Tags: Distribution Testing *Written By: Nimitt*

White Noise vs Data

This document describes an algorithm to test if data observed from an unknown source has some signal or it is only white noise. Formally, given $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$, we test if $\mu = 0$ versus $\mu \geq \varepsilon$ for some pre-specified threshold $\varepsilon \in (0, 1]$, with probability at least $99/100$.

Algorithm 1 White Noise vs Data

Require: n : number of samples; $X_1, X_2, \cdots, X_n$: samples; $\varepsilon \in (0, 1]$: prespecified threshold

$\tau \leftarrow \varepsilon^2/16$

$T \leftarrow \frac{1}{n(n-1)} \sum_{i_1 \neq i_2 \in [n]} \left(\mathbb{I}(X_{i_1} \geq 0) - \frac{1}{2}\right) \left(\mathbb{I}(X_{i_2} \geq 0) - \frac{1}{2}\right)$

if $T \geq \tau$ **then**

  return *True*

else

  return *False*

end if

Theorem 1. *Algorithm 1 takes $O(1/\varepsilon^2)$ samples from an unknown gaussian distribution $\mathcal{N}(\mu, 1)$, and outputs:*
(i) True*, if $\mu \geq \varepsilon$*
(ii) False*, if $\mu = 0$*
for some pre-specified threshold $\varepsilon \in [0, 1]$. Moreover, it is optimal in number of samples.

Proof. Let for all $i \in [n]$, $Y_i = \mathbb{I}(X_i \geq 0)$. Note that, Y_i are i.i.d Bernoulli samples with bias (let)

$$p := \Pr[Y_i = 1] = \Pr[X_i \geq 0] = \frac{1}{2}\text{Erfc}\left(-\frac{\mu}{\sqrt{2}}\right)$$

Expectation of T. We compute,

$$\mathbb{E}[T] = \frac{1}{n(n-1)} \sum_{i_1, \neq i_2} \left(p - \frac{1}{2}\right)\left(p - \frac{1}{2}\right) = (p - 1/2)^2$$

Variance of T. Now, we bound $\text{Var}[T]$ as follows. Let $Z_i = Y_i - 1/2$ and $S = \sum_{i=1}^n Z_i$, then,

$$n(n-1)T = \left(\sum_{i=1}^n Z_i\right)^2 - \sum_{i=1}^n Z_i^2 = S^2 - n/4$$

where last inequality since $Z_i^2 = 1/4$. Thus, $n^2(n-1)^2\text{Var}[T] = \text{Var}[S^2] = \mathbb{E}[S^4] - \mathbb{E}[S^2]^2$. Now,

$$\mathbb{E}[S^2] = n(n-1)\mathbb{E}[T] + n/4 = n(n-1)(p - 1/2) + n/4$$

$$\mathbb{E}[S^2]^2 = n^2(n-1)^2(p-1/2)^2 + n^2(n-1)(p-1/2)^2/2 + n^2/4$$

Now, we bound $\mathbb{E}[S^4]$. Expanding S^4,

$$\begin{aligned}\mathbb{E}[S^4] = \sum_k \mathbb{E}[Z_k^4] + 4 \sum_{k_1 \neq k_2} \mathbb{E}[Z_{k_1}^3 Z_{k_2}] \\ + 3 \sum_{k_1 \neq k_2} \mathbb{E}[Z_{k_1}^2 Z_{k_2}^2] \\ + 6 \sum_{k_1 \neq k_2 \neq k_3} \mathbb{E}[Z_{k_1}^2 Z_{k_2} Z_{k_3}] \\ + \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \mathbb{E}[Z_{k_1} Z_{k_2} Z_{k_3} Z_{k_4}].\end{aligned}$$

Using

$$Z_k^2 = \frac{1}{4}, \qquad Z_k^4 = \frac{1}{16}, \qquad \mathbb{E}[Z_k] = \tilde{p}_i, \qquad \mathbb{E}[Z_k^3] = \frac{\tilde{p}_i}{4},$$

we obtain

$$\begin{aligned}\mathbb{E}[S^4] = \frac{n}{16} + \frac{3n(n-1)}{16} + n(n-1)\tilde{p}_i^2 \\ + \frac{3}{2} n(n-1)(n-2)\tilde{p}_i^2 \\ + n(n-1)(n-2)(n-3)\tilde{p}_i^4.\end{aligned}$$

Using the above expressions for $\mathbb{E}[S^2]$ and $\mathbb{E}[S^4]$

$$\begin{aligned}n^2(n-1)^2 \mathrm{Var}(T) = \frac{n(n-1)}{8} + n(n-1)(n-2)(p-1/2)^2 \\ - 2n(n-1)(2n-3)(p-1/2)^4.\end{aligned}$$

Using $(p-1/2)^2 = \mathbb{E}[T]$, we get,

$$\mathrm{Var}[T] \leq \frac{1}{8n(n-1)} + \frac{(n-2)\mathbb{E}[T]}{n(n-1)}$$

We are ready to show that the error probability is less than 1/100 using Chebyshev in the following two cases:

Case ($\mu = 0$) Then, $p = 1/2$; samples Y_i are from a uniform Bernoulli distribution and $\mathbb{E}[T] = 0$. Using Chebyshev we have,

$$\Pr[T \geq \tau] = \Pr[T - \mathbb{E}[T] \geq \varepsilon^2/16] \leq \frac{256\mathrm{Var}(T)}{\varepsilon^4} = \frac{32}{n(n-1)\varepsilon^2} < 1/100$$

where last inequality whenever $n > C_1 \cdot (1/\varepsilon^2)$ for some $C_1 > 0$.

Case ($\mu \geq \varepsilon$) Using standard properties of $\mathrm{Erfc}(x)$, we have, $(p-1/2)^2 \geq \varepsilon^2/8$. Then, $\mathbb{E}[T] \geq \varepsilon^2/8$ and we get,

$$\begin{aligned}\Pr[T < \tau] = \Pr[\mathbb{E}[T] - T \geq \mathbb{E}[T]/2] &\leq \frac{1}{2n(n-1)\mathbb{E}[T]^2} + \frac{4(n-2)}{n(n-1)\mathbb{E}[T]} \\ &\leq \frac{8}{n(n-1)\varepsilon^2} + \frac{64(n-2)}{n(n-1)\varepsilon^2} < \frac{1}{100}\end{aligned}$$

where last inequality whenever $n > C_2 \cdot (1/\varepsilon^2)$ for some $C_2 > 0$. □