SuperAGI | Assignment(For AI roles)

Ans-1:

In ordinary linear regression, the model has the flexibility to distribute weights between different columns as it pleases without altering the objective function. However, this flexibility may lead to issues, especially when trying to infer regression weights. If you use a matrix inversion method for fitting, it may fail, or if a general-purpose optimization algorithm is employed, the breakdown of weights between columns can be arbitrary.

For predictive modeling purposes, where the focus is on learning a predictive model rather than inferring specific weights, this arbitrary breakdown may not be a critical concern, unless there are numerical stability issues with matrix inversion.

On the other hand, when employing regularized linear models like Ridge Regression or LASSO, the weight distribution is influenced by the chosen regularization method. This introduces a level of regularization control and helps prevent overfitting.

In the context of logistic regression, it's important to note that due to the nature of the logistic function, the weights for different features (columns) are learned in a way that contributes to the probability prediction. The regularization term, commonly added to the objective function, helps control the magnitude of the weights.

In decision tree models, the splitting of nodes is based on features, and it may seem arbitrary in terms of feature selection. However, for feature importance analysis, this arbitrary choice may impact the interpretation.

Hence, in logistic regression, the weights $w_0, w_1, \ldots, w_{n-1}$ are equivalent to the corresponding weights $w_{\text{new}0}, w_{\text{new}1}, w_{\text{new}2}, \ldots, w_{\text{new}n-1}$. Additionally, the regularization constraint implies that $w_{\text{new}n} + w_{\text{new}n+1} = w_n$.

Ans-2:

1. **Definitions:**

   Given, *A*,*B*,*C*,*D*,*E*: Control template and alternative templates.

   n: Number of emails sent for each template (1000 in this case).

   CTR_A,CTR_B,$CTR_C$$CTR_D$,$CTR_E$ - Click-through rates for templates A, B, C, D, E, respectively.

2. **Hypotheses:**

   H_0: There is no significant difference between the click-through rates of the templates (Null Hypothesis).

   H_1: There is a significant difference between the click-through rates of the templates (Alternative Hypothesis).

3. **Mathematical Formulation:**

- **Mean Click-Through Rates:** μ_A, μ_B, μ_C, μ_D, μ_E

- **Standard Deviations:** σ_A, σ_B, σ_C, σ_D, σ_E
- **Z-Scores:**

  $Z_A = \frac{CTR_A - \mu_A}{\sigma_A/\sqrt{n}}$

  $Z_B = \frac{CTR_B - \mu_B}{\sigma_B/\sqrt{n}}$

  $Z_C = \frac{CTR_C - \mu_C}{\sigma_C/\sqrt{n}}$

  $Z_D = \frac{CTR_D - \mu_D}{\sigma_D/\sqrt{n}}$

  $Z_E = \frac{CTR_E - \mu_E}{\sigma_E/\sqrt{n}}$

5. **Z-Score Threshold for 95% Confidence:**

   $Z_{threshold}$=1.96

6. **Decision Rule:**

   If $|Z_i| > Z_{\text{threshold}}$, reject H_0 in favor of H_1.

6. **Results:**

   Template E is determined to be significantly better than A if $|Z_E| > Z_{\text{threshold}}$.

   Template B is determined to be significantly worse than A if $|Z_B| > Z_{\text{threshold}}$.

   Templates C and D require additional data or testing to establish significance.

7. **Conclusion:**

   We can be 95% confident that template E is better than template A, and that template B is worse than template A. However, we need to run the test for longer to compare templates C and D to template A with 95% confidence.

   As you can see, the z-scores for templates B, C, and D are all less than 1.961.96, which means that we cannot be 95% confident that they are different from template A. However, the z-score for template E is greater than 1.96, which means that we can be 95%95% confident that it is better than template A.

Final Answer

The correct option is :- 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.

Ans-3 :

The time complexity of gradient descent is expressed as (O(knd)), where:

- (k) is the number of iterations,
- (n) is the number of samples, and
- (d) is the number of features or parameters being updated during the iterative optimization process of gradient descent.

