Assignment

Sugata Ghorai

17 December 2023

1 Section QA:

Answer 1: In the scenario where a feature n is duplicated to create a new feature $n+1$, and a logistic regression model is retrained, the weights $W_{\text{new},n}$ and $W_{\text{new},n+1}$ would likely adjust such that their sum is close to the original weight W_n. This is because the logistic regression model would distribute the total importance that was previously assigned to feature n across both the original and the duplicated feature.

Answer 2:

a. Yes , this statement is correct . From the figure we get that B and C are worse than A and D and E are better than A . But without proper statistical test we cannot identify the position of A with respect to others.

b. The statement is correct about condition on A and E. But it is wrong about A and B . We need a sophisticated statistical testing about A and B . Also D and E need much more number of tests for coming to conclusion which statistically valid.

c. This statement claims that both D and E are better than A with 95% confidence and that both B and C are worse than A with over 95% confidence. Without statistical testing, we cannot make these claims with a specified level of confidence. Based on the CTRs alone, Template E appears to be the best performer, and Template B the worst. However, the assertion of 95% confidence for any comparison requires statistical testing. The most accurate option, given typical statistical standards and without seeing the actual p-values or confidence intervals from a statistical test, would likely be a version of option b.

Answer 3: The computational cost for a single iteration of gradient descent in logistic regression with sparse feature vectors can be calculated by considering the operations involved in computing the gradient of the loss function with respect to the weights. Given:

1. m training examples
2. n features
3. k average non-zero entries in each feature vector (with $k \ll n$)

Here are the steps for a single iteration of gradient descent, focusing on the computational cost:

1. Compute the Prediction:
   - For each training example, we compute the predicted probability. This involves a dot product between the feature vector and the weight vector. For sparse vectors, this dot product only needs to be taken over non-zero entries.
   - Cost per example: $O(k)$
   - Total cost for all examples: $O(m \cdot k)$
2. Compute the Gradient:
   - The gradient of the loss function with respect to the weights is the average over all training examples of the product of the prediction error and the feature vector. Again, for sparse feature vectors, we only consider non-zero entries.
   - Cost per example: $O(k)$
   - Total cost for all examples: $O(m \cdot k)$
3. Update the Weights:
   - The weights are updated by subtracting the product of the learning rate and the gradient. The update is only applied to weights corresponding to non-zero features.
   - Cost: $O(k)$

Answer 5:

a. Maximum Likelihood Estimate : $\frac{k}{n}$.

b. Bayesian Estimate : $\frac{k+1}{n+2}$.

c. MAP estimate: $\frac{k-1}{n-2}$