Universal Neyman–Pearson Classification with a Known Hypothesis

Parham Boroumand
University of Cambridge
pb702@cam.ac.uk

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—We propose a universal classifier for binary Neyman–Pearson classification where null distribution is known while only a training sequence is available for the alternative distribution. The proposed classifier interpolates between Hoeffding's classifier and the likelihood ratio test and attains the same error probability prefactor as the likelihood ratio test, i.e., the same prefactor as if both distributions were known. Similarly to Hoeffding's universal hypothesis test, the proposed classifier is shown to attain the optimal error exponent tradeoff attained by the likelihood ratio test whenever the ratio of training to observation samples exceeds a certain value.**

I. Prelimenaries

Consider the following binary classification problem where an observation $\boldsymbol{x} = (x_1, \ldots, x_n)$ is generated in an i.i.d. fashion from either of two possible distributions P_0 or P_1 defined on the probability simplex $\mathcal{P}(\mathcal{X})$ with alphabet size $|\mathcal{X}| < \infty$. We assume that the distribution P_0 is known while only a sequence of training samples $\boldsymbol{z} = (z_1, \ldots, z_k) \sim P_1^k$ generated in an i.i.d. fashion from P_1 is available; training and test sequences are sampled independently from each other. We also assume that both $P_0(x) > 0, P_1(x) > 0$ and $\frac{P_0(x)}{P_1(x)} \leq c$ for each $x \in \mathcal{X}$ for some positive c. Also we let k, the length of the training, be such that $k = \alpha n$ for some positive α.

The type of an n-length sequence $\boldsymbol{y}$ is defined as $\hat{T}_{\boldsymbol{y}}(a) = \frac{N(a|\boldsymbol{y})}{n}$, where $N(a|\boldsymbol{y})$ is the number of occurrences of symbol $a \in \mathcal{X}$ in sequence $\boldsymbol{y}$. The types of the observation and training sequences $\boldsymbol{x}, \boldsymbol{z}$ are denoted by $\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}$ respectively. The set of all sequences of length n with type P, denoted by $\mathcal{T}_P^n$, is called the type class. The set of types formed with length n sequences on the simplex $\mathcal{P}(\mathcal{X})$ is denoted as $\mathcal{P}_n(\mathcal{X})$.

Let $\phi(\boldsymbol{z}, \boldsymbol{x}) : \mathcal{X}^k \times \mathcal{X}^n \to \{0, 1\}$ be a classifier that decides the distribution that generated the observation $\boldsymbol{x}$ upon processing the training sequence $\boldsymbol{z}$. We consider deterministic classifiers ϕ that decide in favor of P_0 if $\boldsymbol{x} \in \mathcal{A}_0(P_0, \boldsymbol{z})$, where $\mathcal{A}_0(P_0, \boldsymbol{z}) \subset \mathcal{X}^n$ is the decision region for the first hypothesis and is a function of P_0 and the training samples $\boldsymbol{z}$. We define $\mathcal{A}_1(P_0, \boldsymbol{z}) = \mathcal{X}^n \setminus \mathcal{A}_0$ to be the decision region for the second hypothesis. If we assume no prior knowledge on either distribution, the two possible pairwise error probabilities determine the performance of the classifier. Specifically, the type-I and type-II error probabilities are defined as

$$\epsilon_0(\phi) = \sum_{\boldsymbol{z} \in \mathcal{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathcal{A}_1(P_0, \boldsymbol{z})} P_0(\boldsymbol{x}), \tag{1}$$

$$\epsilon_1(\phi) = \sum_{\boldsymbol{z} \in \mathcal{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathcal{A}_0(P_0, \boldsymbol{z})} P_1(\boldsymbol{x}). \tag{2}$$

In the case where both distributions are known, the training sequence is not needed and the classifier becomes a hypothesis test. In this case, the classifier is said to be optimal whenever it achieves the optimal error probability tradeoff given by

$$\min_{\phi : \epsilon_0(\phi) \leq \xi} \epsilon_1(\phi), \tag{3}$$

where $\xi \in [0, 1]$. It is well known that likelihood ratio test

$$\phi^{\mathrm{lrt}}(\boldsymbol{x}) = \mathbb{1}\left\{ \frac{P_1^n(\boldsymbol{x})}{P_0^n(\boldsymbol{x})} \geq e^{n\gamma} \right\}, \tag{4}$$

attains the optimal tradeoff (3) for every γ. This is the well-known Neyman–Pearson lemma [1]. The likelihood ratio test can also be expressed as a function of the type of the observation $\hat{T}_{\boldsymbol{x}}$ as e.g.[2], [3]

$$\phi^{\mathrm{lrt}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{ D(\hat{T}_{\boldsymbol{x}} \| P_0) - D(\hat{T}_{\boldsymbol{x}} \| P_1) \geq \gamma \big\} \tag{5}$$

where $D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the relative entropy between distributions P and Q. The optimal error exponent tradeoff (E_0, E_1) is defined as

$$E_1^*(E_0) \triangleq \sup \big\{ E_1 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t.} \\ \forall n > n_0, \epsilon_0(\phi) \leq e^{-nE_0} \quad \text{and} \quad \epsilon_1(\phi) \leq e^{-nE_1} \big\}. \tag{6}$$

