Nearest Neighbor Classifier with Optimal Stability

Wei Sun

Department of Statistics
Purdue University

June 10, 2014
Duke University
Joint work with Xingye Qiao and Guang Cheng

Outline

- Motivations
- Classification instability and its minimax properties
- Stabilized nearest neighbor classifier
- Experiments

Motivation

- Begley and Ellis (Nature, 2012) found that 47/53 medical research papers on the subject of cancer were irreproducible.
- Marcia McNutt, Editor-in-Chief of *Science*:

  Reproducibility

  SCIENCE ADVANCES ON A FOUNDATION OF TRUSTED DISCOVERIES. REPRODUCING AN EXPERIMENT is one important approach that scientists use to gain confidence in their conclusions.

- Reproducibility is important for scientific conclusions.

Motivation

- In the paper "Stability" (Yu, 2013), Bin Yu wrote
  ...reproducibility manifests itself in the stability of statistical results relative to reasonable perturbations to data...

Motivation

- In the paper "Stability" (Yu, 2013), Bin Yu wrote
  ...reproducibility manifests itself in the stability of statistical results relative to reasonable perturbations to data...
- Stability has been of a great concern in statistics:
  - Breiman (1996) on instability for model selection
  - Bousquet and Elisseeff (2002) derived GE bound via stability
  - Ben-Hur et al. (2002) on stability for structure detection
  - Wang(2010) on stability for selecting number of clusters
  - Meinshausen and Bühlmann (2010) on stability selection
  - Sun et al. (2013) on stability for tuning parameter selection

Motivation

- In the paper "Stability" (Yu, 2013), Bin Yu wrote *...reproducibility manifests itself in the stability of statistical results relative to reasonable perturbations to data...*
- Stability has been of a great concern in statistics:
  - Breiman (1996) on instability for model selection
  - Bousquet and Elisseeff (2002) derived GE bound via stability
  - Ben-Hur et al. (2002) on stability for structure detection
  - Wang(2010) on stability for selecting number of clusters
  - Meinshausen and Bühlmann (2010) on stability selection
  - Sun et al. (2013) on stability for tuning parameter selection
- There has been little systematic and rigorous theoretical study of stability in the classification context.

Classification Instability (CIS)

- $(X, Y) \sim P$ be a random couple in $\mathbb{R}^d \otimes \{1, 2\}$
- Denote a classifier $\widehat{\phi}_n$ learned from $\mathcal{D} = \{(X_i, Y_i)_{i=1}^n\}$

Classification Instability (CIS)

- $(X, Y) \sim P$ be a random couple in $\mathbb{R}^d \otimes \{1, 2\}$
- Denote a classifier $\widehat{\phi}_n$ learned from $\mathcal{D} = \{(X_i, Y_i)_{i=1}^n\}$

Definition

Define the instability of one classification procedure Ψ *as*

$$CIS(\Psi) = \mathbb{E}_{\mathcal{D}_1, \mathcal{D}_2}\left[\mathbb{P}_X\left(\widehat{\phi}_{n1}(X) \neq \widehat{\phi}_{n2}(X)\right)\right] \quad (1)$$

where $\widehat{\phi}_{n1}$ *and* $\widehat{\phi}_{n2}$ *are classifiers obtained by applying* Ψ *to* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *which are i.i.d. copies of* $\mathcal{D}$.

A classification procedure is reliable if the classifiers trained from multiple homogeneous samples yield similar predictions.

