# A FLEXIBLE TEST FOR DISTRIBUTION CHANGE

## 1. The Problem

Suppose we have a stream of events drawn independently from some unknown distribution $D$ over a space $X$. At some the distribution shifts to a new distribution $D'$, and we would like to detect this shift.

Naturally, we do not care about all changes from $D$ to $D'$ because many of them are simply unimportant. Consequently, we constrain ourselves to look at changes between $D$ and $D'$ through a particular language $L = \{\sigma\}$ where each $\sigma \subseteq X$ is a subset of the events. Our goal is to construct a test which detect if there is a $\sigma$ such the the measure $\Pr_{x\sim D}(x \in \sigma)$ is significantly less than the measure, $\Pr_{x\sim D'}(x \in \sigma)$.

## 2. The Method

There are two sets of events that we start with: the set of past events $S$ and the set of new events $S'$. For any particular $\sigma$ and set $S$, we can define:

$$\sigma_S = \Pr_{x\in S}(x \in \sigma) = \frac{1}{|S|}\sum_{x\in S} I(x \in S)$$

It is important to take into account the many $\sigma$ under consideration because running many tests that the distribution has changed can result in too many false positives. To take this into account consider a description language[1] over $\sigma$ given the combined data sets $S$ and $S'$. Let $|\sigma|$ denote the description length of $\sigma$ given $S$ and $S'$. We use the PAC-MDL bound to mediate between these many tests given the description language for $L$.

The precise bound is defined in terms of a cumulative hypergeometric distribution. To understand this distribution, imagine a bucket with $m$ red balls and $n$ blue balls, from which $a + b$ balls are drawn without replacement. Now, define $\mathrm{Bucket}(m, n, a, b)$ to be the probability that at least $b$ blue balls are drawn. That is,

$$\mathrm{Bucket}(m, n, a, b) \quad = \quad \sum_{t=b}^{a+b} \frac{\binom{n}{t}\binom{m}{a+b-t}}{\binom{n+m}{a+b}}.$$

The bound is actually defined in terms of a "worst case" over the value of $b$ defined according to:

$$\mathrm{bmax}(m, n, a, \delta) \quad = \quad \max\{b : \mathrm{Bucket}(m, n, a, b) \geq \delta\}$$

**Theorem 1.** *(PAC-MDL bound) For any distribution $D$, for any label description language $\{\sigma\}$, with probability at least $1 - \delta$ over the draw of the two sets $S, S' \sim D^{m+n}$:*

$$\forall\sigma \quad \hat{\sigma}_{S'} \leq \mathrm{bmax}(m, n, \hat{\sigma}_S, 2^{-|\sigma|}\delta))\}$$

[1]For example, we might consider the language which specifies the index of 3 points defining a circle.

For details on this theorem (and it's proof) see [1]. For now, it is important to note that this theorem holds with probability $1-\delta$ whenever the train and test sets are each drawn independently from the distribution $D$. Since we can precisely evaluate $\sigma_{S'}$ and $\sigma_S$, we can explicitly test to see whether or not this theorem holds. When $\delta$ is reasonably small, bound failures only occur when the distribution producing $S'$ differs from the distribution producing $S$.

## References

[1] Avrim Blum and John Langford, "PAC-MDL bounds", COLT 2003