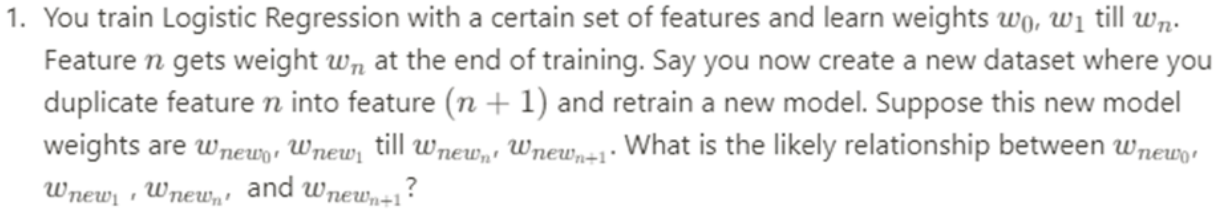
Q/A Assignment

Name: Hardik Jain
Roll Number: BT20ECE094 [IIIT NAGPUR]
Mail:jainh445@gmail.com



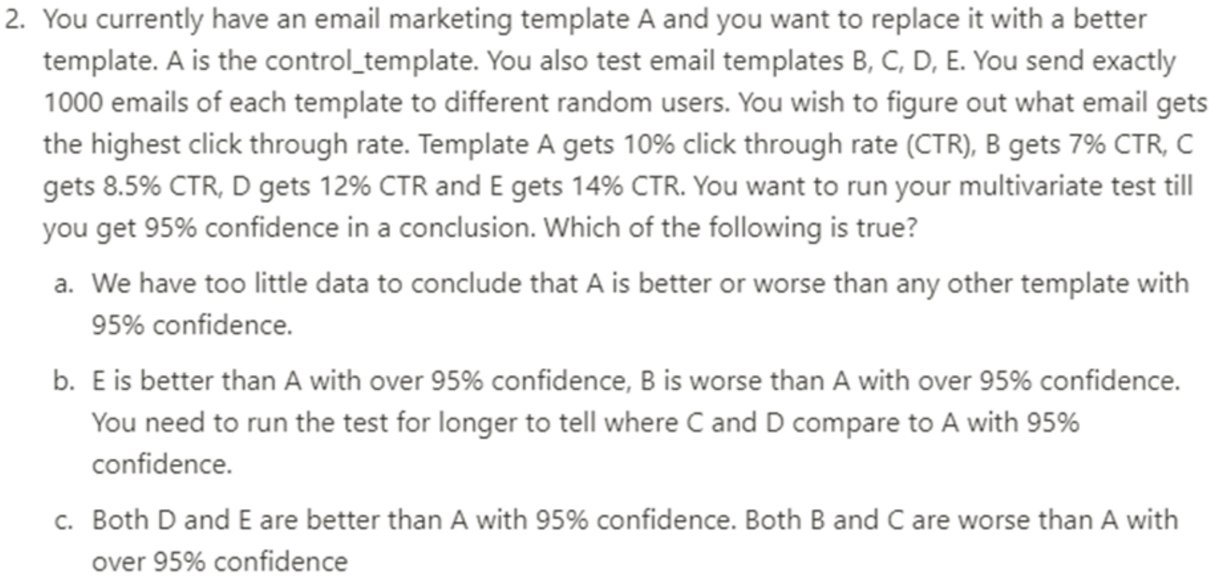
1. You train Logistic Regression with a certain set of features and learn weights w_0, w_1 till w_n. Feature n gets weight w_n at the end of training. Say you now create a new dataset where you duplicate feature n into feature $(n+1)$ and retrain a new model. Suppose this new model weights are w_{new_0}, w_{new_1} till w_{new_n}, $w_{new_{n+1}}$. What is the likely relationship between w_{new_0}, w_{new_1} , w_{new_n}, and $w_{new_{n+1}}$?

Solution

When duplicating feature n into n+1 and retraining logistic regression, weights $(w_{new})_n$ and $(w_{new})_{n+1}$ likely distribute to handle redundant information, influenced by optimization factors like learning rate and regularization.

If there is collinearity between the original feature n and the duplicated feature n+1, the weights may be unstable, and small changes in the training data could lead to significant changes in the learned weights.

The specific relationship depends on the algorithm used and data characteristics.



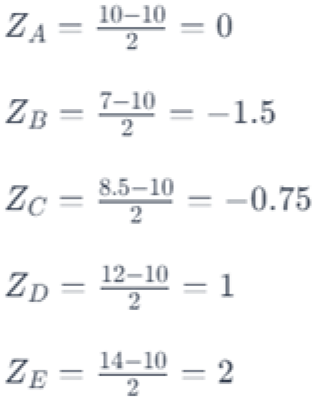
2. You currently have an email marketing template A and you want to replace it with a better template. A is the control_template. You also test email templates B, C, D, E. You send exactly 1000 emails of each template to different random users. You wish to figure out what email gets the highest click 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?
   a. We have too little data to conclude that A is better or worse than any other template with 95% confidence.
   b. 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.
   c. Both D and E are better than A with 95% confidence. Both B and C are worse than A with over 95% confidence

Solution

B option is correct

To better visualize it, we can normalize (Z-score) each template.

