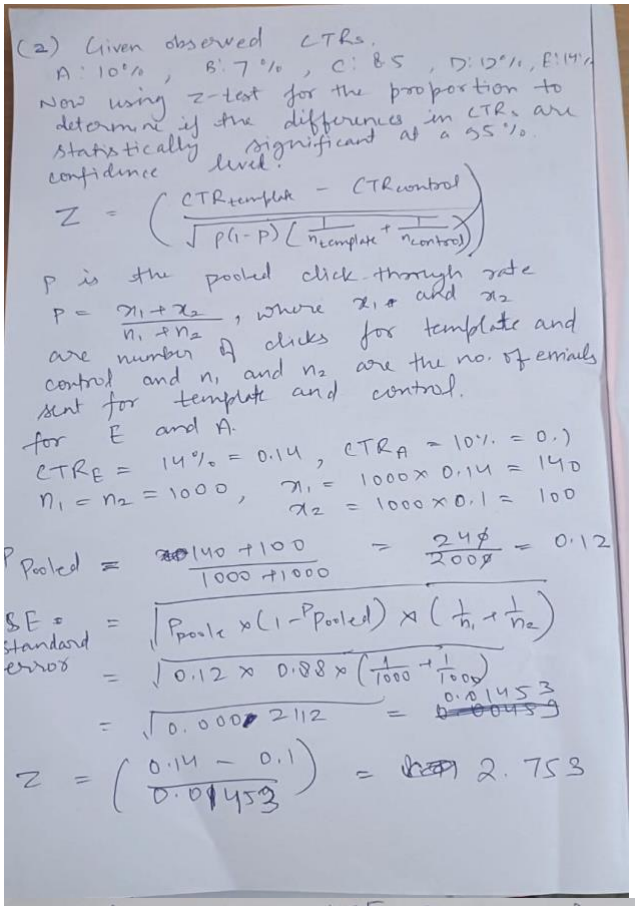
Q/A Assignment

1. When we duplicate a feature in logistics regression and retain the model, the new weights w_{new} and $w_{new_{(n+1)}}$ assigned to the original and the duplicated feature would be almost half the original weight w_n, assuming all other conditions remain the same. This is because the contribution to the prediction that was previously made by feature n is now being shared by feature n and n+1.

$$w_{new_n} + w_{new_{n+1}} \approx w_n$$

2.



(2) Given observed CTRs.
A: 10%, B: 7%, C: 8.5, D: 12%, E: 14%
Now using z-test for the proportion to determine if the differences in CTRs are statistically significant at a 95% confidence level.

$$Z = \left(\frac{CTR_{template} - CTR_{control}}{\sqrt{P(1-P)\left(\frac{1}{n_{template}} + \frac{1}{n_{control}}\right)}} \right)$$

P is the pooled click-through rate

$P = \frac{x_1 + x_2}{n_1 + n_2}$, where x_1 and x_2 are number of clicks for template and control and n_1 and n_2 are the no. of emails sent for template and control.

for E and A.

$CTR_E = 14\% = 0.14$, $CTR_A = 10\% = 0.1$

$n_1 = n_2 = 1000$, $x_1 = 1000 \times 0.14 = 140$, $x_2 = 1000 \times 0.1 = 100$

$P_{Pooled} = \frac{140 + 100}{1000 + 1000} = \frac{240}{2000} = 0.12$

SE = Standard error $= \sqrt{P_{pooled} \times (1 - P_{Pooled}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$= \sqrt{0.12 \times 0.88 \times \left(\frac{1}{1000} + \frac{1}{1000}\right)}$

