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

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

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

(*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]

(*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]

(*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]

(*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]

(*ii*) How does it affect your newly developed test from Part (*d*)(*ii*) above? [2 marks]