Statistical Learning Theory, Exercise 4

Michael Hirsch

January 15, 2015

Bounds on Membership Uncertainty

A sample size of $2t = 2 \times 10^4$ is drawn from some distribution, and this sample is then randomly split up into two half-samples of size $t = 10^4$

1. For any specific event A, these two half-samples define two frequencies, $f_1(A)$ and $f_2(A)$. Find an explicit upper bound on the probability that $|f_1(A) - f_2(A)| > 0.1$.

   In this scenario, we have two samples both of size $t = 10^4$. Further, we know that $E[s_1] = E[s_2] = S/T \times 10^4$ in both samples, where S and T are unknown. Let us assume that S and T are fixed, we have the equality:

$$\begin{aligned} Pr\{|f_1(A) - f_2(A)| > 0.1\} &= Pr\{|f_1(A) - E[f_1(A)]| > 0.05\} \\ &= Pr\{|s_1 - St/T| > 0.05\} \\ &= Pr\{|s_1/t - S/T| > 0.05/t\} \\ &= Pr\{|s_1/t - E[s_1/t]| > 0.05/t\} \\ &< 2e^{-2(0.05)^2 10^4} = 2e^{-50} \approx 1.93 \times 10^{-22} \end{aligned}$$

   That is, there is a very small probability that the frequencies will differ by more than 10% of the mean $\square$

2. We now make such a comparison for each $\Phi(3, 2 \times 10^4)$ different sets. Find an explicit upper bound on the probability that $|f_1(A) - f_2(A)| > \varepsilon$ for at least one A.

   The union bound, also known as Boole's Inequality, is formulated as follows, where $i \in [1, \Phi(3, 2\times 10^4)]$: For any countable number of countable events $A_1, A_2, A_3, \ldots$ we have:

$$\begin{aligned} P(\bigcup_i A_i)\} \leq \sum_i P(A_i) &= \binom{\Phi(3, 2 \times 10^4)}{2} \times 2e^{-50} \\ &= \left(\binom{1000}{0} + \binom{1000}{1} + \binom{1000}{2} + \binom{1000}{3} \atop 2\right) \times e^{-50} \\ &= \binom{1 + 1000 + 49950 + 166167000}{2} \times e^{-50} \\ &\approx 1.3322 \times 10^{-6} \end{aligned}$$

   Here, we have taken $\varepsilon = 0.1$ as in the first example. Each event A_i will actually be the event in which we compare the frequencies defined by two samples. Because of this, there will actually be $\binom{\Phi(3,2\times 10^4)}{2}$ such events. Since each comparison will happen twice, we will need to divide this factor by 2. $\square$