By using Sanov's Theorem [2], [4], the optimal error exponent tradeoff (E_1, E_0), attained by the likelihood ratio is given by

$$E_0(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathcal{Q}_0(\gamma)} D(Q\|P_0), \tag{7}$$

$$E_1(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathcal{Q}_1(\gamma)} D(Q\|P_1), \tag{8}$$

where

$$\mathcal{Q}_0(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \geq \gamma \big\}, \tag{9}$$

$$\mathcal{Q}_1(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \leq \gamma \big\}. \tag{10}$$

By varying the threshold γ in the range $-D(P_0\|P_1) \leq \gamma \leq D(P_1\|P_0)$, Eqs. (7) and (8) fully characterize the error exponent tradeoff in (6).

This work has been funded in part by the European Research Council under ERC grant agreement 725411 and by the Spanish Ministry of Economy and Competitiveness under grant PID2020-116683GB-C22.

The classification problem described above with known P_0 and a training sequence from P_1, can also be viewed as the composite binary hypothesis problem where additional training sequence samples are given for the second hypotheses. In a composite hypothesis testing setting with given P_0 and the other hypothesis is unrestricted to $\mathcal{P}(\mathcal{X})$, Hoeffding proposed the generalized likelihood-ratio test given by [5]

$$\phi^{\text{glrt}}(\boldsymbol{x}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}}\|P_0) > E_0\big\}, \tag{11}$$

By Sanov's theorem, the error exponent of Hoeffding's test is given by

$$E_0(\phi^{\text{glrt}}) = E_0, \tag{12}$$

$$E_1(\phi^{\text{glrt}}) = \min_{\substack{Q\in\mathcal{P}(\mathcal{X}),\\ D(Q\|P_0)\le E_0}} D(Q\|P_1). \tag{13}$$

By varying the threshold E_0 in the range $0 \le E_0 \le D(P_1\|P_0)$, (12) and (13) fully characterize the optimal error exponent tradeoff in (6). Using a large deviations refinement [6], [7], the type-I error probability of the likelihood ratio test can be expressed as

$$\epsilon_0(\phi^{\text{lrt}}) = \frac{1}{\sqrt{n}} e^{-nE_0}\big(c + o(1)\big), \tag{14}$$

while, for Hoeffding's test it can be expressed as [8], [6]

$$\epsilon_0(\phi^{\text{glrt}}) = n^{\frac{|\mathcal{X}|-3}{2}} e^{-nE_0}\big(c' + o(1)\big) \tag{15}$$

where c, c' are constants that only depend on P_0, P_1 and the corresponding test thresholds. Since the likelihood ratio and Hoeffding's tests attain the optimal error exponent tradeoff (6), for any fixed E_0, then $E_1(\phi^{\text{glrt}}) = E_1(\phi^{\text{lrt}})$. As a result, when the number of observations is large, Hoeffding's test, although attaining the optimal error exponent tradeoff, suffers in exponential prefactor when compared to the likelihood ratio's $\frac{1}{\sqrt{n}}$ for observation alphabets such that $|\mathcal{X}| > 2$. For $|\mathcal{X}| = 2$, the decision regions for the likelihood ratio and Hoeffding's tests coincide and thus, (15) is the same as (14).

II. Fixed Sample Sized Universal Classifier

We propose a classifier that interpolates between the likelihood ratio and Hoeffding's tests that attains a prefactor that is independent of the alphabet size and is equal to $\frac{1}{\sqrt{n}}$. In addition, we show that if the ratio of training samples to the number of test samples α exceeds a certain threshold, the proposed test also achieves the optimal error exponent tradeoff.

