Rate-optimal Meta Learning of Classification Error

Morteza Noshad, Alfred O Hero III
Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, Michigan.
{noshad,hero}@umich.edu

Motivation

- Predicting optimal classification error is a fundamental problem in information fusion, sensor management and adaptive learning.
- Learning to classify can be very difficult, especially in high dimensions.
- Learning to predict optimal misclassification error can be much easier as it bypasses the high complexity of designing a classifier.
- The optimal meta-learner structure can lead to insights into optimal classifier design.

Introduction

Bayes error

- Z an observed r.v. with hidden state (label)

$$Z \sim \begin{cases} f_X & w/probability \quad p \\ f_Y & w/probability \quad q = 1-p \end{cases}$$

- Bayes error is the best average probability of error that can be achieved by any classifier of label.
- A sandwich bounds on the Bayes error [2]:

$$\frac{1}{2} - \frac{1}{2}\sqrt{\widetilde{D}_p(f_X, f_Y)} \le \epsilon^{Bayes} \le \frac{1}{2} - \frac{1}{2}\widetilde{D}_p(f_X, f_Y), \tag{1}$$

- $\widetilde{D}_p(f_X, f_Y) = 4pqD_p(f_X, f_Y) + (p-q)^2$.
- $D_p(f_X, f_Y)$ is the **HP-divergence** [3]:

$$D_p = 1 - \int \frac{f_X(x) f_Y(x)}{p f_X(x) + q f_Y(x)} dx. \tag{2}$$

Assumptions

- f_X and f_Y have common bounded support set.
- $0 < C_L \le f_X, f_Y \le C_U < \infty$.
- Densities are differentiable of order d.

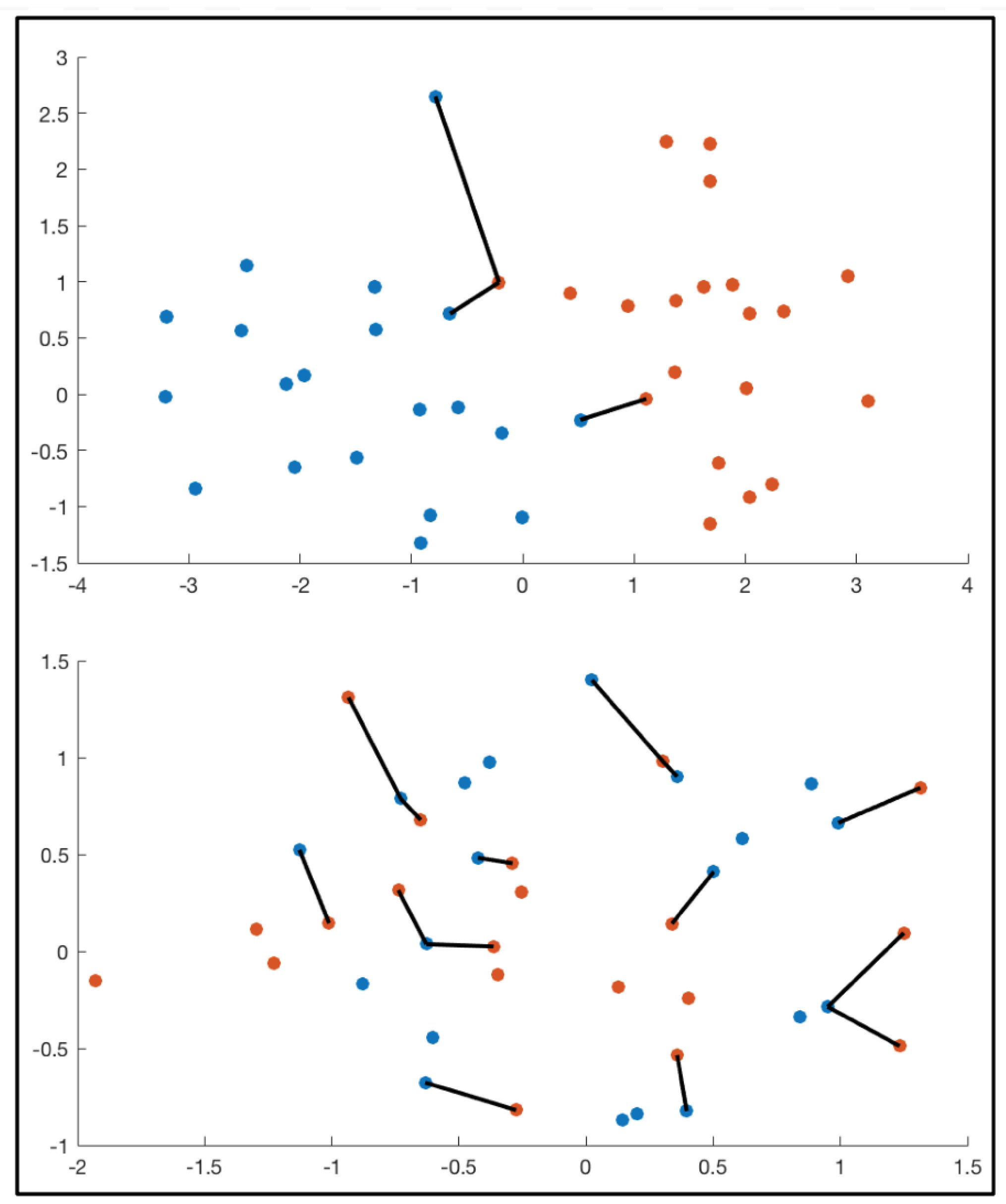
This work was supported by a grant from ARO, number W911NF-15-1-0479.

