**Q/A Assignment**

1. You currently have an email marketing template A and you want to replace it with a better template. A is the control\_template. You also test email templates B, C, D, E. You send exactly 1000 emails of each template to different random users. You wish to figure out what email gets the highest click through rate. Template A gets 10% click through rate (CTR), B gets 7% CTR, C gets 8.5% CTR, D gets 12% CTR and E gets 14% CTR. You want to run your multivariate test till you get 95% confidence in a conclusion. Which of the following is true?
   1. We have too little data to conclude that A is better or worse than any other template with 95% confidence.
   2. 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 compare to A with 95% confidence.
   3. Both D and E are better than A with 95% confidence. Both B and C are worse than A with over 95% confidence

Sol : Option b seems to be true as E is superior to A & A is superior to B.

1. You have m training examples and n features. Your feature vectors are however sparse and average number of non-zero entries in each train example is k and k << n. What is the approximate computational cost of each gradient descent iteration of logistic regression in modern well written packages?

Sol : In modern packages like scikit-learn or TensorFlow, these operations are optimized to make the most of the sparse nature of the data. This optimization significantly reduces the computational cost per gradient descent iteration to roughly m\*k. This is a substantial improvement compared to the dense scenario where the cost would be proportional to m\*n.

1. We are interested in building a high quality text classifier that categorizes news stories into 2 categories - information and entertainment. We want the classifier to stick with predicting the better among these two categories (this classifier won't try to predict a percent score for these two categories). You have already trained V1 of a classifier with 10,000 news stories from the New York Times, which is one of 1000 new sources we would like the next version of our classifier (let's call it V2) to correctly categorize stories for. You would like to train a new classifier with the original 10,000 New York Times news stories and an additional 10,000 different news stories and no more. Below are approaches to generating the additional 10,000 pieces of train data for training V2.
   1. Run our V1 classifier on 1 Million random stories from the 1000 news sources. Get the 10k stories where the V1 classifier’s output is closest to the decision boundary and get these examples labeled.
   2. Get 10k random labeled stories from the 1000 news sources we care about.
   3. Pick a random sample of 1 million stories from 1000 news sources and have them labeled. Pick the subset of 10k stories where the V1 classifier’s output is both wrong and farthest away from the decision boundary.

Ignore the difference in costs and effort in obtaining train data using the different methods described above. In terms of pure accuracy of classifier V2 when classifying a bag of new articles from 1000 news sources, what is likely to be the value of these different methods?How do you think the models will rank based on their accuracy?

**Sol:**

* The third approach seems promising as it addresses cases where the V1 model made errors, and the examples are far from the decision boundary. This could lead to more significant improvements in classifier accuracy.
* The first approach, by selecting examples near the decision boundary, could enhance the model's ability to handle ambiguous cases.
* The second approach, while providing diversity, may not specifically target areas of weakness identified by the V1 model.

1. You wish to estimate the probability, $p$ that a coin will come up heads, since it may not be a fair coin. You toss the coin $n$ times and it comes up heads $k$ times. You use the following three methods to estimate $p$
   1. Maximum Likelihood estimate (MLE)
   2. Bayesian Estimate: Here you assume a continuous distribution uniform prior to $p$ from $[0,1]$ (i.e. the probability density function for the value of $p$ is uniformly $1$ inside this range and $0$ outside. Our estimate for $p$ will be the expected value of the posterior distribution of $p$. The posterior distribution is conditioned on these observations.
   3. Maximum a posteriori (MAP) estimate: Here you assume that the prior is the same as (b). But we are interested in the value of $p$ that corresponds to the mode of the posterior distribution.

What are the estimates?

**Sol :** 1) Maximum Likelihood Estimate = k/n

2) Bayesian Estimate = (k+1)/(n+2)

3) MAP = k/(n+1)