$= \sqrt{0.0002112} = 0.01453$

$Z = \left(\frac{0.14 - 0.1}{0.01453} \right) = 2.753$

Similarly, $Z_B = -2.405$, $Z_C = -1.158$

$Z_D = 1.429$, $Z_E = 2.752$

$Z_{critical} = 1.96$

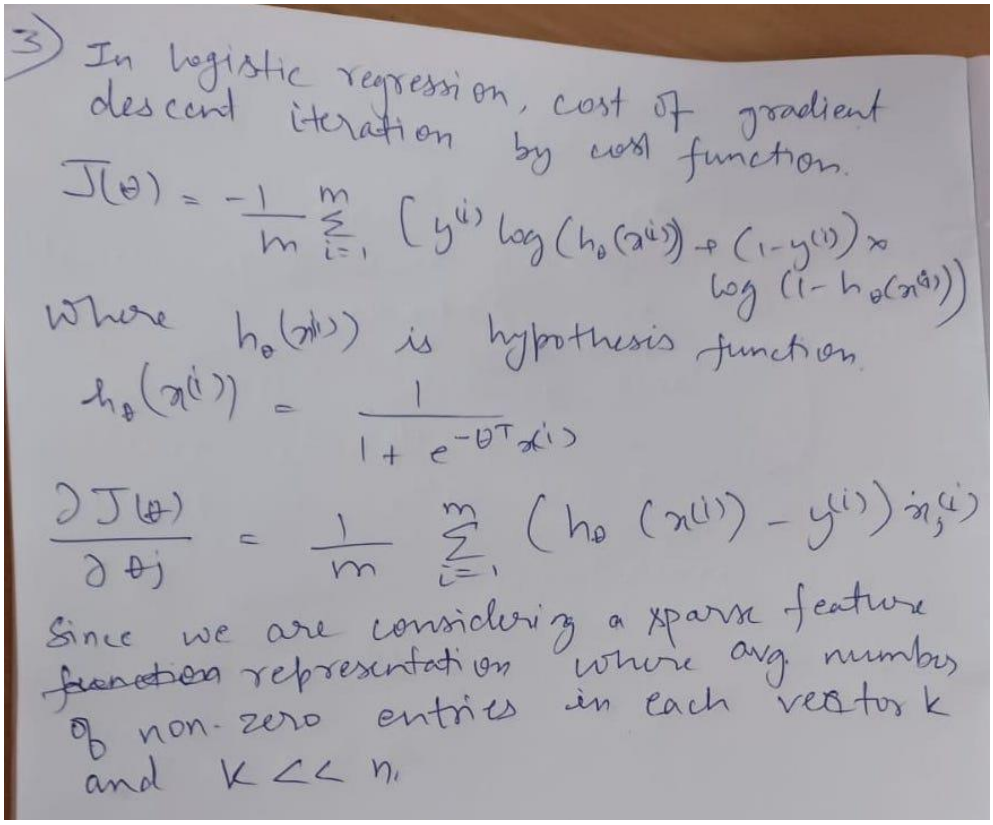
Hence if Z-score is greater than 1.96 or less than -1.96, the difference is considered statistically significant at 95% confidence level.

Hence correct option.

(b) E is better than A with over 95% confidence, B is worse than A with 95% confidence.

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.



For each training example i, the hypothesis function $h_\theta(x^i)$ requires computing $\theta^T x^i$ which has a computational cost proportional to the number of non-zero entries, k, since we ignore the zeros in the sparse matrix-vector multiplication. This makes the cost for computing the hypothesis for all m examples O(mk). The gradient computation also requires us to multiply $(h_\theta(x^i) - y^i \; with \; each \; x_j^i$. Since there are k non-zero features on average for each training example, and we need to update all n features, the naïve computational cost would be O(mnk). However, due to the sparsity, we only need to perform updates for non-zero features, reducing the cost to O(mk) for the gradient computation.

4. Assuming the goal is to improve the accuracy of V2, analyzing how each approach might influence the classifier's performance:

Uncertainty Sampling	Random Sampling	Boosting hard
This method selects examples where the classifier is most uncertain. Training on these examples can help the classifier to better define the decision boundary, which may be particularly useful when the classes are not well-separated. This can lead to a more refined model that performs better on examples where it was previously uncertain.	This method provides a random set of labeled examples. While it ensures a variety of examples, it might not be as efficient as other methods in improving the decision boundary since it's not targeted. This approach might still improve the classifier's overall performance but potentially requires more data to achieve significant improvements compared to targeted methods.	By selecting examples that the classifier currently gets wrong and are farthest from the decision boundary, this method focuses on the most challenging cases. Training on these hard examples can significantly improve the classifier's performance, especially if the current errors are systematic or if there's a particular subset of the data that V1 struggles with.

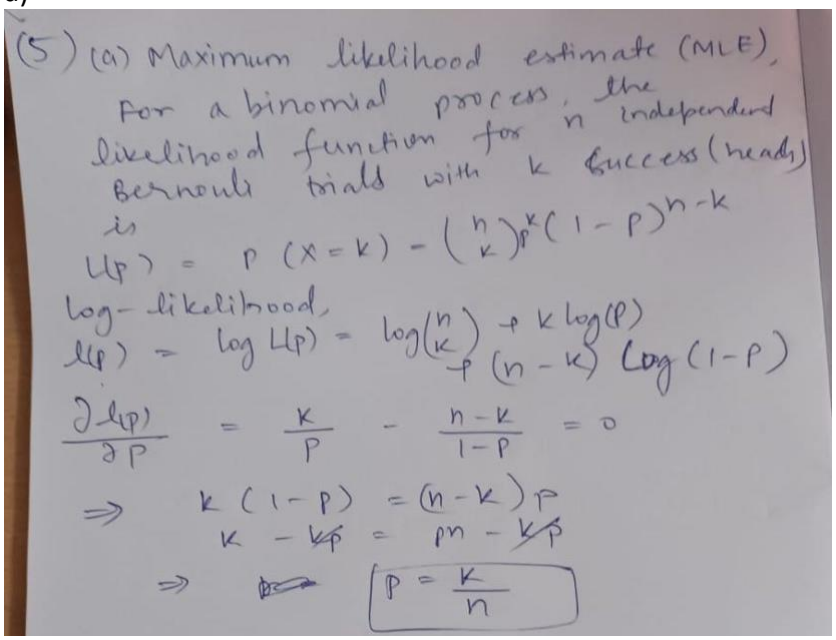
Hence in terms of pure accuracy, without considering the cost and efforts.

