Learning to Aggregate Information for Sequential Inferences

Diyan Teng and Emre Ertin
The Ohio State University

THE OHIO STATE UNIVERSITY

Problem Statement

Given samples $\{\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_M^{(0)}\}$ and $\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_N^{(1)}\}$ from hypothesis H_0 and H_1 respectively.

Goal: Learn a mechanism $\{\delta, \gamma, \eta\}$ to sequentially classify testing sequence, where $\delta : \mathbb{R} \to \{0,1\}$ is the stopping rule, $\gamma : \mathbb{R} \to \{0,1\}$ is the final decision rule, $\eta : \mathbb{R}^d \to \mathbb{R}$ is the information accumulation rule using a single real number.

Criterion: Minimize $\pi_0 N_0 + \pi_1 N_1$ given P_M and P_F constraints.

Assumption: $\{\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_M^{(0)}\}$ and $\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_N^{(1)}\}$ are conditionally iid, same for testing sequence.

Background

Wald's Sequential Probability Ratio Test (SPRT) [1] is known to be optimal for a binary sequential detection problem in the sense that it achieves the minimal sampling cost given a fixed error rate requirement. Under *zero-overshoot* assumption, the relationship between termination boundary and expected cost are given in [2] as:

$$P_F = \frac{1-A}{B-A}$$

$$P_M = \frac{A(B-1)}{B-A}$$

$$N_0 = \frac{1}{-D_{01}} \frac{(1-A)\log(B) - (B-1)\log(A)}{B-A}$$

$$N_1 = \frac{1}{D_{10}} \frac{(B-1)A\log(A) - (1-A)B\log(B)}{B-A}$$

where B and A are respectively the upper and lower terminating boundary in likelihood ratio domain.

Our Approach

Kernel Functional Estimation: Construct kernel based estimate for the log-likelihood ratio function

Log-likelihood Ratio Accumulation: Use sum of log-likelihood ratio with termination boundaries using zero-overshoot assumption

Cost Approximation: Derive error and sampling cost using Martingale theory

Kernel Weights Optimization: Optimize kernel weights to minimize sampling cost with constraints

Theorem 1. *The process* $\hat{M}_n = \frac{\exp(\mu \sum_{i=1}^n \log(\hat{r}_i))}{E[\exp(\mu \log(\hat{r}))]^n}$ *is a Martingale.*

Evaluate $E[\hat{M}_N] = 1$ for a terminating time N for each hypothesis. Error rate:

$$\hat{P}_F = \frac{E[\hat{r}|H_0]^N - A}{B-A}$$

$$\hat{P}_M = \frac{A(B - E[\hat{r}^{-1}|H_1]^N)}{B-A}$$

Constraints:

$$E[\hat{r}|H_0] = 1$$

and

$$E[\hat{r}^{-1}|H_1] = 1$$

Objective: Choose $\hat{r}$ such that:

$$C(\hat{r}) = \pi_0 \hat{N}_0 + \pi_1 \hat{N}_1$$

is minimized. From Martingale property:

$$C(\hat{r}) = \frac{\omega_0}{E[-\log(\hat{r})|H_0]} + \frac{\omega_1}{E[-\log(\hat{r}^{-1})|H_1]}$$

Idea developed in [3, 4]:

$$\begin{aligned} C(r) &\le \frac{\omega_0}{\int \sup_f \{f \cdot r + 1 + \log(-f)\} p_0} + \frac{\omega_1}{\int \sup_g \{g \cdot \frac{1}{r} + 1 + \log(-g)\} p_1} \\ &\le \inf_f \frac{\omega_0}{\int \{f \cdot r + 1 + \log(-f)\} p_0} + \frac{\omega_1}{\int \{\frac{1}{f} \cdot \frac{1}{r} + 1 + \log(-\frac{1}{f})\} p_1} \\ &= \inf_f \frac{\omega_0}{\int \{f \cdot p_1 + p_0 + \log(-f) p_0\}} + \frac{\omega_1}{\int \{\frac{1}{f} \cdot p_0 + p_1 + \log(-\frac{1}{f}) p_1\}} \end{aligned}$$

