Infinitely imbalanced logistic regression

Art B. Owen
Stanford University

In binary classification problems it is common for the data sets to be very imbalanced: one class is very rare compared to the other. In this work we consider the infinitely imbalanced case where the rare class has fixed finite sample size n, while the common class has sample size N —> сю. For logistic regression, the infinitely imbalanced case often has a useful solution. The logistic regression intercept typically diverges to —сю as expected. But under mild conditions, the rest of the coefficient vector approaches a non trivial, interpretable and useful limit. Perhaps surprisingly, the limiting parameter vector depends on the n points from the rare class only through their sample mean.