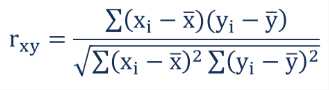
**ICT583 Data Science Applications**

**Murdoch University**

**Exercise 3**

1. **Identify three successful and well-used scoring functions in areas of personal interest to you. For each, explain what makes it a good scoring function and how it is used by others. (20 points)**
2. **Correlation**

Correction gives the relationship between the variables. Correlation is represented by “r” and is calculated by



Where x and y are the observation in a dataset.

If,

* + r < 1, then the two observations are negatively correlated. If one increase, others decreases.
  + r > 1, then the two observations are positively correlated. If one increase, others also increases.
  + r = 0, then there is no relationship between the observations.

This is a balanced formula in numerator and denominator. In the numerator, we find the difference between mean and the observation of x and y, and calculate the sum of their products, gives you a small value. In the denominator, we find the square of the difference and square root which is also a small value. So the division of these values gives a small value, which lies between (-1, +1)

This can be used to find the relationship between the attributes. For instance, In a machine learning model design, highly correlated attributes are removed to reduce the multicollinearity effect (large variance).

1. **The nature of roots for a quadratic equation**



Discriminant

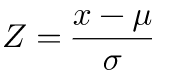
* If Δ < 0, Roots non-real i.e., imaginary
* If Δ > 0, roots are not equal and there are 2 roots possible
* If Δ = 0, roots are equal and only one root
* If Δ >= 0, roots are real and unequal

In the above equation, we are, squaring the coefficient "b" and subtracting from the four times the product of coefficients "a" and "c". Which helps in find the nature of roots in quadratic equations.

This is important to double-check the roots when calculated in different methods such as using quadratic formulae, factoring and square roots method etc.

1. **P-value from Z score**

Z score is given by



Where

* Z is a standard score
* *X* is the Observed value
* µ is the mean of the sample
* σ is the Standard Deviation

Z score is positive if the values lie above mean, and negative otherwise.

This is one of the examples of a good score function because, in the numerator, we are subtracting the test score with the mean of the samples, which gives the distance from the mean. And then dividing the result with Standard deviation. This helps in calculating the test statistics, one tail and two-tail tests.

Z score helps in calculating the P-value in the Hypothesis testing. Ideally

* If the P-value is less than 0.05 then it is statistically significant
* If the p-value is great than 0.05 then it is statistically not significant

1. **Suppose we want to train a binary classifier where one class is very rare. Give two examples of such a problem. How should we train this model? What metrics would you consider to measure the performance? (20 points)**

The examples of imbalance classes are

1. Credit card fraud detection

Fraud credit card cases can be very less when compared to non-fraud cases, resulting in class imbalance

1. Diagnosing a patient for cancer or not (Malignant and Benign)

The unwanted cell growth in an area can result in tissue formation. This tissue can be due to cancer cells or just cell growth. If the tissue is cancer, then it is called as Malignant, if it is due to cell growth, then it is called Benign. Cancer cells were grown in size and could be dangerous to life, on the other hand, benign tissue stops the growth after a certain point of time and can be removed through surgeries.

In general malignant cases will be less compared to benign cases. Diagnosing cancer patients is a challenging task.

1. Identifying spam and non-spam emails.

**How should we train this model?**

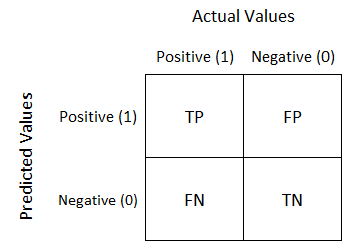
The following methods/techniques can be used to train the imbalanced data:

1. Resampling:
2. Oversampling the minority class
3. Under-sampling majority class
4. Finding the minority class.
5. Removing some of the samples randomly in the majority class.
6. Weigh the minority class heavily, by taking care of oversampling issues.
7. Replication minority class, with random perturbation.

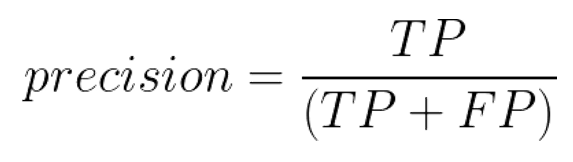
**What metrics would you consider to measure the performance?**

Accuracy is not the best way to find the performance of the model, so the following performance metrics such as:

