**SGI ASSIGNMENT**

**Q/A Assignment:**

1. In logistic regression, the weights associated with each feature determine the impact of that feature on the outcome predicted. We trained the logistic regression with a certain set of features and learned weights W0, W1,….Wn. Feature n gets Wn at the end of the training. Now we created a new dataset where we duplicated feature n into n+1 and retrained the model, then likely relationship of the new weights will be:

* The new weights of the new models will be as that of the weights of the old model because we trained the new model also on the same set of features as that of the old model only change is the duplication of nth feature.
* The weight n+1 which is the new duplication of the weight n is likely to like weight n of the old feature n. This is because the model while training will assign almost the same weights as that of the old model as it is training on the same set of data. One extra duplication will not impact the data huge.

Finally, the weights of the new model will be same as that of the old model and the new weight for the duplicate feature “n+1” will be same as that of the old weight “n**”.**

**2)** With 95% confidence, we may say that template B is inferior to template A and that template E is superior to template A. For a 95% confidence level, we must run the test longer to compare templates C and D to template A.

The z-scores for templates B, C, and D are all smaller than 1.96, as you can see, indicating that we are not 95% certain that they differ from template A. On the other hand, template E's z-score is higher than 1.96, indicating a 95% confidence level that it performs better than template A.

Thus, it would be advisable to keep running the test for a longer amount of time until we are 95% certain of the outcomes for templates C and D.

So, the correct option is **option B. (i.e.,** E is better than A with over 95% confidence, B is worse than A with over 95% confidence. You need to run the test for longer to tell where C and D compared to A with 95% confidence.)

**3**) The number of non-zero entries in each training example (k), the total number of training examples (m), and the number of features (n) determine the computational cost of each logistic regression gradient descent iteration. This is because the gradient of the cost function with respect to each feature for every training example is computed during each gradient descent iteration, and this computation is only required for non-zero entries. As a result, the computing cost is **O(kmn).**

**4**) **Ranking:**

**Approach 3**: This approach, which concentrates on fixing mistakes caused by V1 and choosing cases that are distant from the decision border, may be the most successful. It directly tackles V1's issues.

**Approach 1**: Although it deals with circumstances that are unclear, it might not offer a wide range of examples. Nevertheless, it might be useful for fine-tuning the model close to the decision border.

**Approach 2**: This offers a random sample, which is useful for generalization, but it may not concentrate on mistake correction or enhancing the model's performance close to the decision boundary.

**5) MLE Estimate =** k/n.

**Bayesian Estimate** = k+1/n+2.

**MAP Estimate** = k/n.