Assignment

Name: Rajni Poonia (2020CS10371)

Q1.

Ans: The likely relationship between the new weights would be as follows:

- Weights of Non-Duplicated Features : These weights are likely to be similar to the original weights but might not be exactly the same due to the changed dynamics in the feature space.
- Weights of Duplicated Features : The interesting part is here. Since feature n and feature n+1 are duplicates, they provide the same information to the model. The logistic regression model will try to distribute the importance (weight) it initially gave to the single feature n across both w_new(n) and w_new(n+1).

As a result: The weights w_new(n) and w_new(n+1) are likely to be roughly half of w_n or at least their sum would be close to the original w_n . This is because the contribution to the model's decision that was previously captured by w_n is now shared between two weights. However, this division is not guaranteed to be exactly equal due to the optimization process and the interaction with other features.

Q2.

Ans: To determine which statement is true regarding the comparison of email template click-through rates (CTRs) with 95% confidence, we need to conduct a statistical significance test. A common approach is using a chi-square test for comparing proportions. The null hypothesis in each case is that there is no difference between the CTRs of the templates being compared.

Let's calculate the statistical significance of the differences in CTRs between templates A (control) and each of the other templates (B, C, D, E). The data is as follows:

- Template A (Control): 10% CTR from 1000 emails, so 100 clicks.
- Template B: 7% CTR from 1000 emails, so 70 clicks.
- Template C: 8.5% CTR from 1000 emails, so 85 clicks.
- Template D: 12% CTR from 1000 emails, so 120 clicks.
- Template E: 14% CTR from 1000 emails, so 140 clicks.

We can calculate the chi-square value and corresponding p-value for each comparison to see if the differences are statistically significant at the 95% confidence level (p-value < 0.05). Let's perform these calculations. Based on the chi-square test results, the p-values for the comparisons are:

- Template A vs. Template B: $p=0.0201$
- Template A vs. Template C: $p=0.2799$
- Template A vs. Template D: $p=0.1745$
- Template A vs. Template E: $p=0.0073$

So option (2) is true.

Q3.

Ans: The approximate computational cost of each gradient descent iteration in logistic regression for sparse data is $O(m \times k)$.

This is because for sparse data, modern logistic regression implementations typically optimize computations by only considering non-zero entries. Since the average number of non-zero entries per training example is k and k is much smaller than n (the total number of features), the cost per iteration scales with the number of non-zero entries rather than the total number of features.

Q4.

Ans: To evaluate the potential impact of each approach on the accuracy of classifier V2, let's consider them individually:

1) Using 10k Stories Closest to the Decision Boundary of V1:
    a) **Value**: This method targets stories where the V1 classifier is most uncertain. Including these in the training set for V2 can help refine the decision boundary, potentially improving accuracy on ambiguous cases.
    b) **Impact on Accuracy:** Likely to be high. By focusing on borderline cases, V2 can learn to better differentiate between 'information' and 'entertainment' in more ambiguous scenarios.
2) Using 10k Random Labeled Stories from 1000 News Sources:
    a) **Value**: This approach provides a broader and potentially more representative sample of the types of stories from the 1000 news sources. It can introduce a diverse set of examples which might not be covered in the New York Times dataset.
    b) **Impact on Accuracy:** Likely to be moderate to high. The diversity of this dataset can help V2 generalize better across different sources, but it might not specifically target the weaknesses of V1.
3) Using 10k Stories Where V1 is Most Confidently Wrong:
    a) **Value**: This method focuses on correcting the most significant errors of V1. It targets cases where V1 is not only wrong but also confident in its incorrect predictions, which could indicate systematic errors in V1's approach.
    b) **Impact on Accuracy**: Likely to be high, especially in terms of improving upon the specific weaknesses of V1. By addressing its most glaring errors, V2 can significantly enhance its overall performance.

Ranking Based on Likely Accuracy Improvement:

1) Approach 3 (Correcting Most Confident Errors): Targeting the most confident errors of V1 is likely to result in significant improvements, as it directly addresses the areas where V1 is most deficient.
2) Approach 1 (Borderline Cases): This method helps refine the decision boundary, which is crucial for improving performance in ambiguous cases.

3) Approach 2 (Random Diverse Stories): While this approach aids generalization, it might not be as targeted as the other two in addressing specific weaknesses of V1.

Q5.

Ans. Let's calculate the estimates for p using the three methods:

1) Maximum Likelihood Estimate (MLE):
    a) MLE for a binomial distribution (coin tosses) is given by $p_{MLE} = k/n$.
2) Bayesian Estimate::
    a) With a uniform prior, the posterior distribution for p in a binomial scenario is a Beta distribution: Beta (k+1,n-k+1).
    b) The expected value (mean) of a Beta distribution Beta(a,b) is a/(a+b).
    c) So $p_{bayesian} = (k+1)/(n+2)$
3) Maximum a Posteriori (MAP) Estimate:
    a) The mode of a Beta distribution Beta (a,b) for a>1 and b>1 is (a-1)/(a+b-2).
    b) For k+1 > 1 and n-k+1>1, $p_{MAP} = (k/n)$
    c) If k=0 or k=n, the mode is at the boundary, and p_{MAP} would be 0 or 1 respectively.

These estimates provide different perspectives on estimating the probability p of a coin turning up heads, considering different statistical approaches.