1. Mean, Median, Mode

Mean: the average of a set of values. The mean is used for the imputation when the data points in the given data set are not far apart. This is so that we get an appropriate value rather than a one which is influenced by a data point which is far from the other points. The mean can be used only on numerical data. When a numerical column/feature has missing values, we impute it with mean.

Mathematically represented as = (Sum of n elements)/n

Median: It’s the middle element in a given set of data values. Median is used when we have one or more data values that can influence the mean towards it because of the large difference with other data points. In this case, we first arrange the elements in ascending order and then choosing the middle number. In case where there the number of data points is even, then choose the 2 middle number and find the average between them. Again, median is used only for numerical data. When a numerical column/feature has missing values, we impute it with mean.

Mode: It’s a method of choosing the most frequently occurring element in the given list of data values. Mode is applicable for only categorical data. No mathematical operation is performed other than just keeping the count of the occurrence of elements unlike mean. When a categorical column/feature has missing values, we impute it with mode.

1. ChiSquare Test:

This test works only for categorical data. It tests whether distributions of categorical data vary from one another. Mathematically represented as ((O-E)\*(O-E))/(E) where O is the observed value and E is the expected value with respect to ‘c’ independent parameters.

A low value of chi square means, high correlation between the two sets of data.

1. Naïve Bayes

P(A|B) = P(B|A) \* P(A)/P(B)

It’s a relation between the conditional probability and independent probability between two events. This helps us find reverse probabilities. This makes an assumption of conditional independence. that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors

that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors. that the effect of the value of a predictor (*B*) on a given class (*A*) is independent of the values of other predictors.

1. Confusion Matrix

A confusion matrix is table of numerical representation of how many data points are correctly predicted, incorrectly predicted. Its generated for a classifier problem. It’s a performance metric that describes how well a classifier performs. It compares the actual with the predicted values. Following are the components of the matrix

|  |  |  |
| --- | --- | --- |
| CONFUSION MATRIX | Predicted True | Predicted False |
| Actual True | True positive (TP) | False negative (FN) |
| Actual False | False positive (FP) | True negative (TN) |

TP: Actually positive and correctly predicted as positive.

TN: Actually negative and correctly predicted as negative.

FP: Actually negative but incorrectly predicted as positive.

FN: Actually positive but incorrectly predicted as negative.

Following are few other metrics obtained from the confusion matrix

Accuracy = (TP+TN)/(TP + TN + FP + FN)

Recall = (TP) + (TP +FN) i.e total number of correctly classified as positive/ total number of actual positive.

Precision = (TP) + (TP + FP) i.e total number of correctly classified as positive/ total number of predicted positive.

F1 score = (2 \* recall \* precision)/(recall + precision). This is better than accuracy in case of uneven class distribution.

1. Normal distribution and 3 sigma ranges. Assumptions of Linear regression

It’s a probability distribution which is symmetric about the mean, indicating the data around the mean is more frequent than the one’s further from the mean.

A close up of a device

Description automatically generated

A data set that has a normal distribution, usually has its data spread out in the following ranges

* 68% of data lies between -1 to +1std deviation
* 95% of data lies between -2 to +2 std deviation
* 99.7% data lies between -3 t0 +3 std deviation

Assumptions of linear regression:

* The model must be linear i.e plot of feature/target must be straight line.
* The error terms must be independent.
* The error terms must have constant variances (homoscedastic)
* Error terms must be normally distributed.

1. Outlier check using z-score, IQR and cook’s distance

Z-score: It’s a numerical representation of data points relative distance to the groups mean in terms of standard deviation.

Z-score = (Point – mean)/ standard deviation of group of points. It tells you how far the data point is from the mean in terms of standard deviation.

IQR (Interquartile range): It’s a measure of where the bulk of the values lie. For every group of values, we can find the IQR. With the help of this IQR we can determine the lower and upper whisker of a data set using box plot. The lower and upper whisker are the limits/threshold beyond which if any data point lies, can be treated as an outlier.

Cook’s distance: It’s a measure of an observations tendency of influencing the Linear regression. It is used to filter outliers. Influence is a combination of leverage and distance (residuality on y axis). For each point leverage is calculated. Sum of all leverages is equal to number of parameters being estimated in regression model.