Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 7th, 2023

Problem 2: Doping in professional cyclist races

The Union Cycliste Internationale (UCI) routinely conducts blood-based tests during professional cycling events to detect potential doping cases. However, this year, the UCI aims to introduce a faster and more cost-effective preliminary urine-based test. This urine test will serve as an initial screening to identify individuals with abnormal urine characteristics who are likely to be doping. Subsequently, those flagged by the urine test will undergo the more comprehensive and accurate blood-based testing.

The file `doping.txt` contains measurements of four urine parameters for 50 cyclists: `pH` the potential of hydrogen, `creatinine` the creatinine concentration in mg/dL, `rdensity` the relative density with respect to water density and `turbidity` which measures urine opacity on a scale from 0 (transparent) to 1 (opaque). Additionally, the variable `result` indicates whether these athletes were found to be doping or not according to the blood-based testing.

The aim is to build a classifier for discriminating doping cases vs clean cases, based on the four urine parameters listed above.

The cost associated with the blood-based test is 1000€ per cyclist. This cost is equivalent to the estimated economic benefit of correctly identifying a true doping case. The economic loss incurred due to failing to detect a doping case is quantified as a cost of 50,000€. The UCI estimates that 1% of professional cyclists are doping. The classifier should be optimal in minimizing the expected cost of misclassification.

a) Which classification method should be used? Verify the underlying assumptions.

b) Build the corresponding classifier, providing an estimate of its actual error rate through leave-one-out cross-validation.

The UCI intends to use the classifier developed in (b) to identify potential doping cases during the upcoming professional race, which will have 200 cyclists.

c) How much should be budgeted for the cost of the advanced blood-based tests?

d) Compared to the previous strategy of conducting accurate blood-based tests on all cyclists, what are the cost savings associated with the new two-fold testing approach?

Upload your results here:
`https://forms.office.com/e/QMWKwRqxr9`