics of dispersion and shape statistics for the following dataset:

suppose that the data for analysis includes the attribute age. the age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

A) Central Tendency:

The measures of central tendency are Mean, Median, Mode.

1) Mean:

Mean is defined as the sum of all of the numbers divided by the number of numbers in a given data.

$$\text{Mean} = \frac{\text{Sum of observations}}{\text{Total no of observations}}$$

$$= \frac{809}{27} = 29.963$$

2) Median:

Median is the middle number in a sorted, ascending or descending order.

If n is odd then Median is $\left(\frac{n+1}{2}\right)^{th}$ term

If n is even then $\frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$

Here $n = 27$ (odd)

So Median $= \left(\frac{n+1}{2}\right)^{th}$ term

$= \left(\frac{27+1}{2}\right)^{th}$ term

$= 14^{th}$ term

$\therefore$ Median = 25

3) **Mode:**

Mode is the value that appears most frequently in a data set.

In given data 25 & 35 are repeated most times with frequency '4'

∴ Mode = 25, 35.

Statistics & Dispersion:

The measures of Dispersion are Range, Quartiles, Mean Deviation, Standard Deviation and variance

1) **Range:**

Range is defined as the difference between Maximum value and minimum value in a given data

In given data, Maximum value = 70, minimum value = 13

Range = 70 − 13 = 57

2) **Quartiles:**

The formula for Quartiles is given by

Lower Quartile $(Q_1) = (n+1) * \frac{1}{4}$

Middle Quartile $(Q_2) = (n+1) * \frac{2}{4}$

Upper Quartile $(Q_3) = (n+1) * \frac{3}{4}$

Inter Quartile Range $= Q_3 - Q_1$

Here $n = 27$.

$Q_1 = \left(\frac{27+1}{4}\right) \times \frac{1}{4} = 7^{th}$ term $= 20$

$Q_2 = (27+1) \times \frac{2}{4} = 14^{th}$ term $= 25$

$Q_3 = (27+1) \times \frac{3}{4} = 21^{th}$ term $= 35$

Inter Quartile Range $= Q_3 - Q_1$

$= 35 - 20 = 15$

3) Mean Deviation:

It is defined as statistical measure that is used to calculate average deviation from mean value of given data set

Formula is $\frac{1}{n}\sum_{i=1}^{n} |x_i - x_m|$

where x_m = mean

here mean = 29.963

$\cong 30$

$\therefore$ Mean Deviation $= \frac{17+13+14+14+11+10+10+9+3+3+5+5+5+\text{total}=\ldots}{27}$

$= 10.03$

4) Standard Deviation & variance:

The standard Deviation is a statistic that measures dispersion of dataset relative to its mean and is calculated as squareroot of variance.

Variance refers to statistical measurement of spread between numbers in a data set. It measures how far each number in sets is from mean and thus from every other number in set.

variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$

$= \frac{17^2+13^2+14^2+14^2+\ldots}{27}$

$= 155.29$

standard deviation $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2}$

$= \sqrt{155.29}$

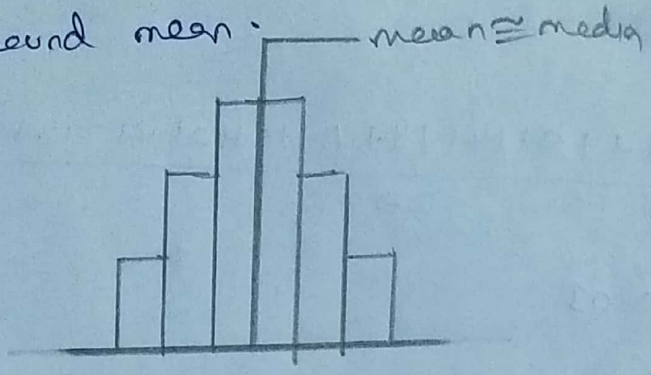
$= 12.46$

Shape statistics:

Symmetric Data Distribution, skewed Data Distributions are shape statistics

Symmetric Data Distribution:

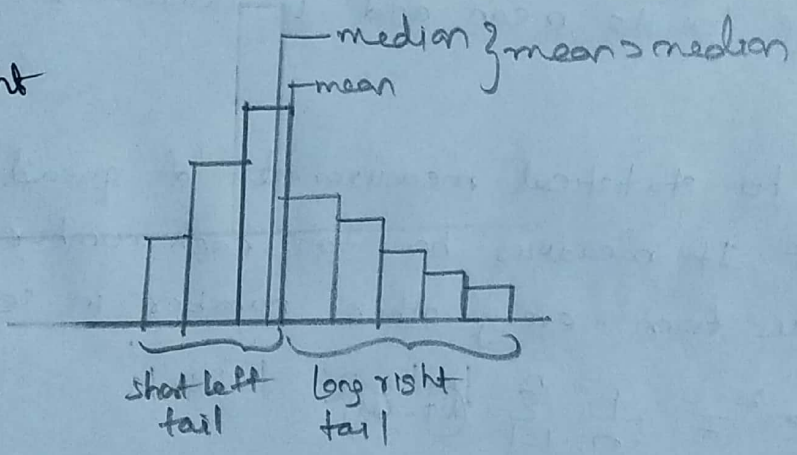
Where left & right hand sides of Distribution are roughly equally balanced around mean.



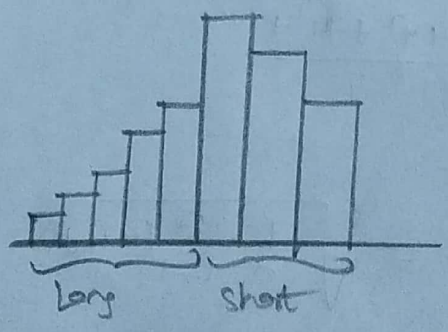
Skewed Data Distributions:

A Distribution that is skewed right (known as positively skewed)

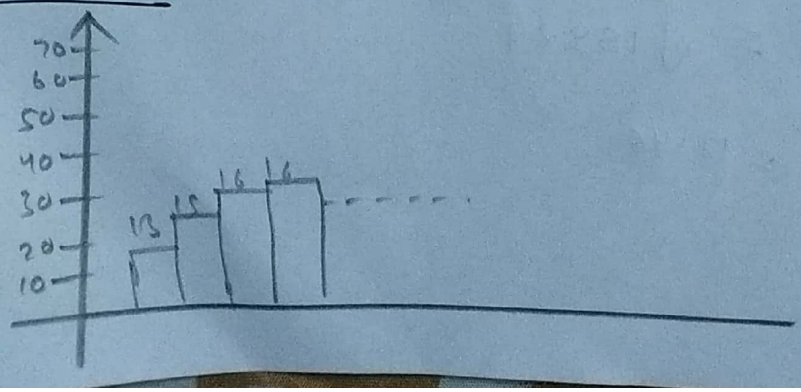
Skewed right



Skewed Left



For given data:



2) Compare bagging ensemble models and boosting ensemble models

A) Bagging:

* Bagging is used when the goal is to reduce the variance of a decision tree classifier.
* Suppose there are N observations and M features in training Data set. A Sample from training data set is taken randomly with replacement.
* A subset of M features are selected randomly and whichever feature gives best split is used to split node iteratively.
* The tree is grown to largest.
* Above steps are repeated n times and prediction is given based on aggregation of predictions from n number of trees.

Advantages:

* Reduces over-fitting of the model.
* Handles higher dimensionality data very well.
* Maintains accuracy for missing data.

Disadvantages:

* It may result in high bias if it is not modelled properly and thus may result in underfitting.
* since we must use multiple models, it becomes computationally expensive and may not be suitable in various use cases.
* since final prediction is based on the mean predictions from subset trees, it won't give precise values for classification and regression model.

Boosting:

* Boosting is used to create a collection of predictors.
* When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model.
* Draw a random subset of training samples d_1 without replacement from training set D to train a weak learner C_1.
* Draw second random training subset d_2 without replacement from training set and add 50 percent of samples that were previously falsely classified/misclassified to train a weak learner C_2.
* Find the training samples d_3 in training set D on which C_1 & C_2 disagree to train a third weak learner C_3.
* Combine all weak learners via majority voting.

Advantages:

* Supports different loss function (we have used 'binary:logistic' for this example)
* works well with interactions.
* It is good at handling the missing data.