- Method a might increase accuracy significantly because it helps the classifier to refine the decision boundary.
- Method b might yield a less significant accuracy improvement because it might include many examples that are easy to classify and thus may not contribute much to learning.
- Method c is likely to yield a significant improvement in accuracy, particularly if V1's mistakes are not random but systematic. By correcting these mistakes, V2 can potentially make large gains in accuracy.

  Therefore, in terms of expected improvements in accuracy for V2

1) Method c
2) Method a
3) Method b

5. a)


(5) (a) Maximum likelihood estimate (MLE),
For a binomial process, the likelihood function for n independent Bernoulli trials with k success (heads) is

$L(p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

Log-likelihood,

$l(p) = \log L(p) = \log\binom{n}{k} + k\log(p) + (n-k)\log(1-p)$

$\frac{\partial l(p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0$

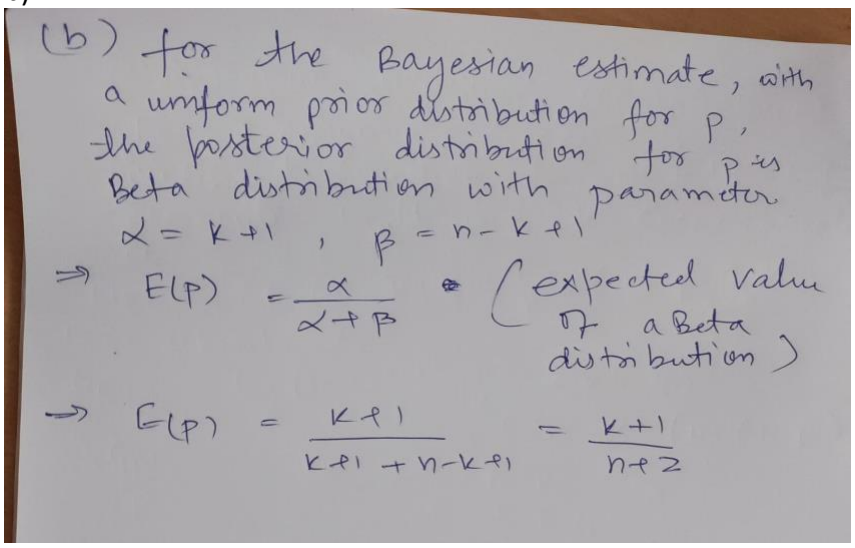
$\Rightarrow k(1-p) = (n-k)p$

$k - kp = pn - kp$

$\Rightarrow p = \frac{k}{n}$

here p estimate is $p = \frac{n}{k}$

b)


(b) for the Bayesian estimate, with a uniform prior distribution for p, the posterior distribution for p is Beta distribution with parameter

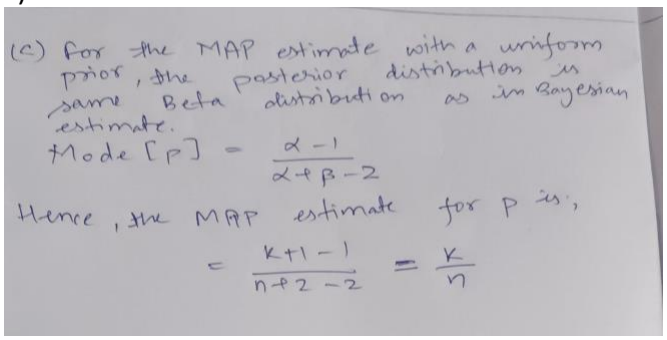
$\alpha = k+1$, $\beta = n-k+1$

$\Rightarrow E(p) = \frac{\alpha}{\alpha+\beta}$ * (expected value of a Beta distribution)

$\rightarrow E(p) = \frac{k+1}{k+1+n-k+1} = \frac{k+1}{n+2}$

Here p estimate is $\frac{k+1}{n+2}$

c)


(c) For the MAP estimate with a uniform prior, the posterior distribution is same Beta distribution as in Bayesian estimate.

Mode $[p] = \frac{\alpha-1}{\alpha+\beta-2}$

Hence, the MAP estimate for p is,

$= \frac{k+1-1}{n+2-2} = \frac{k}{n}$

here p estimate is $p = \frac{n}{k}$

The MLE and MAP turn out to be the same in this case due to the uniform prior, which does not influence the estimation heavily. The Bayesian estimate, however, takes into account the prior and is more conservative, especially when n is small.