Proofs

Here we provided proofs for **Group Anomaly Detection Using Minimum Description Length** paper.

Proof that the KS test is universal

Let F be the CDF corresponding to P and $\tilde{F}$ the CDF corresponding to $\widehat{P}$. If $\widehat{P} \neq P$, $d = \sup_x |\tilde{F}(x) - F(x)| > 0$. Let F_M be the empirical CDF based on M samples. By the Glivenko-Cantelli Theorem (e.g., [1]), under H_0, $\sup_x |F_M(x) - F(x)| \xrightarrow{P} 0$. Under H_1 we can write $|F_M(x) - F(x)| \geq |\tilde{F}(x) - F(x)| - |F_M(x) - \tilde{F}(x)|$ and therefore also by Glivenko-Cantelli, $\sup_x |F_M(x) - F(x)| \xrightarrow{P} \tilde{d} \geq d$. we also know [1] that $\sqrt{M}\sup_x |F_M(x) - F(x)| \xrightarrow{D} B$, where B is the Kolmogorov distribution. Therefore if we choose a threshold $\tau = \frac{\ln M}{\sqrt{M}}$ we get $P_{FA} \to 0$, and since $\tau \to 0$, for large enough M, $\tau < d$ for any d. Then $P_D \to 1$ for any $\widehat{P}$.

Proof of Eq. 8

By using the Taylor series we have

$$D(\hat{p}\|\frac{1}{K})\ln 2 = \sum_{i=1}^{K} \hat{p}_i \ln(\frac{\hat{p}_i}{1/K}) = 0 + \sum_{i=1}^{K} \hat{\hat{p}}_i + \frac{K}{2}\sum_{i=1}^{K} \hat{\hat{p}}_i^2 + \frac{1}{K}\sum_{i=1}^{K}\sum_{n=3}^{\infty} \frac{(-K\hat{\hat{p}}_i)^n}{n(n-1)} = K\sum_{i=1}^{K} \hat{\hat{p}}_i^2 + \frac{1}{K} g_K(\hat{\hat{p}})$$

where $\hat{\hat{p}}_i = \hat{p}_i - \frac{1}{K}$ for $i = 1, \ldots, K$. Note that $\lim_{M\to\infty} \sum_{i=1}^{K} \hat{\hat{p}}_i = 0$. Also,

$$\begin{aligned}
\lim_{M\to\infty} g_K(\hat{\hat{p}}) &= \lim_{M\to\infty} \sum_{i=1}^{K}\sum_{n=3}^{\infty} \frac{(-K\hat{\hat{p}}_i)^n}{n(n-1)} - \frac{K^2}{2}\sum_{i=1}^{K} \hat{\hat{p}}_i^2 + K\sum_{i=1}^{K} \hat{\hat{p}}_i \\
&= \lim_{M\to\infty} \sum_{i=1}^{K}\sum_{n=3}^{\infty} (\frac{1}{n-1} - \frac{1}{n})(-K\hat{\hat{p}}_i)^n - \frac{K^2}{2}\sum_{i=1}^{K} \hat{\hat{p}}_i^2 \\
&= \lim_{M\to\infty} -\sum_{i=1}^{K} \frac{1}{K\hat{\hat{p}}_i}\sum_{n=2}^{\infty} \frac{(-K\hat{\hat{p}}_i)^n}{n} - \sum_{i=1}^{K}\sum_{n=3}^{\infty} \frac{(-K\hat{\hat{p}}_i)^n}{n} - \frac{K^2}{2}\sum_{i=1}^{K} \hat{\hat{p}}_i^2 \\
&= \lim_{M\to\infty} -\sum_{i=1}^{K} \frac{K\hat{\hat{p}}_i}{2} - \sum_{i=1}^{K} (1 + \frac{1}{K\hat{\hat{p}}_i})\sum_{n=3}^{\infty} \frac{(-K\hat{\hat{p}}_i)^n}{n} - \frac{K^2}{2}\sum_{i=1}^{K} \hat{\hat{p}}_i^2 \\
&= \lim_{M\to\infty} \sum_{i=1}^{K} \frac{(1 + K\hat{\hat{p}}_i)\ln(1 + K\hat{\hat{p}}_i)}{K\hat{\hat{p}}_i} - K
\end{aligned}$$

where it is a fair assumption that $|\hat{\hat{p}}_i| < \frac{1}{K}$. Note that $1 - |x| \leq \frac{(1+x)\ln(1+x)}{x} \leq 1 + |x|$. So, $\lim_{M\to\infty} g_K(\hat{\hat{p}}) = 0$.

By noting that $\lim_{M\to\infty} \hat{\hat{p}}_K = -\lim_{M\to\infty} \sum_{i=1}^{K-1} \hat{\hat{p}}_i$ we can summarise the results as:

$$\lim_{M\to\infty} D(\hat{p}\|\frac{1}{K}) = \lim_{M\to\infty} \frac{K}{\ln 2} \hat{\hat{p}}^T (J_{K-1} + I_{K-1}) \hat{\hat{p}}$$

where, $\hat{\hat{p}} = [\hat{\hat{p}}_1, \ldots, \hat{\hat{p}}_{K-1}]$, and I_{K-1} and J_{K-1} are the identity matrix and unit matrix of size $K-1$, respectively.

As a result, $D(\hat{p}\|\frac{1}{K}) \xrightarrow{d} \frac{K}{\ln 2}\hat{\hat{p}}^T(J_{K-1}+I_{K-1})\hat{\hat{p}}$ [2]. Lets recall the distribution of $\hat{\hat{p}} \sim \mathcal{N}(\frac{1}{K}\mathbf{1}_{K-1}, \mathbf{C})$ where $\mathbf{C} = \frac{1}{MK^2}(-J_{K-1} + KI_{K-1})$. Note that for any $\alpha \in \mathbb{R}$ and $n \in \mathbb{N}$,

$$(J_n + \alpha I_n)^{-1} = \frac{1}{\alpha(n+\alpha)}(-J_n + (n+\alpha)I_n)$$

Thus, $J_{K-1} + I_{K-1} = (KM\mathbf{C})^{-1}$ and

$$M \ln 2 D(\hat{p}\|\frac{1}{K}) \xrightarrow{d} MK\hat{\hat{p}}^T(J_{K-1} + I_{K-1})\hat{\hat{p}} \xrightarrow{d} \hat{\hat{p}}^T\mathbf{C}^{-1}\hat{\hat{p}} \sim \chi^2_{K-1}$$

References

[1] V. N. Vapnik, *Statistical Learning Theory.* John Wiley, 1998.

[2] R. J. Serfling, *Approximation theorems of mathematical statistics.* John Wiley & Sons, 2009.