Let's μ is 10% and the standard deviation σ is 2%.



$$Z_A = \frac{10-10}{2} = 0$$

$$Z_B = \frac{7-10}{2} = -1.5$$

$$Z_C = \frac{8.5-10}{2} = -0.75$$

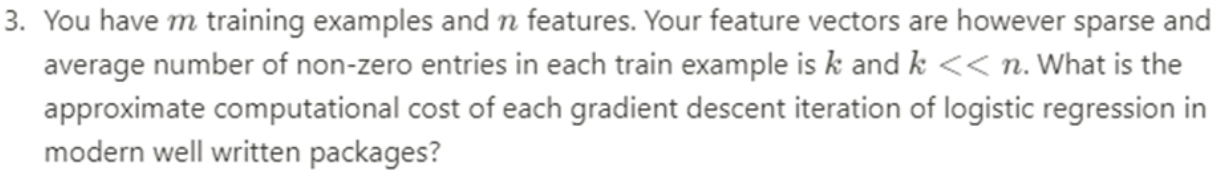
$$Z_D = \frac{12-10}{2} = 1$$

$$Z_E = \frac{14-10}{2} = 2$$

By analysing z-scores,

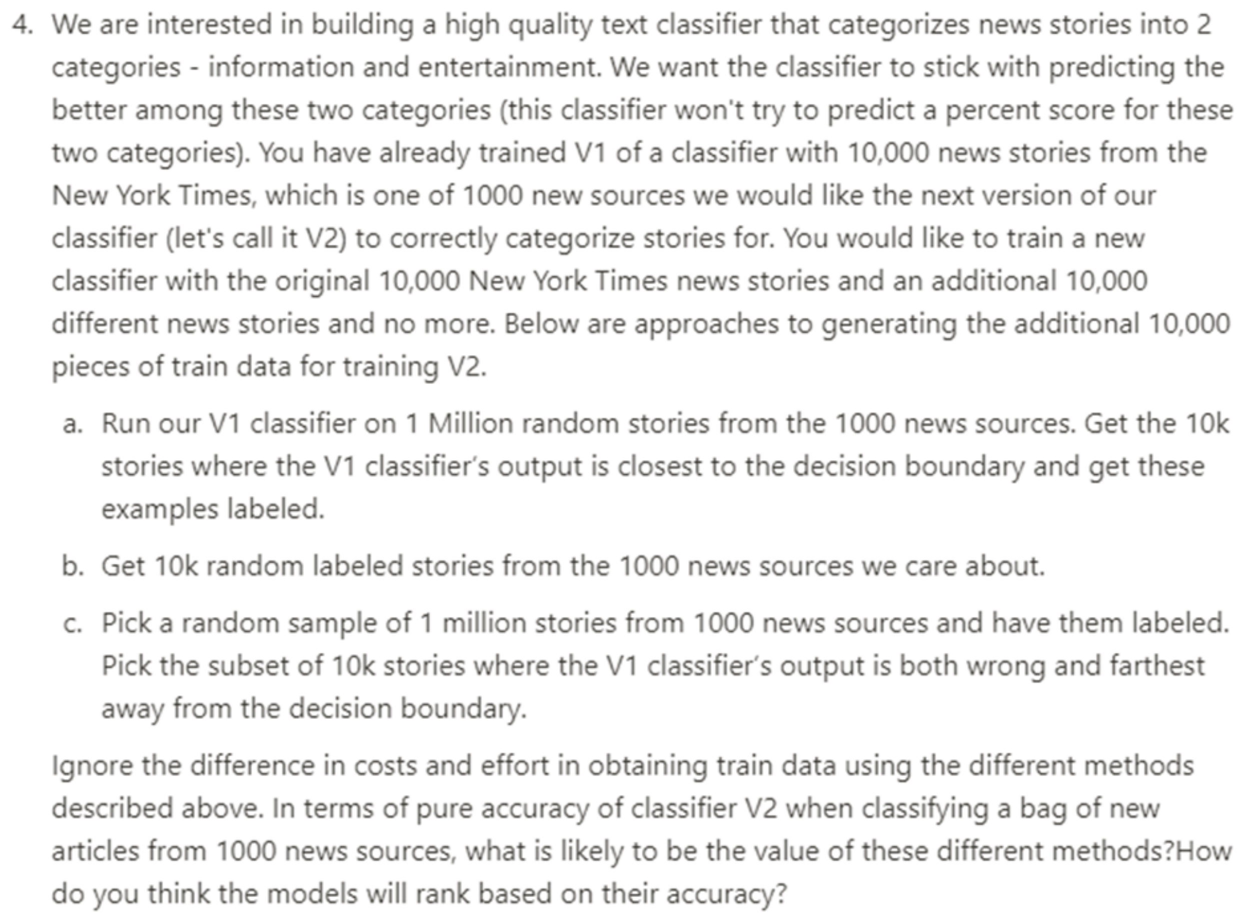
We can conclude that E is 95% more significant than A as it has received more clicks than A, assuming that the difference is statistically significant.