Final objective:

$$\hat{r}^\star = \arg\min_{\hat{r}} \frac{\omega_0}{\int -\log(\hat{r}) p_0} + \frac{\omega_1}{\int \log(\hat{r}) p_1}$$
$$\text{s.t.} \int \hat{r} p_0 = 1$$
$$\text{and} \int \hat{r}^{-1} p_1 = 1$$

Density Ratio Estimation with a Reproducing Kernel Structure

Impose reproducing kernel structure:

$$\log(\hat{r}(\mathbf{x})) = -\sum_{i=1}^{l} \alpha_i \cdot K_i(\mathbf{x})$$

Kernel based objective:

$$\hat{r}^\star = \min_{\alpha} \frac{\omega_0}{\frac{1}{M}\sum_{j=1}^M \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})} - \frac{\omega_1}{\frac{1}{N}\sum_{i=1}^N \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})}$$
$$\text{s.t.} \frac{1}{M}\sum_{j=1}^M \exp(-\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})) = 1$$
$$\text{and} \frac{1}{N}\sum_{i=1}^N \exp(\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})) = 1$$

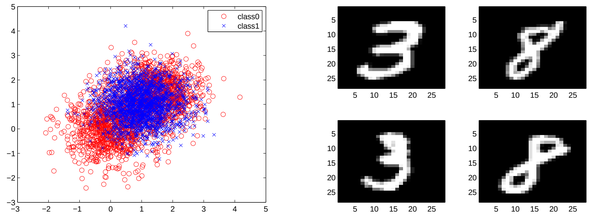
Applying inequality relaxation to obtain a convex problem:

$$\hat{r}^\star = \min_{\alpha} \frac{\omega_0}{\frac{1}{M}\sum_{j=1}^M \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})} - \frac{\omega_1}{\frac{1}{N}\sum_{i=1}^N \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})}$$
$$\text{s.t.} \frac{1}{M}\sum_{j=1}^M \exp(-\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})) \le 1$$
$$\text{and} \frac{1}{N}\sum_{i=1}^N \exp(\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})) \le 1$$

Experimental Results

1. Synthetic data with Gaussian mixture densities
2. MNIST hand written digits database

We compare our technique with the Wald-boost algorithm[5] that uses an Adaboost based likelihood ratio function estimate to aggregate information.



(a) Synthetic 2D Gaussian Mixture

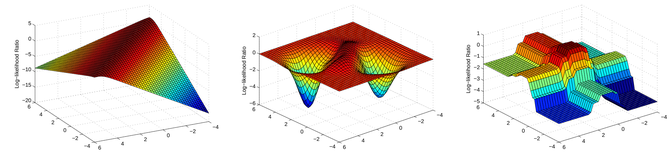
(b) MNIST Handwritten Digits

Figure 1: Experiments Data

Synthetic example:

$$H_0 : 0.5\mathcal{N}(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}) + 0.5\mathcal{N}(\begin{bmatrix}1.5\\1.5\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix})$$
$$H_1 : \mathcal{N}(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix})$$

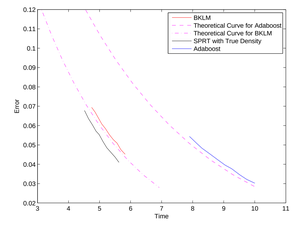
Log-likelihood ratio function and estimation:



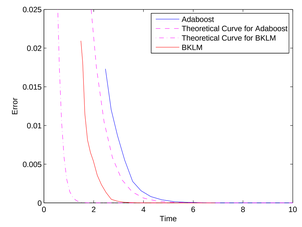
(a) True Log-likelihood Function (b) BKLM Estimation (c) Adaboost Estimation

Figure 2: 2D Log-likelihood Ratio Plot

Sequential classification experiments for both data:



(a) Error vs Time for Synthetic Data



(b) Error vs Time for MNIST Digits Data

Figure 3: Sequential Classification Experiments

References

[1] A. Wald, *Sequential analysis*. Courier Corporation, 1973.

[2] B. C. Levy, *Principles of signal detection and parameter estimation*. Springer, 2008.

[3] T. Kanamori, T. Suzuki, and M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation," *Machine Learning*, vol. 86, no. 3, pp. 335–367, 2012.

[4] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847–5861, 2010.

[5] J. Sochman and J. Matas, "Waldboost-learning for time constrained sequential detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005.*