Question 3 solution

Zhenrui Liao

1 a

Think of Bagging n items out of n as sampling n times from the "bag" of all n items with replacement (because clearly if you sample without replacement you'll get each item once!). Suppose the probability of getting each item is uniform, i.e. $\frac{1}{n}$.

The number of times each item $i \in [n]$ shows up is a random weight $M_1, ..., M_i, ..., M_n$. What is $\Pr[M_i = m_i]$? This should be familiar: it's the probability you choose item i exactly m_i times in n trials with replacement (you don't care what you get the other $n - m_i$ times, as long as they're not i) - in other words, a Binomial distribution.

$$\Pr[M_i = m_i] = \binom{n}{m_i} \frac{1}{n^{m_i}} (\frac{n-1}{n})^{n-m_i} = \binom{n}{m_i} \frac{(n-1)^{n-m_i}}{n^n}$$

However, these variables are not independent - obviously if you get element i $n-1$ times (to take an extreme example) you can only get $j \neq i$ either 0 or 1 time. We can ask what the correlation is between m_i, m_j. Note that by symmetry this correlation is the same for all pairs $i \neq j$.

The term of interest is the cross term $\mathbb{E}\{M_i M_j\}$ (from this you get Pearson's correlation ρ_{ij} by simple algebraic manipulations). To compute this, we need to figure out the joint distribution of M_i, M_j. Without loss of generality, suppose you draw item i first from n total, and then draw element j from the $n - m_i$ that remain. Then your joint probability is

$$\Pr(M_i = m_i, M_j = m_j) = \binom{n}{m_i} \frac{1}{n^{m_i}} \binom{n-m_i}{m_j} \frac{1}{n^{m_j}} (\frac{n-2}{n})^{n-m_i-m_j}$$

$$= \frac{n!}{m_i! m_j! (n-m_i-m_j)!} \frac{(n-2)^{n-m_i-m_j}}{n^n}$$

This is an example of the Multinomial distribution with uniform probabilities. Hence, the correlation of two terms M_i, M_j is

$$\rho_{ij} = -\sqrt{\frac{\frac{1}{n} \cdot \frac{1}{n}}{(1-\frac{1}{n})(1-\frac{1}{n})}} = -\frac{1}{n-1}$$

2 b

Suppose that T is even. $\frac{1}{2} - \gamma$ represents the probability that one weak learner is "wrong". Then the probability that the majority vote of T weak learners is wrong is

$$\Pr[wrong] = \sum_{k=T/2+1}^{T} \binom{T}{k} (\frac{1}{2}-\gamma)^k (\frac{1}{2}+\gamma)^{T-k} < \sum_{k=T/2}^{T} \binom{T}{k} (\frac{1}{2}-\gamma)^k (\frac{1}{2}+\gamma)^{T-k}$$

We need to establish a bound of the form $\Pr[wrong] \leq \varepsilon$. However, this expression is quite difficult to work with, so we approximate. Observe that for each $k = T/2 \dots T$,

$$(\frac{1}{2} - \gamma)^k (\frac{1}{2} + \gamma)^{T-k} \leq (\frac{1}{2} - \gamma)^{T/2} (\frac{1}{2} + \gamma)^{T/2}$$

Now we need to deal with the $\binom{T}{k}$ term. Recall from combinatorics that

$$\sum_{k=0}^{T} \binom{T}{k} = 2^T, \binom{T}{k} = \binom{T}{T-k}$$

So

$$\sum_{k=\frac{T}{2}}^{T} \binom{T}{k} \leq 2^{T-1}$$

Now we plug and chug

$$\Pr[wrong] = \sum_{k=T/2}^{T} \binom{T}{k} (\frac{1}{2} - \gamma)^k (\frac{1}{2} + \gamma)^{T-k} < 2^{T-1} (\frac{1}{2} - \gamma)^{T/2} (\frac{1}{2} + \gamma)^{T/2}$$

Clearly, establishing the bound

$$2^{T-1} (\frac{1}{2} - \gamma)^{T/2} (\frac{1}{2} + \gamma)^{T/2} < \varepsilon$$

will automatically establish $\Pr[wrong] < \varepsilon$. We take the logarithm on both sides.

$$T - 1 + (\frac{T}{2})\lg((\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)) < \lg\epsilon$$

$$T + (\frac{T}{2})\lg(\frac{1}{4} - \gamma^2) < \lg\epsilon + \lg 2$$

We solve for T:

$$T(1 + \frac{1}{2}\lg(\frac{1}{4} - \gamma^2)) < \lg 2\epsilon$$

Here we must be careful: $\lg\frac{1}{4} = -2$ and $\frac{1}{4} - \gamma^2 < \frac{1}{4} \implies \lg(\frac{1}{4} - \gamma^2) < -2$. Thus $1 + \frac{1}{2}\lg(\frac{1}{4} - \gamma^2) < 0$ and we must reverse the inequality upon dividing

$$T > \frac{\lg 2\epsilon}{(1 + \frac{1}{2}\lg(\frac{1}{4} - \gamma^2))} \qquad (*)$$

With some more algebra,

$$T > \frac{\lg 2\epsilon}{(\frac{1}{2}\lg 4 + \frac{1}{2}\lg(\frac{1}{4} - \gamma^2))}$$

$$T > \frac{2\lg 2\epsilon}{\lg(1 - 4\gamma^2)}$$

Since $\gamma \leq \frac{1}{2}$, $0 \leq 1 - 4\gamma^2 < 1$ so $\lg(1 - 4\gamma^2) \leq -4\gamma^2$ (using the first-order Taylor expansion, which tells us that $\lg(1 - x) \leq -x$ on [0,1)). Thus we have

$$T > -\frac{2\lg 2\epsilon}{4\gamma^2}$$

So $\boxed{C = \frac{1}{2}}$.

2.1 Notes

- Correct solutions may differ by log of a constant if student chose to use the logarithm of a different base
- Full credit to be awarded for arriving at $(*)$ (even if student misses further algebraic simplifications or use of Taylor approximation)
- If the student begins the proof by assuming T odd, the final result will differ by $+1$ (assuming correct intermediate reasoning). Still award full credit.