The same is true for B and A; if there is a difference between them, A is more clicked than B. If A is 95% more significant than A, then A is statistically significant. Since there is less of a difference now between C and D between A, we may require additional testing to determine the importance.



3. 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?

To implement gradient descent for logistic regression using a well-maintained package, we must perform two key steps: computing the gradient and updating the weights. The gradient computation involves a dot product between the feature and weight vectors. Given the sparsity of the feature vectors (with most entries being zeros), the time complexity for this step is approximately O(nk), where n is the size of the feature vectors and k is an approximation for the number of non-zero entries. The subsequent weight update, occurring once per iteration, takes O(n) time. Therefore, the overall time complexity for a single iteration is approximately O(mkn + n), where m represents the number of training examples.Since mkn is generally larger than n, the dominant term in the complexity is O(mkn).



4. 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.

   a. 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.

   b. Get 10k random labeled stories from the 1000 news sources we care about.

   c. 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?

Approach A: This approach can be beneficial for improving V2's handling of uncertain cases, but requires caution to avoid amplifying existing biases.

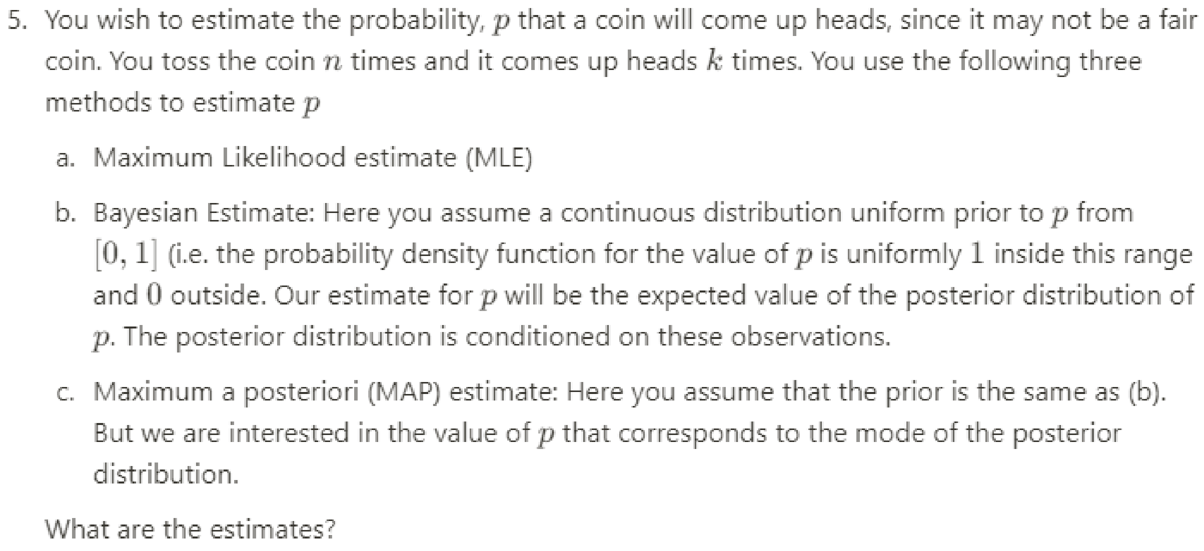
> Output/Accuracy: 10,000 labeled stories from the 1 million random stories, where V1's confidence was closest to the decision boundary. These stories would likely reflect cases where V1 struggled and could be valuable for V2 to learn .

Approach B: This approach can be effective for diversifying V2's training data and reducing overall bias, but might not address V1's specific weaknesses as effectively.

> Output/Accuracy: 10,000 randomly labeled stories from the 1000 sources. These stories would introduce diversity and guaranteed labels, but their quality might vary depending on the external labeling process.

Approach C: This approach can be highly effective for rectifying V1's weaknesses and boosting accuracy in specific areas, but requires more resources and might miss other important aspects of the data.

Output: 10,000 randomly labeled stories from the 1000 sources. These stories would introduce diversity and guaranteed labels, but their quality might vary depending on the external labeling process.



5. 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

   a. Maximum Likelihood estimate (MLE)

   b. 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.

   c. 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?

Answer

MLE = k/n

Bayesian= k+1/n+2

MAP= k+1/n+1