Worksheet 4 - Statistics
Submitted By: Jwala R

1) D
2) A
3) A
4) C
5) A
6) A
7) C
8) B
9) B
10) A special type of bar graph called a histogram displays bars for a range of data values rather than just one. A box plot is a type of data visualization where the interquartile range of the data is shown by drawing a box over a number line. A box plot's "whiskers" display the data set's lowest and highest values.
11) Best Predicted vs Human, BPH:
Kendall's tau coefficient shows the correlation between the two lists of ranked items based on the number of concordant and discordant pairs in a pairwise: in each case we have two ranks (machine and human prediction). Firstly, the ranked items are turned into a pairwise comparison matrix with the correlation between the current rank and others. A concordant pair means an algorithm rank correlates with a human rank. Otherwise, this will be a discordant pair. Therefore, this coefficient is defined as following:

The values of τ varies from 0 to 1. The closer |τ| is to 1, the better ranking is. For instance, when τ-value is close -1, the ranking is just as accurate, however the order of its items should be vice-a-versa. This is quite consistent with estimate indicators which assign the highest rank to the best values, whereas during manual human ranking the best ones receive the lowest ranks. τ-value=0 indicates the lack of any correlation between ranks.

REGRESSION
performance metrics

MEAN ABSOLUTE ERROR (MAE)

This regression metric indicates the average sum of absolute difference between the actual and predicted value.

MEAN SQUARE ERROR (MSE)

Mean Squared Error (MSE) calculates the average sum of squared difference between the actual and predicted value for the entire data points. All related values are raised to the second power therefore all of negative values are not compensated by positives. Moreover, due to the features of this metric, the impact of errors is higher. For example, if the error in our initial calculations is 1/2/3, MSE will equal 1/4/9 respectively. The less MSE is, the more accurate our predictions is. MSE = 0 is the optimal point in which our forecast is perfectly accurate.

MSE has some advantages over MAE:
1. MSE highlights large errors over small ones.
2. MSE is differentiable which helps find minimum and maximum values using mathematical methods more effectively.

ROOT MEAN SQUARE ERROR (RMSE)

RMSE is a square root of MSE. It is easy to interpret compared to MSE and it uses smaller absolute values which is helpful for computer calculations.

CONFUSION MATRIX
This matrix is used to evaluate the accuracy of a classifier and is presented in the table below.
Some examples

False Positive (FP) moves a trusted email to junk in an anti-spam engine.
False Negative (FN) in medical screening can incorrectly show desease absense, when it is actually positive.

ACCURACY METRIC
This metric is the basis one. It indicates the number of correctly classified items compared to the total number of items.
Keep in mind that accuracy metric has some limitations: it doesn't work well with unbalanced classes that can have many items of the same class and few other classes.

RECALL/SENSITIVITY METRIC

Recall Metric shows how many True Positives the model has classified from the total number of positive values.

PRECISION METRIC

This metric represents the number of True Positives which are really positive compared to the total number of positively predicted values.

F1 SCORE

This metric is a combination of precision and recall metrics which serves as a comprise. The best F1 score equals 1, while the worst one is 0.

12) The use of hypothesis testing is necessary to determine statistical significance. The alternative and null hypotheses would be presented first. Second, you would determine the p-value, which represents the probability of receiving the test's results if the null hypothesis is correct. When the p-value is less than the threshold of significance (alpha), indicating that the result is statistically significant, you would then choose to reject the null hypothesis.

13) There is no Gaussian or log-normal distribution for exponential distributions. In actuality, categorical data of any kind won't have these distributions either.

14) Since income has a skewed distribution, it is a prime example of a situation in which the median should be used instead of the mean. According to the median, 50% of all incomes are below 27581 and 50% are above it. The mean overestimates the range of household incomes for these data.

15) The likelihood is not a probability density over the parameter; rather, it is the likelihood that a specific event is seen when the real value of the parameter is, which is identical to the probability mass. The posterior probability of giving the data is not to be confused with the likelihood, which is not the same thing.