In scenarios where the number of features ((d)) becomes substantial, for instance, when (d > 10^4), the computation of the matrix (\theta = (X^{TX)}{-1}X^Ty) (known as the Normal Equation) in linear regression, particularly the ((X^{TX)}{-1}) part, becomes computationally challenging.

In such cases, it becomes more practical to employ the gradient descent method to find the optimal parameters for linear regression. The computational cost of gradient descent is determined by the number of iterations needed to reach convergence, and its complexity is (O(kn^2)).

This distinction becomes crucial when (n) (the number of samples) is very large, making it more efficient to use gradient descent instead of the closed form of linear regression due to its lower computational cost and scalability.

Ans-4 : Let's analyze the three approaches to generating additional training data for V2 and discuss their potential impact on the accuracy of the new classifier:

1. **Run V1 Classifier on 1 Million Random Stories:**
    - **Methodology:** Utilize the V1 classifier to predict categories for 1 million random stories. Select the 10,000 stories where the V1 classifier's output is closest to the decision boundary.
    - **Potential Impact:** This approach focuses on the ambiguous cases where V1 is uncertain about the category. It may capture examples that are challenging for the current classifier. However, relying solely on the decision boundary may not guarantee diversity in the dataset.
2. **Get 10,000 Random Labeled Stories:**
    - **Methodology:** Randomly select 10,000 labeled stories from the 1000 news sources.
    - **Potential Impact:** This method introduces diversity in the training dataset. It includes various examples, not necessarily focusing on ambiguous cases. It provides a broad overview of the data, potentially capturing a wide range of language use and topics.
3. **Pick Random Sample of 1 Million Stories and Label Ambiguous Cases:**
    - **Methodology:** Randomly select 1 million stories and label them. Pick the subset of 10,000 stories where the V1 classifier's output is both wrong and farthest away from the decision boundary.
    - **Potential Impact:** This approach combines randomness with a focus on cases where V1 is both wrong and far from the decision boundary. It aims to include challenging examples while ensuring diversity in the dataset.

Ranking Based on Accuracy:

- The second approach, obtaining 10,000 random labeled stories, is likely to contribute to the overall accuracy by providing a diverse set of examples. This approach does not rely on the decision boundary and captures various linguistic nuances and topics.
- The third approach, picking a random sample and labeling ambiguous cases, also has the potential to improve accuracy by addressing challenging instances. However, the success of this approach depends on the effectiveness of identifying cases where V1 is both wrong and far from the decision boundary.
- The first approach, focusing on examples close to the decision boundary, may be helpful in refining the classifier's performance on ambiguous cases. However, it may not provide sufficient diversity, and the improvement in accuracy might be more incremental compared to the other methods.

In summary, the second approach, obtaining random labeled stories, is likely to have the most significant impact on improving the accuracy of the V2 classifier. The third approach, combining randomness with a focus on challenging cases, also has potential, while the first approach may offer more incremental improvements.

ANS-5 : Coin Toss Probability Estimates

1. **Maximum Likelihood Estimate (MLE):**
    - MLE is given by the ratio of the number of heads to the total number of tosses.
      $[\text{MLE: } \hat{p}_{\text{MLE}} = \frac{k}{n}]$
2. **Bayesian Estimate:**
    - Assuming a uniform prior, the posterior distribution is a Beta distribution with parameters $\alpha = k + 1$ and $\beta = n - k + 1$. The expected value of this distribution is the Bayesian estimate.
      Bayesian Estimate: $\hat{p}_{\text{Bayesian}} = \frac{\alpha}{\alpha+\beta} = \frac{k+1}{n+2}$
3. **Maximum a Posteriori (MAP) Estimate:**
    - The mode of the posterior distribution is given by the peak of the Beta distribution, which occurs at (\frac{k}{n}).
      $MAP Estimate : \hat{p}_{\text{MAP}} = \frac{k}{n}$

In summary:

- MLE is simply the observed probability of getting heads.
- Bayesian Estimate incorporates a prior, resulting in a smoother estimate that tends towards the observed probability but is influenced by the prior.
- MAP Estimate, assuming the mode of the posterior, is the same as the MLE in this case due to the choice of the uniform prior.