Minimax Upper Bound of CIS for Plug-in Classifiers

- The plug-in classifier first estimates $\eta(x) := \mathbb{P}(Y = 1|X = x)$ and then predicts x as $\widehat{\phi}_n(x) = 1$ iff $\widehat{\eta}_n(x) \geq 1/2$.
- We say distribution P satisfies the *margin condition* if there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for any $\epsilon > 0$,

$$\mathbb{P}(0 < |\eta(X) - 1/2| \leq \epsilon) \leq C_0 \epsilon^{\alpha}.$$

Minimax Upper Bound of CIS for Plug-in Classifiers

- The plug-in classifier first estimates $\eta(x) := \mathbb{P}(Y = 1|X = x)$ and then predicts x as $\widehat{\phi}_n(x) = 1$ iff $\widehat{\eta}_n(x) \geq 1/2$.
- We say distribution P satisfies the *margin condition* if there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for any $\epsilon > 0$,

$$\mathbb{P}(0 < |\eta(X) - 1/2| \leq \epsilon) \leq C_0 \epsilon^{\alpha}.$$

Theorem

(Minimax Upper Bound) Let $\mathcal{P}$ be a set of p.d. on $\mathcal{R} \otimes \{1, 2\}$ satisfying the margin condition and for some sequence $a_n \to \infty$, for any $n \geq 1$, $\delta > 0$, and almost all x w.r.t. marginal dist. of X,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{\mathcal{D}}\Big(|\widehat{\eta}_n(x) - \eta(x)| \geq \delta\Big) \leq C_1 \exp(-C_2 a_n \delta^2) \qquad (2)$$

Then we have: $\sup_{P \in \mathcal{P}} CIS(\Psi) \leq C a_n^{-\alpha/2}$.

Minimax Upper Bound of CIS

- Condition (2) holds for various types of estimators.
  - The local polynomial estimator (Audibert and Tsybakov, 2007) with bandwidth $h = n^{-\frac{1}{2\gamma+d}}$ satisfies it with $a_n = n^{\frac{2\gamma}{2\gamma+d}}$.
  - Our to-be-introduced estimator satisfies it with the same rate.
  - In both cases, the upper bound is $O(n^{-\frac{\alpha\gamma}{2\gamma+d}})$.
- Next we will show this rate is minimax-optimal.

Minimax Lower Bound of CIS

Definition

(Audibert and Tsybakov, 2007) For $\alpha \geq 0$, $\gamma > 0$, denote $\mathcal{P}_{\alpha,\gamma}$ the class of p.d. P on $\mathcal{R} \otimes \{1, 2\}$ s.t.
(i) P satisfies the margin assumption with parameter α;
(ii) $\eta(x)$ belongs to the Holder class with parameter γ;
(iii) the marginal dist. P_X satisfies the strong density assumption.

Minimax Lower Bound of CIS

Definition

(Audibert and Tsybakov, 2007) For $\alpha \geq 0$, $\gamma > 0$, *denote* $\mathcal{P}_{\alpha,\gamma}$ *the class of p.d.* P *on* $\mathcal{R} \otimes \{1, 2\}$ *s.t.*
(i) P *satisfies the margin assumption with parameter* α*;*
(ii) $\eta(x)$ *belongs to the Holder class with parameter* γ*;*
(iii) the marginal dist. P_X *satisfies the strong density assumption.*

Theorem

(Minimax Lower Bound) Let α, γ *be positive constants satisfying* $\alpha\gamma \leq d$*. Assume* $\mathcal{P}_{\alpha,\gamma}$ *satisfies* (2) *with* $a_n = n^{2\gamma/(2\gamma+d)}$*. Then there exists a constant* $C' > 0$ *such that for any* $n \geq 1$*, we have*

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} CIS(\Psi) \geq C' n^{-\alpha\gamma/(2\gamma+d)}.$$

Some Comments of Minimax Rates

- The requirement $\alpha\gamma \leq d$ implies that α and γ can not be large simultaneously. A very large γ implies a very smooth η, while a large α implies that η cannot stay very long near $1/2$, and hence when η hits $1/2$, it should take off quickly.

