Q/A Assignment

Harshita Badiyasar 2020MT10807

1 December 2023

1 Answer 1

If feature n is duplicated into feature $n+1$, the weights initially assigned to just feature n are distributed to n and $n+1$. Therefore, w_0 to w_{n-1} will be the same, and $\lambda w_{\text{new}_n} + (1-\lambda) w_{\text{new}_{n+1}} = w_n$ where $\lambda \in [0,1]$.

2 Answer 2

To determine the statistical significance of the click-through rates (CTR) for the different email templates, we can perform hypothesis testing. The most common approach is to use a hypothesis test for comparing two proportions (in this case, the CTRs).

Let's use the following notation:

- p_A, p_B, p_C, p_D, p_E: True CTRs for templates A, B, C, D, and E, respectively.

The null hypothesis (H_0) is that there is no difference in CTR between template A and each of the other templates:

$$\begin{aligned} H_0 &: p_A = p_B \\ H_0 &: p_A = p_C \\ H_0 &: p_A = p_D \\ H_0 &: p_A = p_E \end{aligned}$$

The alternative hypothesis (H_1) is that there is a difference:

$$\begin{aligned} &H_1 : p_A \neq p_B \text{ (two-tailed test)} \\ &H_1 : p_A \neq p_C \\ &H_1 : p_A \neq p_D \\ &H_1 : p_A \neq p_E \end{aligned}$$

To determine which statements are true based on the confidence level, we can perform a hypothesis test for each template and calculate the confidence interval for the difference in proportions.

1. "We have too little data to conclude that A is better or worse than any other template with 95% confidence." - This is not necessarily true. We can perform hypothesis tests with the given data.

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." - This statement is not accurate. Without performing the tests, we cannot conclude the direction of the differences or make statements about confidence levels.

3. "Both D and E are better than A with 95% confidence. Both B and C are worse than A with over 95% confidence." - This statement is too specific and makes assumptions about the direction of the differences without performing the tests.

In conclusion, based on the information provided, we cannot make specific statements about the direction of differences or confidence levels without performing hypothesis tests. To draw conclusions, conduct hypothesis tests for each template comparison and calculate confidence intervals for the differences in proportions.

3 Answer3

In a sparse setting, where the average number of non-zero entries in each training example is k and k ¡¡ n, the key advantage comes from the fact that the gradient computation can be efficiently implemented to consider only the non-zero entries.

The computational cost of each gradient descent iteration for logistic regression in a sparse setting is often dominated by the multiplication of the feature matrix by the error vector.

The time complexity of sparse matrix-vector multiplication (SpMV) is approximately O(k * m), where k is the average number of non-zero entries per row and m is the number of training examples. This is significantly more efficient than O(n * m), which would be the complexity for dense matrices.

So, in summary, the approximate computational cost of each gradient descent iteration for logistic regression in a sparse setting is O(k * m), where k is the average sparsity level of the feature vectors and m is the number of training examples.

4 Answer 4

Using V1 classifier on 1 Million random stories:

Pros:

It targets examples near the decision boundary, potentially focusing on challenging cases that the initial model found difficult. Can capture nuances in the data that might be ambiguous or on the edge of the two categories.

May improve accuracy on cases that are challenging for the initial model.

Could adapt to new patterns or topics in the news.

Cons:

The approach relies on the assumption that the V1 classifier is a good starting point and can effectively identify cases near the decision boundary. There is a risk of overfitting to the specific characteristics of the initial 10,000 New York Times stories. The added complexity might not significantly improve accuracy if the initial V1 model is already performing well. Getting 10k random labeled stories:

Pros:

Provides a diverse set of labeled examples that could cover a wide range of topics and writing styles. Reduces the risk of biasing the model towards specific patterns present in the initial 10,000 New York Times stories. Cons:

May include many examples that are relatively easy for the model, potentially not pushing the model to generalize better. May not specifically target challenging cases or those near the decision boundary. Could introduce noise or irrelevant patterns if the random sample is not representative of the broader data distribution. Random sample with V1 classifier on 1 Million stories:

Pros:

Targets cases where the V1 classifier is both wrong and far from the decision boundary, potentially addressing cases where the model is confident but incorrect. Focuses on challenging cases where the initial model made mistakes, encouraging better generalization. Cons:

Similar to the first approach, it assumes that the V1 classifier is a good starting point and can effectively identify cases where it is wrong and far from the decision boundary. May introduce biases based on the errors of the initial model. Potential Ranking based on Accuracy:

Approach 1 could potentially lead to improved accuracy by focusing on challenging cases and the decision boundary. Approach 3 might also be effective if the errors made by the initial model are particularly informative and can be corrected. Approach 2 may provide a good baseline but might not push the model to its limits in terms of handling challenging cases.

5 Answer 5

Let's denote the probability of getting heads as p. When you toss the coin n times and it comes up heads k times, the estimates for p using the three methods are as follows:

1. **Maximum Likelihood Estimate (MLE):** The MLE for p is simply the ratio of the number of heads observed (k) to the total number of coin tosses (n):

$$\text{MLE: } \hat{p}_{\text{MLE}} = \frac{k}{n}$$

2. **Bayesian Estimate:** With a uniform prior, the posterior distribution is a Beta distribution, and the expected value (mean) of the posterior distribution

is used as the Bayesian estimate for p. The Beta distribution parameters are updated based on the number of heads (k) and tails ($n-k$) observed:

$$\text{Bayesian Estimate: } \hat{p}_{\text{Bayesian}} = \frac{k+1}{n+2}$$

3. **Maximum a Posteriori (MAP) Estimate:** The MAP estimate corresponds to the mode of the posterior distribution, which, in the case of a Beta distribution, is given by:

$$\text{MAP Estimate: } \hat{p}_{\text{MAP}} = \frac{k}{n}$$

It's important to note that in the case of a uniform prior Beta distribution, the MAP estimate coincides with the MLE estimate.

$$\text{MLE: } \hat{p}_{\text{MLE}} = \frac{k}{n}$$
$$\text{Bayesian Estimate: } \hat{p}_{\text{Bayesian}} = \frac{k+1}{n+2}$$
$$\text{MAP Estimate: } \hat{p}_{\text{MAP}} = \frac{k}{n}$$

These estimates represent different approaches to infer the probability p based on the observed data and prior beliefs in the case of Bayesian estimation. The choice between these methods may depend on the specific assumptions and goals of your analysis.