Disadvantages:

* Boosting is hard to implement in real-time due to increased complexity of algorithm.
* Prone to over-fitting.
* Requires careful tuning of different hyper-parameters.

3) Compare Gradient Boosting algorithm and XGBoost

A)

Gradient Boosting Algorithm:

* The gradient is used to minimise the loss function (error difference between the actual values and predicted values).
* It is basically the partial derivative of loss function, so it describes the steepness of our error function.
* In each round of training, the weak learner is built and its predicted values are compared to actual values. The distance between prediction and reality represents the error rate of our model.
* Take the derivative (gradient) of loss function (error) of each parameter. Calculate the step size and learning rate and calculate new parameters based on that. In this way, you will create a new weak learner.
* keep repeating the steps (descending the gradient) and keep generating new learners until step size is very small or maximum number of steps are completed.
* By using gradient descent and updating our predictions based on learning rate (the stepsize with which we descend gradient), we can find the values where loss function is minimum.
* So, we are basically updating predictions such that sum of our residuals is close to 0 (or minimum) and predicted values are sufficiently close to actual values.

XGBoost (Extreme Gradient Boosting):

* XGBoost stands for Extreme Gradient Boosting. XGBoost is a specific implementation of the Gradient Boosting method which delivers more accurate approximations by using the strengths of second order derivative of loss function, L1 and L2 regularisation and parallel computing.

* XGBoost is particularly popular because it has been the winning algorithm in a number of recent Kaggle competitions.

* XGBoost is more regularised form of Gradient Boosting.

* XGBoost uses advanced regularisation (L1 & L2) which improves model organisation capabilities.

* XGBoost delivers high performance as compared to Gradient Boosting. Its training is very fast and can be parallelized/ distributed across clusters.

* XGBoost computes second-order gradients. i.e second parallel derivatives of the loss function, which provides more information about the direction of gradients and how to get to the minimum of our loss function.

* XGBoost also handles missing values in the dataset. So, in data ~~wrangler~~ wrangling, you may or may not do a seperate treatment for missing values, because XGBoost is capable of handling missing values internally.

4) Explain Feature Construction and Selection.

A) Feature Construction:

* Feature Construction (also known as constructive induction or attribute discovery) enriches data by adding derived features.

* These are useful in a data analysis pipeline if they capture relevant relationships within the data that downstream processes are unable to readily model or exploit.

* They may also be useful in explainable AI if they make explicit relationships that would otherwise be implicit and difficult to comprehend.

* Our pioneering research demonstrated that feature construction can empower machine learning systems to construct more accurate models across a wide range of learning tasks.

* It is done to create new features based on original descriptors to improve the accuracy of the predictive model.

* It involves transforming a given set of input features to generate a new set of more powerful features which are used for prediction.

Feature Selection:

* Feature selection is one of the main components of feature engineering, is the process of selecting the most impotant features to input in machine learning algorithms

Feature selection methods:

Feature selection methods are categorised as either supervised or unsupervised. upsupervised techniques are classified as filter methods, wrapper methods, embedded methods.

a) Filter methods:

Filter methods select features based on statistics rather than feature selection cross-validation performance. A selected metric is applied to identify irrelevant attributes and perform recursive feature selection.

b) Wrapper methods:

Wrapper feature selection methods consider the selection of a set of features as a search problem, whereby their quality is assessed with the preparation, evaluation and comparision of a combination of features to other combination of features.

c) Embedded Methods:

Embedded feature selection methods integrate the feature selection machine learning algorithm as part of learning algorithm, in which classification and feature selection are performed simultaneously. The features that will contribute the most to each iteration of the model training process are carefully extracted. Random forest feature selection, decision tree feature selection, and LASSO feature selection are common embedded methods.

Advantages of feature selection:

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret
* It improves the accuracy of a model if the right subset is chosen.
* It reduces overfitting.

5) Compare Bagging and random forests.

A) Bagging:

* Bagging, short for 'bootstrap aggregating', is a simple but highly effective ensemble method that creates diverse models on different random samples of the original dataset. These samples are taken uniformly with replacement and are known as bootstrap samples. Because samples are taken with replacement the bootstrap sample will in general contain duplicates. Hence some of the original data points will be missing even if the bootstrap sample is same size as original data set.

