CSDS 440: Assignment 8

Shaochen (Henry) ZHONG, `sxz517`

Mingyang TIE, `mxt497`

Due and submitted on 10/30/2020
Fall 2020, Dr. Ray

Problem 35

We denotes $F = e_A - e_B$ to represent the difference between the error rate of the two classifers. We know that $E(F) = e_1 - e_B$ and the variance will be $V(F) = \frac{e_A(1-e_A)+e_B(1-e_B)}{n}$ according to the property of binomial distribution.

As we may model our calculation as a Gaussian distribution, a 95% CI is around 1.96σ. We want to make 0 will not in this interval, thus $0.1 - 1.96\sigma > 0 \Longrightarrow 0.1 > 1.96\sigma$. Now by substitute the $V(F)$ in, we have:

$$
\begin{aligned}
0.1 &> 1.96\frac{e_A(1-e_A)+e_B(1-e_B)}{n} \\
&> 1.96\frac{\sqrt{e_A(1-e_A)+e_B(1-e_B)}}{\sqrt{n}} \\
\sqrt{n} &> \frac{1.96}{0.1}\sqrt{e_A(1-e_A)+e_B(1-e_B)} \\
\Longrightarrow n &> 384.16(e_A(1-e_A)+e_B(1-e_B))
\end{aligned}
$$

Known that the 95% CI are $[x_B, y_B]$ and $[x_N, y_N]$ respectively for two datasets, we may deduce the error rates to be $e_B = \frac{x_B+y_B}{2}$ and $e_N = \frac{x_N+y_N}{2}$ for their respective dataset.

Assume Prof. Bob's dataset has n_B examples and Prof. Nan has n_N examples, Prof. Scoop may simply derive his version of error rate to be a combination of the both, which is $e_S = \frac{n_B e_B + n_N e_N}{n_B + n_N}$. This implies $n_S = n_B + n_N$, we may get this n_S (and also n_B, n_N) by doing:

$$\begin{aligned}
x_B &= e_B - 2\sigma_B \\
2\sqrt{\frac{e_B(1-e_B)}{n_B}} &= e_B - x_B \\
\frac{4e_B(1-e_B)}{n_B} &= (e_B - x_B)^2 \\
\implies n_B &= \frac{4e_B(1-e_B)}{(e_B - x_B)^2}
\end{aligned}$$

By doing the same calculation for n_N we have:

$$\begin{aligned}
n_N &= \frac{4e_N(1-e_N)}{(e_N - x_N)^2} \\
\implies n_S &= \frac{4e_B(1-e_B)}{(e_B - x_B)^2} + \frac{4e_N(1-e_N)}{(e_N - x_N)^2}
\end{aligned}$$

Since e_B, n_B, e_N, n_N, n_S are now all known, we can calculate e_S accordingly. Now, assume Prof. Scoop will land with a 95% CI of $[x_S, y_S]$, then there must be:

$$\begin{aligned}
x_S &= e_S - 2\sigma_S \\
&= e_S - 2\sqrt{\frac{e_S(1-e_S)}{n_S}} \\
y_S &= e_S + 2\sigma_S \\
&= e_S + 2\sqrt{\frac{e_S(1-e_S)}{n_S}}
\end{aligned}$$

The above two equations suggest we may derive x_S and y_S with just e_S and n_S, which are two known value to us and also to Prof. Scoop. So Prof. Scoop – while not doing any experiment – may derive his 95% CI base on Prof. Bob and Prof. Nan's results at the cost of his academic integrity.