Evaluation of Early Stopping in A/B Testing

Shan Huang

May 22, 2017

This documentation shows the evaluation of early stopping algorithms.

1 Problem

Given samples **x** from treatment group, samples **y** from control group, We are interested in whether there is a significant difference between the mean the two variants $\delta = \mu(y) - \mu(x)$. To save the cost of long-running experiments, we want to stop early if we are already certain that there is a significant result.

2 Significance Decision

Given H_0 represents the null hypothesis of no difference, H_1 represents the alternative hypothesis meaning there is a difference, we can decide whether the result is statistically significant using either:

- **Confidence interval**: If 0 is outside confidence interval of δ, it is statistically significant. Vice versa.
- or **credible interval**: Credible interval is the Bayesian version of confidence interval. If 0 is outside credible interval of δ, it is statistically significant. Vice versa.
- or **Bayes factor**: Theoretically, Bayes factors higher than 3 can be interpreted as support for the null hypothesis (significant no difference), whereas values smaller than 1/3 can be interpreted as support for the alternative hypothesis (significant difference). Values between 1/3 and 3 are inconclusive. **In our test, we will only use Bayes factor less than 1/3 as decision criteria**. See reasons below.

The ability of each metric, i.e., types of significance it can detect, is shown in the following table.

	Paradigm	Significant H_1	Significant H_0	No Significant Result
Confidence Interval	frequentist	✓		✓
Credible Interval	Bayesian	✓		✓
Bayes factor	Bayesian	✓	✓	✓

Table 2.1: Comparison of Decision Ability

To be consistent with all mehods, we draw a binary conclusion that either there is a significant difference or not. In other words, the first column means that there is a significant difference, combining the second and third column means that there is no significant difference. Thus, we only use the part where Bayes factor less than 1/3 as decision criteria.
It is worth noting that a **typical conclusion of Bayes factor** would be "*There is a significant difference corresponds to Cauchy prior and a threshold of Bayes factor=3*". This might be quite difficult to explain to non-tech users. While the **conclusion based on interval** can be more intuitive such as "*You can be 95% sure that the significant difference is not due to chance*".

3 Early Stopping Criteria

We can stop the experiment by either

- **Confidence interval**: Calculate the new significance level for each day based on alpha-spending function in group sequential method. Stop the test If 0 is outside confidence interval of δ.
- or **credible interval**: Stop the test if 0 is outside credible interval of δ.
- or **Bayes factor**: Stop the test if Bayes factor is smaller than 1/3 (significant difference).
- or **Bayes precision**: Stop the test if credible interval width is smaller than 0.08.

4 Evaluation

4.1 Evaluate Significance Decision

We are going to compute FP(Type I error), FN(Type II error), TP, TN rates for the three significance decision algorithms described in Section 2.

4.2 Evaluate Early Stopping Criteria

It is obvious that if we use the frequentist approach in early stopping, we should also use frequentist's approach (that is confidence interval) to decide significance. Moreover, when using Bayes factor to stop, we often find a conflicting result if we draw significance conclusion based on credible interval. Therefore we don't evaluate all combinations of early stopping algorithms and significance decision algorithms, find below a table of combination we will evaluate.

Significant Based On	Early Stopping Based On			
	Confidence Interval	Credible Interval	Bayes factor	Bayes precision
Confidence Interval	✓			
Credible Interval		✓		
Bayes factor			✓	✓

Table 4.1: Combinations of Early Stopping and Significance Decision To Evaluate

For each combination shown in the table above, we will evaluate the following metrics:

- **False positive rate**: Percentage of wrongly stopped tests. i.e., Early stop says there is a significant difference and stops the test, but the test to the end day (as if there is no early stopping) will tell you it is not significant.
- **Run time reduced(all)**: Percentage of run time reduced for all tests.
- **Run time reduced(true positive)**: Percentage of run time reduced for only the correctly stopped tests.
- **Bias**: Percentage of the difference of effect size(delta) between stopping day and end day of test(as if there is no early stopping).

5 Evaluation

The evaluation is conducted on both simulation data and real data. Code can be found here.

5.1 Simulation Data

We generate 2000 simulation tests based on Gaussian distributed KPIs. We calculate the minimal detectable effect size from power analysis. 1000 tests have a significant difference between control and treatment, and the other 1000 tests have no significant difference.
We model the frequency of visits of an entity by a Poisson distribution with 3 days per entity in average. We simulate the A/B testing for a period of 20 days, and decide whether to stop on each day.
First of all, let's see the performance of different **significance decision** algorithms.

	FPR	TNR	FNR	TPR
confidence interval	4.6%	95.4%	0.9%	99.1%
credible interval	4.6%	95.4%	1.0%	99.0%
Bayes factor	0.2%	99.8%	18.9%	81.1%

Table 5.1: Performance of Significance Decision on Simulation Data

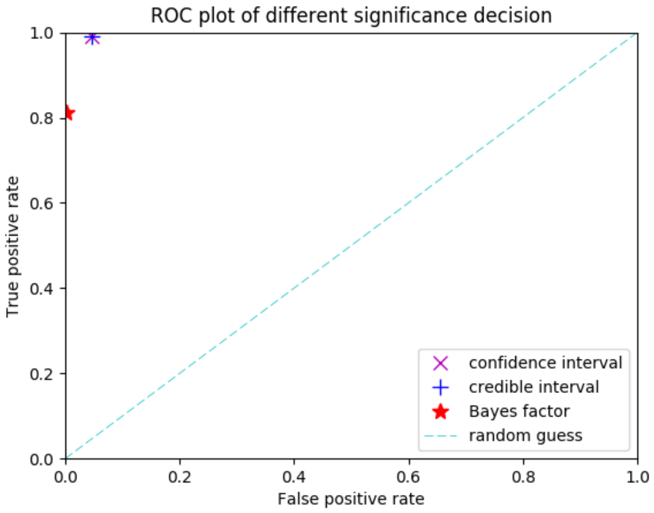

Figure 5.1: ROC plot on simulation data

We can observe in Figure 5.1 that Bayes factor works very good on false positive rate, but the two interval based approaches works better on true positive rate. In general, the error rate of all three methods are acceptable(FPR < 5% and FNR < 80%) for simulation data.

Next, let's look at the results of **early stopping** algorithms. To make notations compact, in the following table we use FPR for false positive rate, RTR(all) for run time reduced(all), RTR(TP) for run time reduced(true positive) and CI for confidence interval. The definition of each metric is described previously in Section 4.2.

Significance	**Early Stopping**	**Paradigm**	**FPR**	**RTR(all)**	**RTR(TP)**	**Bias**
CI	CI	frequentist				
Credible Interval	Credible Interval	Bayesian				
Bayes factor	Bayes factor	Bayesian				
Bayes factor	Bayes precision	Bayesian				

Table 5.2: Performance of Early Stopping on Simulation Data

5.2 Real Data

Repeat the evaluation process on real data from previous A/B testings.

6 Conclusion

Choose a significance decision and an early stopping algorithm.