COMPUTER SCIENCE TRIPOS Part IB – mock – Paper 6

3 Data Science (DJW)

A researcher has a dataset of $n = 500$ records, each record a pair (x_i, y_i) where x_i is the predictor variable and $y_i \in \mathbb{R}$ is the response variable. They trained two supervised learning algorithms, A and B, to try to predict y_i given x_i. Algorithm A's prediction was closer in $n_A = 260$ cases, and B's prediction was closer in $n_B = 240$ cases. They wish to know if there truly is any difference between the two, or if the result can be attributed to chance.

You advise the researcher that n_A should be modelled as a $\text{Bin}(n, \theta)$ random variable, with unknown parameter θ.

(a) Let $\hat{\theta}$ be the maximum likelihood estimator for θ. Give an expression for $\hat{\theta}$. [1 mark]

(b) Explain what is meant by "a 95% confidence interval for $\hat{\theta}$". Give pseudocode to compute it. In your answer, you should explain whether a one-sided or a two-sided interval is more appropriate. [7 marks]

(c) Explain how to conduct a hypothesis test of the hypothesis "$\theta = 1/2$". Give pseudocode. In your answer, you should define p-value, and explain whether a one-sided or a two-sided test is more appropriate. [8 marks]

(d) A second researcher asks you if A is better than B. Do you the same advice to this researcher as you gave in parts (b) and (c)? Explain why or why not. [4 marks]

[*Bonus material: If the response variable we want to predict is discrete, then there are three possible outcomes for each case. Suppose A's prediction was better in* $n_A = 70$ *cases, B's prediction was better in* $n_B = 50$ *cases, and they were both equally good or bad in* $n_0 = 380$ *cases. We can model* (n_A, n_B, n_0) *as drawn from a multinomial random variable with parameters* $(\theta_A, \theta_B, 1 - \theta_A - \theta_B)$*, and test the hypothesis "*$\theta_A = \theta_B$*". A good choice of test statistic is* $n_A/(n_A + n_B)$*. This test is discussed in Part IA/IB Machine Learning and Real World Data, though with the less powerful test statistic* $n_A + n_0/2$*.*]