K-NN Estimator

- $\mathbf{X} = \{X_1, ..., X_N\}$ i.i.d. drawn from f_X.
- $\mathbf{Y} = \{Y_1, ..., Y_M\}$ i.i.d. drawn from f_Y.
- $M = \lfloor Nq/p \rfloor$
- Construct **k-th nearest neighbors (k-NN)** graph over $\mathbf{X} \cup \mathbf{Y}$.
- $\mathcal{E}(X, Y)$ are edges edges of k-NN graph connecting **dichotomous points**.

Direct k-NN estimator $\widehat{D_p}$ (FR [4] with K-NN):

$$\boxed{\widehat{D_p}(X, Y) = 1 - |\mathcal{E}(X, Y)| \frac{N+M}{2NM}.} \tag{3}$$



Mean-shifted (top), and identical (bottom) Normal realization from f_X and f_Y with $\mathcal{E}(X, Y)$.

Theorem: Bias Presentation

If f_1 and f_2 are differentiable up to order d, the bias of the direct k-NN estimator is

$$\mathbb{B}\left[\widehat{D}_p(X, Y)\right] = \sum_{i=1}^{d} C_i (k/N)^{i/d} + o(k/N) \tag{4}$$

Theorem: Variance

The variance of the direct k-NN estimator is bounded as

$$\mathbb{V}\left[\widehat{D}_p(X, Y)\right] \le O\left(\frac{1}{N}\right). \tag{5}$$

Runtime:

- Constructing exact KNN graph using Kd-tree algorithm requires $O(kN \log N)$ time complexity.

Ensemble Bias Reduction

- Fix a constant T where $T > d$.
- Let $\{\widehat{D}_p^{k(t)}\}_{t \in \mathcal{T}}$ be T base k-NN estimators.
- $\mathcal{T} := \{t_1, ..., t_T\}$ is a set of index values.
- $k(t) := \lfloor t\sqrt{N} \rfloor$.