Hoeffding's test can favor the second hypothesis for test sequences with types close to P_0 while far from P_1. Suppose we have a training sequence type $\hat{T}_{\boldsymbol{z}}$, we can relax the Hoeffding's test from a ball centered at P_0 to a hyperplane tangent to the Hoeffding's test ball, directed towards the type of the training sequence – this is precisely what enables the improvement in the prefactor of the type-I probability of error. We propose the following classifier

$$\phi_\beta(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) = \mathbb{1}\big\{\beta D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_{\boldsymbol{z}}) - D(\hat{T}_{\boldsymbol{x}}\|P_0) \ \le \gamma(E_0, \hat{T}'_{\boldsymbol{z}})\big\}, \tag{16}$$

where $0 \le \beta \le 1$, the threshold $\gamma(E_0, Q_1)$ is given by

$$\gamma(E_0, Q_1) = \beta \min_{\substack{Q\in\mathcal{P}(\mathcal{X}),\\ D(Q\|P_0)\le E_0}} D(Q\|Q_1) - E_0, \tag{17}$$

and the perturbed training type $\hat{T}'_{\boldsymbol{z}}(a)$ is

$$\hat{T}'_{\boldsymbol{z}}(a) = \big(1-\delta_n\big)\hat{T}_{\boldsymbol{z}}(a) + \frac{\delta_n}{|\mathcal{X}|}, \tag{18}$$

where, δ_n can be chosen as any function of the order $o(n^{-1})$. We add this small perturbation of the training type to avoid the cases where some of the alphabet symbols have not been observed in the training sequence. We define the decision regions of the proposed classifier by

$$\mathcal{A}_0(\hat{T}_{\boldsymbol{z}}, \beta) = \{Q : Q \in \mathcal{P}(\mathcal{X}), \phi_\beta(Q, \hat{T}_{\boldsymbol{z}}) = 0\}, \tag{19}$$

$$\mathcal{A}_1(\hat{T}_{\boldsymbol{z}}, \beta) = \{Q : Q \in \mathcal{P}(\mathcal{X}), \phi_\beta(Q, \hat{T}_{\boldsymbol{z}}) = 1\}. \tag{20}$$

Since parameter β controls how much the training weights in the decision, we have that when $\beta = 0$ we recover Hoeffding's test while for $\beta = 1$ the test is reminiscent of a likelihood ratio test where instead of P_1, we have the perturbed training type $\hat{T}'_{\boldsymbol{z}}(a)$. Intuitively, as long as we have enough training samples, the training type $\hat{T}'_{\boldsymbol{z}}(a)$ will be close to P_1 and we will attain the optimal error exponent tradeoff.

Next, we find a refined expression for the type-I error probability and show that the error probability prefactor is of order $O(\frac{1}{\sqrt{n}})$, i.e., of the same order of the prefactor achieved by the likelihood ratio test.

Theorem 1*:* For $P_0, P_1, 0 < \beta \le 1$ and fixed E_0, the classifier ϕ_β defined in (16) attains a type-I error probability such that

$$\epsilon_0(\phi_\beta) = \frac{1}{\sqrt{n}} e^{-nE_0}(c + o(1)), \tag{21}$$

In addition, for every $P_0, P_1, E_0, \beta \in (0, 1]$, there exists a finite training to sample size ratio α^*_β such that for any $\alpha > \alpha^*_\beta$

$$\epsilon_1(\phi_\beta) = \frac{1}{\sqrt{n}} e^{-nE_1^*(E_0)}(c' + o(1)), \tag{22}$$

where c, c' are positive constants that only depend on the data distributions and E_0.

Theorem 1 shows that the classifier proposed in (16) not only achieves the optimal error exponent tradeoff for $\alpha > \alpha^*_\beta$ but also achieves the same prefactor of the type-I error probability of the likelihood ratio test. This is a significant improvement with respect to the Hoeffding's universal test for observation alphabets $|\mathcal{X}| > 2$, cf. (15). The result also shows that the proposed classifier achieves the same type-II error probability prefactor as the likelihood ratio test, establishing the optimality of the proposed classifier up to a constant. The proof of the result, as well as upper and lower bounds to α^*_β and an extension to the sequential case can be found in [9].

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933. [Online]. Available: http://www.jstor.org/stable/91247

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[3] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004. [Online]. Available: http://dx.doi.org/10.1561/0100000004

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 01 2010, vol. 95.

[5] W. Hoeffding, *Asymptotically Optimal Tests for Multinomial Distributions*. New York, NY: Springer New York, 1994, pp. 431–471.

[6] M. Iltis, "Sharp asymptotics of large deviations," *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, Jul 1995. [Online]. Available: https://doi.org/10.1007/BF02218041

[7] G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Lancho, "Saddlepoint approximation of the error probability of binary hypothesis testing," *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2306–2310, 2018.

[8] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, July 2014.

[9] P. Boroumand and A. Guillén i Fàbregas, "Universal Neyman–Pearson classification with a partially known hypothesis," *submitted to Information and Inference, a journal of the IMA, https://arxiv.org/abs/2206.11700*, 2022.