DESCRIPTION OF KAWAHARA AND SUGIYAMA'S ONLINE CHANGEPOINT DETECTION
ALGORITHM BASED ON DIRECT DENSITY-RATIO ESTIMATION

Christopher Natoli

1   Framework of the solution

Suppose $\{\boldsymbol{y}(t)\}_{t=1}^{T} \subset \mathbb{R}^{d}$ is a d-dimensional time series and let $\boldsymbol{Y}(t) \in \mathbb{R}^{dk}$ be a k-long subsequence of the data beginning at time t:

$$\boldsymbol{Y}(t) := \left[\boldsymbol{y}(t)^{\top}, \boldsymbol{y}(t+1)^{\top}, \ldots, \boldsymbol{y}(t+k-1)^{\top}\right]^{\top},$$

i.e., $\boldsymbol{Y}(t)$ is the concatenation of the k vectors $\boldsymbol{y}(t), \ldots, \boldsymbol{y}(t+k-1)$. Rather than consider the "feature space" or "state space" $\mathbb{R}^{d}$, i.e., the space in which the datapoints live, we will consider the space $\mathbb{R}^{dk}$ in which the subsequences $\boldsymbol{Y}(t)$ live, since these subsequences can describe some of the temporal structure of the time series that points in $\mathbb{R}^{d}$ cannot.

This online algorithm makes use of a sliding window. Divide the sliding window into two intervals, one before the possible changepoint called the reference interval and one after the possible changepoint called the test interval. Denote the beginnings of the reference and test intervals by t_{rf} and t_{te}, respectively, and denote their lengths by n_{rf} and n_{te}. If t is the current time, i.e., the front of the sliding window, then $t = t_{\text{te}} + n_{\text{te}}$ and $t_{\text{te}} = t_{\text{rf}} + n_{\text{rf}}$. Define $\boldsymbol{Y}_{\text{rf}}(i) = \boldsymbol{Y}(t_{\text{rf}}+i-1)$ and $\boldsymbol{Y}_{\text{te}}(i) = \boldsymbol{Y}(t_{\text{te}}+i-1)$. Let p_{rf} and p_{te} be the probability densities of the reference and test intervals, respectively.

We want to test

$$H_0 \colon p(\boldsymbol{Y}(i)) = p_{\text{rf}}(Y(i)) \qquad \text{for } i = t_{\text{rf}}, \ldots, t-1$$

$$\text{vs} \qquad H_1 \colon p(\boldsymbol{Y}(i)) = p_{\text{rf}}(Y(i)) \qquad \text{for } i = t_{\text{rf}}, \ldots, t_{\text{te}}-1$$

$$p(\boldsymbol{Y}(i)) = p_{\text{te}}(Y(i)) \qquad \text{for } i = t_{\text{te}}, \ldots, t-1.$$

The likelihood ratio and log-likelihood ratio for these hypotheses are therefore

$$\Lambda = \frac{\prod_{i=1}^{n_{\text{rf}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{rf}}(i)) \prod_{i=1}^{n_{\text{te}}} p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{\prod_{i=1}^{n_{\text{rf}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{rf}}(i)) \prod_{i=1}^{n_{\text{te}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))}$$

$$= \frac{\prod_{i=1}^{n_{\text{te}}} p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{\prod_{i=1}^{n_{\text{te}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))}$$

$$S := \log \Lambda = \sum_{i=1}^{n_{\text{te}}} \log \frac{p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))}.$$

We choose the log-likelihood ratio S as our score. If the score is greater than a certain threshold, then we say that tme t_{te} is a changepoint.

2 Kullback-Leibler importance estimation

Rather than estimate the density ratio $w := \frac{p_{\text{te}}}{p_{\text{rf}}}$ by first individually estimating the densities p_{te} and p_{rf} and then dividing, the ratio itself will be directly estimated by employing the Kullback-Leibler Importance Estimation Procedure (KLIEP) developed by Sugiymama, et. al. [cite!].

Suppose the estimate $\hat{w}$ lives in a finite-dimensional function space spanned by the basis functions $\{\phi_\ell\}_{\ell=1}^{b}$, where $\phi_\ell \colon \mathbb{R}^{dk} \to \mathbb{R}^{+}$ for all ℓ. Then $\hat{w}(\cdot) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\cdot)$, and so the parameters we want to learn are the coordinates $\{\alpha_\ell\}_{\ell=1}^{b}$. We do so by minimizing the Kullback-Leibler divergence of $\hat{p}_{\text{te}} = p_{\text{rf}}\hat{w}$ from the true distribution p_{te}, i.e.,

$$D_{KL}(p_{\text{te}}\|\hat{p}_{\text{te}}) = \int p_{\text{te}}(\boldsymbol{Y}) \log \frac{p_{\text{te}}(\boldsymbol{Y})}{p_{\text{rf}}(\boldsymbol{Y})\hat{w}(\boldsymbol{Y})}\, d\boldsymbol{Y}.$$

The minimization problem is subject to two constraints. First, w is nonnegative: requiring $\alpha_\ell \geq 0$ for all ℓ implies $\hat{w} \geq 0$. Second, p_{te} is a probability density and therefore integrates to 1: the same should hold for $p_{\text{rf}}\hat{w}$. Approximation reduces this constrained minimization problem, which is convex, to the following maximization problem:

$$\max_{\{\alpha_\ell\}_{\ell=1}^{b}} \left\{ \sum_{i=1}^{n_{\text{te}}} \log \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{Y}_{\text{te}}(i)) \right\}$$

$$\text{subject to } \frac{1}{n_{\text{rf}}} \sum_{i=1}^{n_{\text{rf}}} \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{Y}_{\text{rf}}(i)) = 1$$

$$\alpha_\ell \geq 0, \quad \ell = 1, \ldots, b.$$

For this changepoint algorithm, we choose that the basis functions $\phi_\ell(\cdot)$ are Gaussian kernels $K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(\ell))$ centered at the test subsequences, where

$$K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}') = \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{Y}'\|_2^2}{2\sigma^2} \right).$$

It follows that $b = n_{\text{te}}$. The rationale for centering kernels on the test points is that it is effective to place kernels where the target function w is large. The ratio w is large when p_{te} is large and/or when p_{rf} is small. The density p_{te} of the test distribution is probably large when its argument is pulled from the test distribution, i.e., a test point $\boldsymbol{Y}_{\text{te}}(i)$. Kernels are therefore centered at the test points $\{\boldsymbol{Y}_{\text{te}}(i)\}_{i=1}^{n_{\text{te}}}$.

The kernel parameter σ is selected by cross-validation. The maximization problem is solved by gradient ascent.

3 Updating $\{\alpha_\ell\}$ online

Kawahara and Sugiyama propose a way to update $\{\alpha_\ell\}$ in an online manner rather than repeating the above procedure with every new datapoint.