Ensemble Weighted K-NN (WNN) estimator:

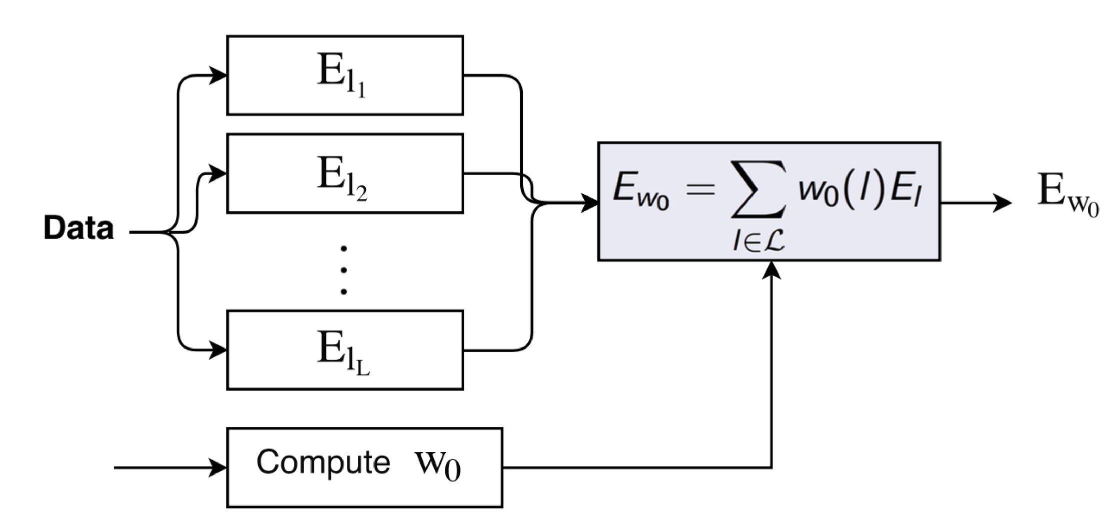
$$\widehat{D}_p^w := \sum_{t \in \mathcal{T}} w(t) \widehat{D}_p^{k(t)}, \tag{6}$$

- Bias of ensemble K-NN estimator:

$$\boxed{\mathbb{B}\left[\widehat{D}_p^w(X, Y)\right] = \sum_{i=1}^{d} C_i N^{-i/2d} \sum_{t=1}^{d} w(t) t^i + O\left(\frac{1}{\sqrt{N}}\right)} \tag{7}$$

$\Rightarrow$ Becomes $O(1/\sqrt{N})$ if $w(t)$ is selected (offline) as

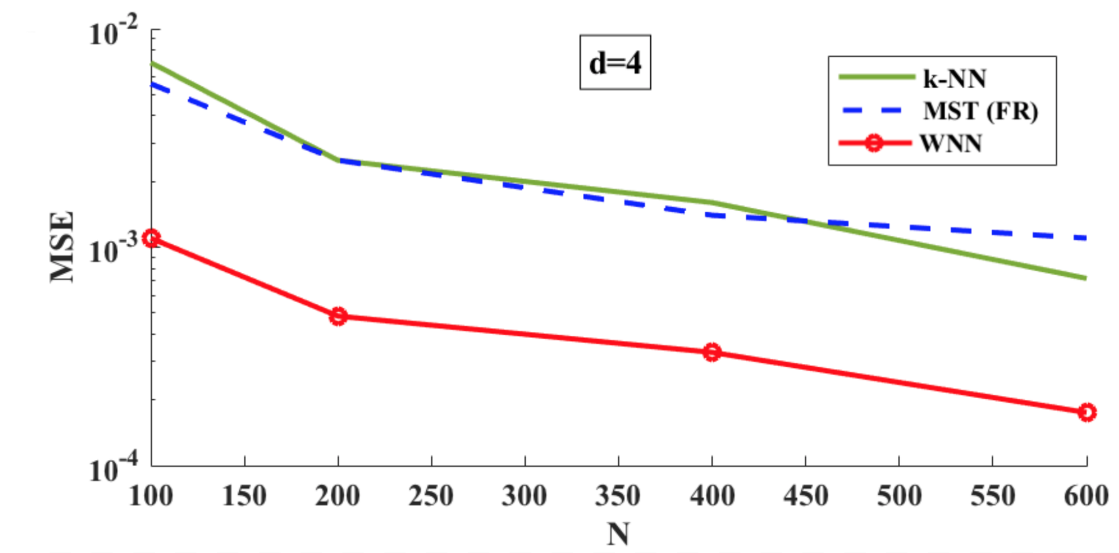
$$\boxed{\begin{aligned} &\min_{w} && \|w\|_2 \\ &subject\ to && \sum_{t \in \mathcal{T}} w(t) = 1, \\ &&& \sum_{t \in \mathcal{T}} w(t) t^i = 0, i \in \mathbb{N}, i \le d, \end{aligned}} \tag{8}$$



