Theoretical Ananlysis

Pramodh Gopalan

June 2021

1 Problem

We are given a dataset with n data points, each having d features, out of which b are backdoored. There are T models labeled from $M_1, \dots M_T$ given to us, which are trained on k features chosen at random from the dataset.

2 Analysis

Given a single model M_i, let E_i be the event that there is no intersection in between the k features and any backdoor in the dataset. Let X_i be a random variable defined as follows.

$$X_i = \begin{cases} 1 & \text{if } E_i \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Clearly, we have

$$P(E_i) = \frac{\binom{d-b}{k}}{\binom{d}{k}} = \alpha$$

Let X be a random variable that is defined as the number of successes of E_i over all is. Verbosely, it denotes the number of models (out of T) that have no intersection with a backdoor. Trivially, X can be seen as the sum of T independent Bernoulli variables X_i.

Now, we require $P[X \geq \frac{T}{2}]$.

2.1 Chernoff Bound based analysis

One way to approach this is through the Chernoff's bound, For definiton and corollaries, please refer to Corollary 13.3 linked. Here is a derivation of the bound in our case. From the chernoff bound, we get this corollary:

Given $X_1, \dots X_n$ with $X = \sum_{i=1}^{n} X_i$ be independent RV, not necessarily from the same distribution, and some real number $\epsilon \in [0, 1]$. Setting $\mu = E[\sum_{i=1}^{n} X_i]$, we have:

$$P(X \leq (1-\epsilon)\,\mu) \leq \exp\left(-\frac{\epsilon^2}{2}\mu\right)$$

.

In our case, with $X_i s$ being i.i.d Bern(α), $\mu = T\alpha$. We require $P[X \geq \frac{T}{2}]$, so $(1-\epsilon)\,\mu = \frac{T}{2}$. thus giving the equations:

$$(1-\epsilon)\,T\alpha = \frac{T}{2}$$

Implying

$$\epsilon = 1 - \frac{1}{2\alpha}$$

Note that we also have the constraint $\epsilon \geq 0$, hence giving us:

$$1 \geq \frac{1}{2\alpha} \implies \alpha \geq \frac{1}{2}$$

The other constraint $\epsilon \leq 1$ is satisfied automatically since $\alpha > 0$:

$$\epsilon = 1 - \frac{1}{2\alpha} < 1$$

Substituting back into the bound, we get:

$$P[X \leq \frac{T}{2}] \leq e^{-\frac{T\alpha}{2}(1-\frac{1}{2\alpha})^2}$$

$$P[X \leq \frac{T}{2}] \leq e^{-\frac{T}{2\alpha}(\alpha-\frac{1}{2})^2}$$

Consequently, we get:

$$P[X > \frac{T}{2}] \geq 1 - e^{-\frac{T}{2\alpha}(\alpha-\frac{1}{2})^2}$$

and

$$P[X \geq \frac{T}{2}] \geq 1 - e^{-\frac{T}{2\alpha}(\alpha-\frac{1}{2})^2} + P(X = \frac{T}{2})$$

The latter term in the above equation is 0, if T is odd. We need $P(X \geq \frac{T}{2})$, and not $P(X > \frac{T}{2})$ since in we are fine with a tie in between the models at prediction time.

Hence, this holds only when $\alpha \geq \frac{1}{2}$. For the other case, please refer to the next page. For most practical cases, such as MNIST and DREBIN, $\alpha << \frac{1}{2}$

For the other case, when $\alpha < \frac{1}{2}$, we use the second part of Corollary 13.3, which states:

Given $X_1, \ldots X_n$ with $X = \sum_{i=1}^{n} X_i$ be independent RV, not necessarily from the same distribution, and some real number $\epsilon > 0$. Setting $\mu = E[\sum_{i=1}^{n} X_i]$, we have:

$$P(X \geq (1+\epsilon)\,\mu) \leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right)$$

.

Like before, $\mu = T\alpha$, and we need $P[X \geq \frac{T}{2}]$, so $(1+\epsilon)\,\mu = \frac{T}{2}$

$$(1+\epsilon)\,T\alpha = \frac{T}{2}$$

Implying

$$\epsilon = \frac{1}{2\alpha} - 1$$

Note, here too we need $\epsilon > 0$, hence we have

$$1 < \frac{1}{2\alpha} \implies \alpha < \frac{1}{2}$$

Hence, this holds only when $\alpha < \frac{1}{2}$ which covers our other case, and we are done.

Substituting back into the bound, we get:

$$P[X \geq \frac{T}{2}] \leq e^{-\frac{T\alpha}{1+\frac{1}{2\alpha}}(\frac{1}{2\alpha}-1)^2}$$

$$P[X \geq \frac{T}{2}] \leq e^{-\frac{T}{\alpha+\frac{1}{2}}(\alpha-\frac{1}{2})^2}$$

which holds if $\alpha < \frac{1}{2}$

NOTE: One important thing to notice is that when $\alpha < \frac{1}{2}$, we get an upper bound, and when $\alpha \geq \frac{1}{2}$, we get a lower bound. This is problematic.

2.2 Gaussian approximation to the Binomial distribution

Given a binomial distribuion $Binomial(n, p)$ with n trials, it can be approximated with normal distribution $\mathcal{N}(np, np(1-p))$, under the conditions that either $p \approx 1/2$ or n is large. So, in our case, the approximating normal distribution is given by $\mathcal{N}(T\alpha, T\alpha(1-\alpha))$. We approximate the quantity we need as the follows

$$P(X \geq T/2) \approx 1 - cdf(T/2)$$

where cdf returns the cumulative distribution function of the Normal approximation defined above.

2.3 Exact numerical computation

Since we need to evaluate $P(X > T/2)$, we directly compute it as follows:

$$P(X \geq T/2) = \sum_{i=T/2}^{T} \binom{T}{i} \alpha^i (1-\alpha)^{T-i}$$

NOTE: This method can result in non-exact answers (due to floating point approximations) when $\alpha << 1$.