Bagging Algorithm:

Input: dataset D; ensemble size T; learning algorithm A.

output: ensemble of models whose predictions are to be combined by ~~wrong~~ voting or averaging.

1 for t=1 to T do
2 build a bootstrap sample Dt from D by sampling |D| data points ~~with~~ with replacement;
3 run A on Dt to produce a model Mt;
4 end
5 return {Mt | 1 ≤ t ≤ T}

Random forests:

Bagging is particularly useful in combination with tree models, which are quite sensitive to variations in the training data. When applied to tree models, bagging is often combined with another idea; to build each tree from a different random subset of the features, a process also referred to as subspace sampling. This encourages the diversity in the ensemble even more and has the additional advantage that the training time of each tree is reduced. The resulting ensemble method is called random forests.

Random Forests Algorithm:

Input: data set D; ensemble size T; subspace dimension d.

Output: ensemble of tree models. whose predictions are to be combined by voting or averaging.

1. for $t=1$ to T do
2. build a bootstrap sample D_t from D by sampling $|D|$ data points with replacement;
3. select d features at random and reduce dimensionality of D_t accordingly;
4. train a tree model M_t on D_t without pruning;
5. end
6. return $\{M_t \mid 1 \leq t \leq T\}$

6) Explain the role of thresholding and discretisation in feature transformations.

A) Thresholding and discretisation :

* Discretisation transforms a quantitative feature into an ordinal feature.
* Each ordinal value is called a bin and corresponds to an interval of the Orisinal quantitative feature.
* There are two approaches to discretization: supervized and unsupervized.
* Unsupervised discretisation methods typically require one to decide the number of bins before hand.

Ex: Equal frequency discretization, Equal width discretization and univariate clustering.

Unsupervised discretization:

* A simple method that often works reasonably well is to choose the bins so that each bin approximately the same number of instances; this is referred to as equal frequency ~~distribution~~. discretization.
* another unsupervised discretization method is equal width discretization, which chooses the bin boundaries so that each interval has same width.

* An interesting alternative is to treat feature discretization as a univariate clustering problem.
* For example, in order to generate k bins we can uniformly sample k initial bin centres and run k-means until convergence.

Supervised Discretization:

* There are of two types: top-down or divisive discretization and bottom-up or agglomerative.
* Divisive methods work by progressively splitting bins.
* Agglomerative methods proceed by initially assigning each instance to its own bin and successively merging bins.
* In either case an important role is played by the stopping criterion, which decides whether a further split or merge is worthwhile.
* A natural generalisation of thresholding leads to top down recursive partitioning algorithm.
* This discretization algorithm finds the best threshold according to some scoring function Q, and proceeds to recursively split left & right bins.
* One scoring function that is often used is information gain.

7) Demonstrate Adaboost and Gradient Boosting

A) Adaboost:

* This is a type of ensemble technique, where a number of weak learners are combined together to form a strong learner. Here, usually each weak learner is developed as decision stumps (A stump is a tree with just a single split and two terminal nodes) that are used to classify observations.

* Here each classifier has diff weights assigned to it based on classifiers performance, weights are also assigned to observations at end of every round, in such a way that wrongly predicted observations have increased weight resulting in their probability of being picked more often in next classifier's sample.

* Thus Adaboost increases the predictive accuracy by assigning weights to both observations at end of every tree and weights to every classifier.

* Hence, in Adaboost every classifier has a different weightage on final prediction contrary to the random forest where all trees are assigned equal weights.

Gradient Boosting:

* Just like AdaBoost, Gradient Boost also combines a no of weak learners to form a strong learner.
* Here, the residual of the current classifier becomes the input for the next consecutive classifier on which trees are built, and hence it is an additive model.
* The residuals are captured in a step-by-step manner by the classifiers, in order to capture the maximum variance within data, this is done by introducing Learning rate of classifiers.
* Initially a tree with single node is built which predicts the aggregated value of Y in case of regression or odds of Y for classification problems after which trees with greater depth are grown on previous classifier's residuals.
* Learning rates are given as constant for every tree so that the model takes small steps in the right direction to capture the variance and the train the classifier on it.
* Unlike Adaboost here all trees are given equal weights.