- An equivalent Weighted Nearest Neighbor (WNN) estimator is proposed which achieves the optimum MSE rate of $O(1/N)$.

Numerical Results

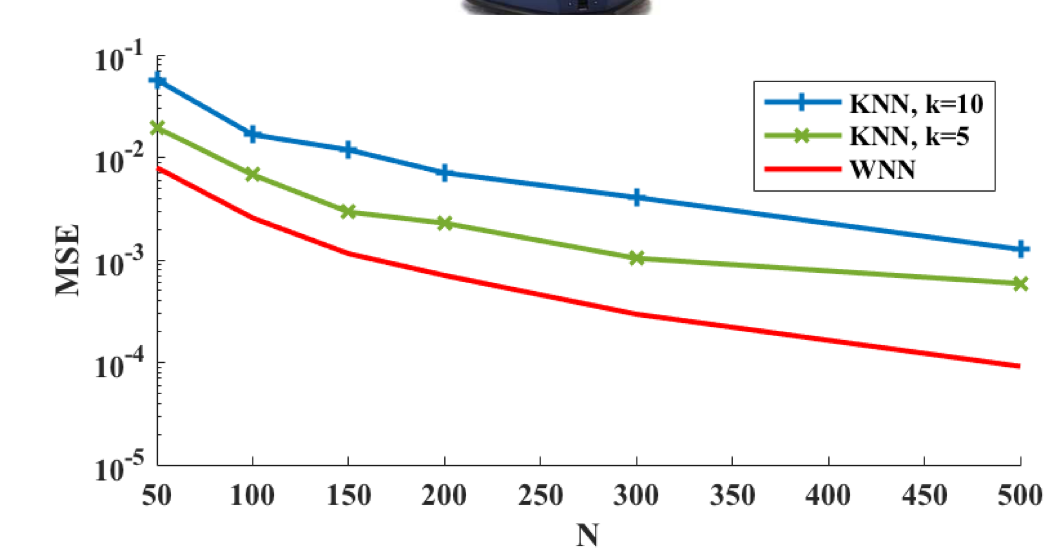
- Theory validated on simulated and real data sets.
- $d = 4$ dimensional normal distributions with the same mean at origin, and $\sigma_1^2 = \sigma_2^2 = I_4$.
- $\mathcal{T} = \{1, ..., 5\}$.



WNN outperforms other HP estimators.

Robot Navigation Dataset:

- Measurements from a set of ultrasound sensors on a navigating robot with four different actions.
- Total number of 5456 instances (corresponding to different timestamps).
- Divergence between the sensor measurements for Slight-Right-Turn and sharp-right-turn classes.



MSE for a set of ultrasound sensors arranged circularly around a robot.

Conclusion

- The WNN HP divergence estimator achieves the optimum MSE rates of $O(1/N)$.
- The computational complexity is $O(kN \log N)$.

References

[1] M. Noshad, K. Moon, S. Sekeh, and A. Hero, Direct estimation of information divergence using NNRs, ISIT 2017
[2] Berisha et al, IEEE Trans Sig Proc 2016.
[3] N. Henze, M. Penrose, Annals of statistics 1999.
[4] J. Friedman and L. Rafsky, Annals of Statistics 1979.
[5] https://archive.ics.uci.edu/ml/datasets/Wall-Following+Robot+Navigation+Data