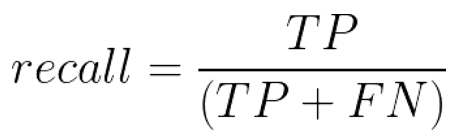
1. Confusion matrix



1. Precision



1. Recall and



1. F1: score

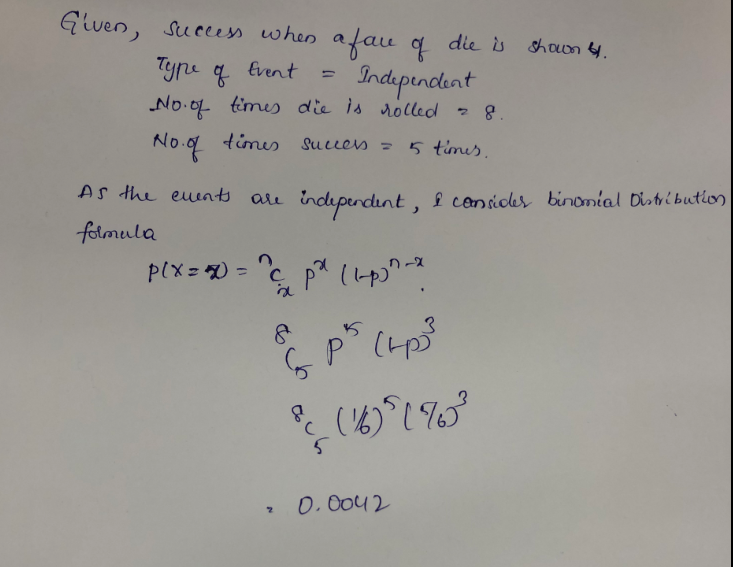


1. ROC-AUC

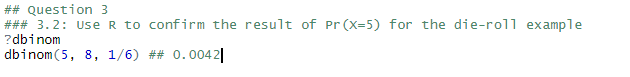
Can be used.

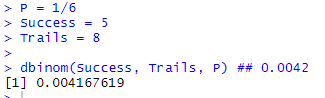
1. **Suppose you are rolling a die with success defined as getting a 4. If you roll the die independently eight times: (30 points)**

**3.1 What is the probability of observing exactly five successes (five 4s) in total? (calculated by hand)**



**3.2 Use R to confirm the result of Pr(X=5) for the die-roll example.**

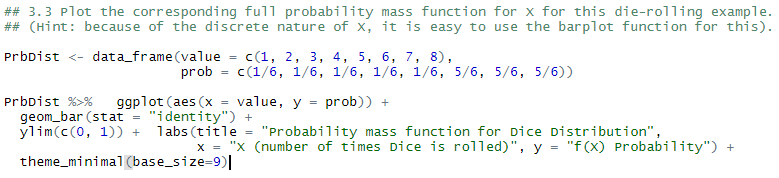


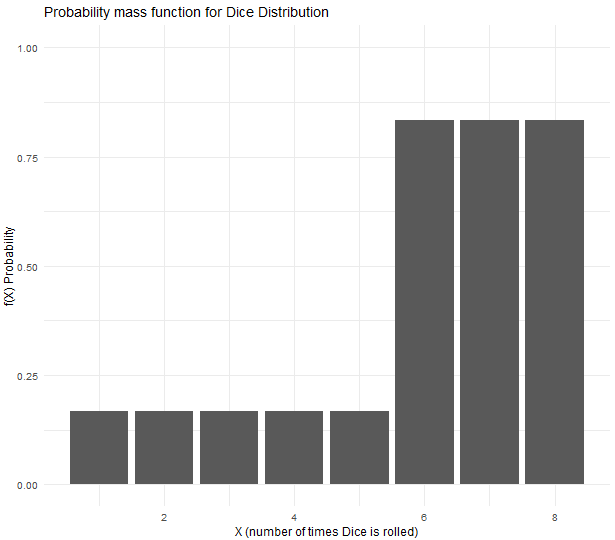


Both the results are balanced.

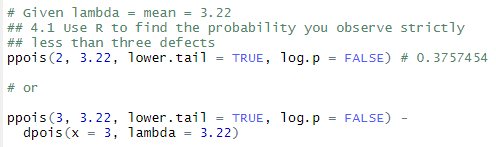
**3.3 Plot the corresponding full probability mass function for X for this die-rolling example.**

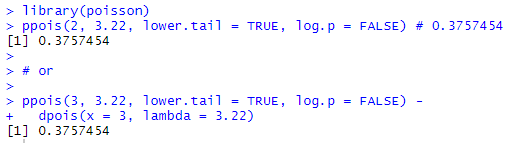
**(Hint: because of the discrete nature of X, it is easy to use the bar plot function for this).**





1. **Consider the example of defects on 1-foot-square sheets of metal coming off a production line. Suppose you’re told that the number of defects found, X, is thought to follow a Poisson distribution with λ = 3.22. In other words, you’d expect to see an average of 3.22 defects on your 1-foot sheets. (30 points)**
   1. **Use R to find the probability you observe strictly less than three defects.**



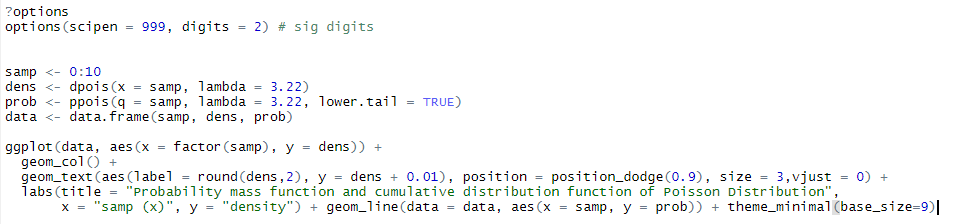


* 1. **Use R to find the probability you observe strictly more than five defects.**





**4.3 Plot the corresponding Poisson probability mass function and cumulative distribution function for λ = 3.22 plotted for the integers 0 ≤ x ≤ 10.**



The probability mass function and Cumulative distribution function for a Poisson distribution is shown below

