COMPUTER SCIENCE TRIPOS Part IA 75%, Part IB 50% – 2019 – Paper 3

7 Machine Learning and Real-world Data (sht25)

You want to compare the performance of two classification systems and perform a significance test on their results. You use six items as detailed in the table below, where the correct answer ("Gold Standard") and the answers of Systems 1 and 2 are listed.

System 1	System 2	Gold Standard
N	0	N
N	P	P
P	0	0
P	N	P
P	0	P
0	N	0

Accuracy as an evaluation metric was introduced in Lecture 1.

(*a*) Name the standard evaluation metric for classification, give its formula, and calculate its value for the two systems' results. [2 marks]

Answer: Accuracy $A = \frac{\text{\# of correctly classified items}}{\text{\# of items}}$. $A_{system1} = \frac{4}{6}.A_{system2} = \frac{2}{6}$

The sign test was introduced in Lecture 4, and students were encouraged to use statistical testing in the entirety of the course.

(*b*) Apply the sign test at significance level $\alpha = 0.05$ to test whether System 1 is significantly better than System 2. [5 marks]

Answer: The Null Hypothesis is that the results of Systems 1 and 2 are created from the same distribution, i.e., that System 1 is not better than System 2.

System 1	System 2	
1	0	+
0	1	-
0	1	-
1	0	+
1	0	+
1	0	+

We calculate the probability p that as many "-" results as observed for System 2, or fewer, would be observed under the Null hypothesis, and we will reject the Null hypothesis if $p <= \alpha$. From the binomial distribution, we can calculate this as:

$$p = P_q(X \leq k|N) = \sum_{i=0}^{k} \binom{N}{i} q^i (1-q)^{N-i}$$

In our case, with $q = 0.5$, p equates to

$p = (60) \times 0.5^0 \times 0.5^6 + (61) \times 0.5^1 \times 0.5^5 + (62) \times 0.5^2 \times 0.5^4 = 0 + (1+6+15) \times 0.5^6 = 0.343750$

This means that we cannot reject the Null hypothesis because $p > \alpha$. According to this test, the results we observe may well be due to chance, given the low sample size.

(*c*) If instead we are testing whether Systems 1 and 2 are statistically different, how does that change your calculations in Part (*b*)? [1 mark]

Answer: We have so far performed a one-tailed significance test, assuming that it is a given that System 1 performs better. In reality, without having looked at the means beforehand, we would not know that. It could actually happen that System 2 was better. We should therefore always perform a two-tailed significance test, as it makes no *a priori* assumptions at all. As the binomial distribution is symmetric, we simply multiply the p we calculated earlier by 2. This means that we are now even more certain that we cannot reject the Null Hypothesis.

(d) A saboteur appears in your laboratory, and creates fake versions of the results table above. She does this by swapping the values of System 1 and 2 in the same row. She decides randomly for each version how many different rows she subjects to this treatment.

(i) How many different fake versions of the table can be generated? [2 marks]

Answer: 2^6, as for each row she can either swap or not (2), and she takes this binary decision 6 times, once for each row.

The students have not come across the test that is introduced here, so understanding what is going on requires some transfer thinking about statistical tests.

(ii) Somebody suggests that the saboteur's actions can be used as the foundation of a new statistical test, based on the idea that if the Null Hypothesis were true, this would imply that results can be randomly swapped without overall changes in the result. Explain how you can use this idea for significance testing. Illustrate how you apply the new test using the table above. [6 marks]

Answer: If the Null hypothesis holds and the results of Systems 1 and 2 are indeed interchangable, we would expect a large number of the copies would perform roughly similar to the original result, some showing a difference in performance in one direction, some in the other. If we find however that the difference between the systems decreases by this manipulation, we know that the systems were different (and swapping made them more similar). In fact, the likelihood that swapping made the samples more similar is directly related to the number of samples where the difference decreased. We therefore take the number of swaps where the difference *increases* as our p:

$$p = \frac{\text{\# copies where } (A_{system1'} - A_{system2'}) > (A_{system1} - A_{system2})}{\text{\# all copies}}$$

(e) The table does not contain any ties, but a high number of ties are often a reality in experiments.

(i) How does the presence of many ties affect the sign test? [2 marks]

Answer: This set of questions is somewhat open-ended. We have in lectures touched upon the fact that when ties can become frequent in the kinds of experiments we have been performing in MLRD. Ties might be very rare when real-valued observations such as length of leaves are taken, but when systems are run against each other on binary success criteria as is often the case in NLP, they can become frequent. In this case, we cannot simply follow the advice to "omit ties". A solution was introduced for the sign test in Lecture 4: count half of ties as belonging to +, half to -. Round up. I expect a short discussion why this is reasonable, or a counter argument.

(ii) How does it affect your newly developed test from Part *(d)*(*ii*) above?

[2 marks]

Answer: As for the permutation test, ties are the rows which when swapped change nothing. Even if nothing is done about them, they naturally balance the amount of (the more, the higher the sensitivity of the test) against ties (the more ties, the lower the sensitivity of the test). This test is overall more sensitive than the sign test, and much less affected by ties than the sign test, another of its advantages. This positive effect comes from the fact that this test more directly works with the null hypothesis through exchanges between Systems 1 and 2.