Foundations of Data Science & Machine Learning

Tutotial 06

April 9, 2021

Question 1. There are some red and blue balls in a jar. Your random experiment involves sampling 100 balls out of it uniformly and independently (with repetition). You repeat this experiment twice and count the number of red balls you get in each repetition, say R_1 and R_2. Show that with probability at least 0.45, $|R_1 - R_2| \leq 20$.

Hint. Hoeffding Bounds.

Hoeffding Bounds. If X is the average of n $\{0,1\}$-random variables $X_1, \ldots, X_n$ with $P(X_i = 1) = p$ for each $i \in [n]$, then

$$P(X > p + \epsilon) \leq e^{-2\epsilon^2 n}, \text{ and}$$
$$P(X < p - \epsilon) \leq e^{-2\epsilon^2 n}.$$

Answer 1. Let $r_1 = R_1/100$ and $r_2 = R_2/100$. Note that both r_1 and r_2 are averages of 100 $\{0,1\}$-random variables with the same mean μ, where μ is the share of red balls in the jar. (The value of μ is still unknown to us.)

$$\begin{aligned}
P[|R_1 - R_2| > 20] &= P[|r_1 - r_2| > 0.2] && \text{scaling} \\
&\leq P[|r_1 - \mu| > 0.1 \text{ OR } |r_2 - \mu| > 0.1] && \text{triangle inequality} \\
&\leq 2e^{-2} + 2e^{-2} && \text{Hoeffding } \epsilon = 0.1,\ n = 100 \\
&< 0.55
\end{aligned}$$

Question 2. Suppose you know that there is a total of 1000 balls in the jar and R_1 was 30 in the above experiment. You want to estimate R, the total number of red balls in the jar. What is the smallest interval $[a, b]$ for which you can guarantee that $R \in [a, b]$ with probability at least 0.9?

Answer 2. Let $a = 1000(r_1 - \epsilon)$ and $b = 1000(r_1 + \epsilon)$. Since r_1 is the average of 100 $\{0,1\}$-random variables with $P(X_i = 1) = p = R/1000$ for each $i \in [100]$, by rewriting the Hoeffding bounds, we get

$$\begin{aligned}
P[R \notin [a, b]] &= P[p < r_1 - \epsilon \text{ OR } p > r_1 + \epsilon] \\
&\leq 2e^{-2\epsilon^2 100}
\end{aligned}$$

Hence we can choose any ϵ which satisfies $2e^{-200\epsilon^2} < 0.1$, that is $\epsilon > \sqrt{\ln(20)/200}$. In particular, $\epsilon = 0.123$ will work. So we can choose the interval to be $[177, 423]$ and guarantee that $R \in [177, 423]$ with probability at least 0.9.