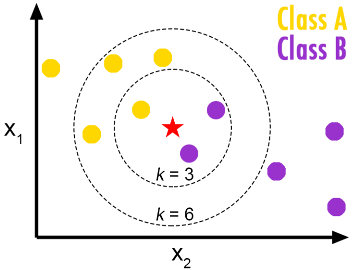
Some Comments of Minimax Rates

- The requirement $\alpha\gamma \leq d$ implies that α and γ can not be large simultaneously. A very large γ implies a very smooth η, while a large α implies that η cannot stay very long near $1/2$, and hence when η hits $1/2$, it should take off quickly.
- When $\alpha\gamma \leq d$, the minimax rate is slower than n^{-1}, and the rate is getting closer to n^{-1} as dimension d increases.
- The optimality of the CIS rate is within the class $\mathcal{P}_{\alpha,\gamma}$.

The to-be-introduced stabilized nearest neighbor classifier can achieve this minimax-optimal rate.

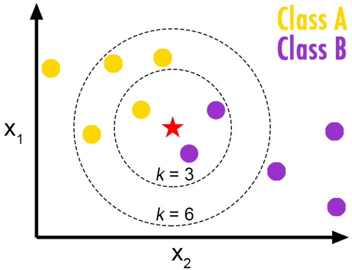
Nearest Neighbor Classifiers

- The knn classifier predicts the class of x to be the most frequent class of its k nearest neighbors.



Nearest Neighbor Classifiers

- The knn classifier predicts the class of x to be the most frequent class of its k nearest neighbors.



- The wnn classifier has weight w_{ni} on the i-th closest neighbor,

$$\widehat{\phi}_n^{\mathbf{w}_n}(x) = 1, \text{iff } \sum_{i=1}^{n} w_{ni} \mathbb{I}_{\{Y_{(i)}=1\}} \geq 1/2.$$

  - When $w_{ni} = \frac{1}{k}\mathbb{I}_{\{1 \leq i \leq k\}}$, wnn reduces to knn.

Asymptotic Expansion of Excess Risk (Regret)

Theorem

(Samworth, 2012) Under regularity assumptions, as $n \to \infty$,

$$\mathit{Regret}(wnn) = \left\{ B_1 \sum_{i=1}^{n} w_{ni}^2 + B_2 \left(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 \right\} \{1 + o(1)\}, \quad (3)$$

where $\alpha_i = i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}}$, B_1 and B_2 are positive constants.

- Minimizing (3) w.r.t. $\mathbf{w}_n$, Samworth (2012) proposed an optimal weighted nearest neighbor classifier (ownn).
- In practice, the ownn classifier is not reliable if its prediction vary much given a small perturbation to the samples.

Asymptotic Equivalent Form of CIS

Theorem

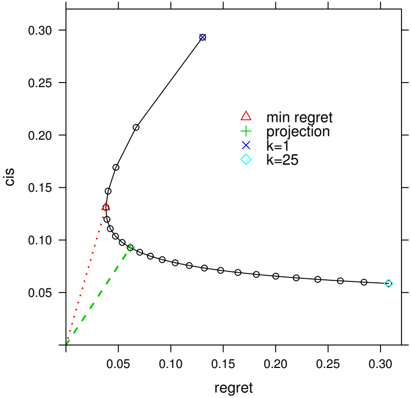
Under the same regularity assumptions, as $n \to \infty$, we have

$$CIS(wnn) = B_3 \Big(\sum_{i=1}^{n} w_{ni}^2 \Big)^{1/2} \{1 + o(1)\}. \tag{4}$$

- The constant $B_3 = 4B_1/\sqrt{\pi}$.
- The CIS of a knn classifier is asymptotically $B_3/\sqrt{k}$.

Regret and CIS of KNN

Figure : *Each dot represents one choice of* $k \in [1, 25]$. *The red triangle obtains minimal regret and the green cross is the projection of the origin to the path.*



Stabilized Nearest Neighbor Classifier

Minimize CIS over the acceptable region where the regret is small:

$$\min_{\mathbf{w}_n} \quad \mathrm{CIS}(\mathrm{wnn})$$

$$\text{s.t.} \quad \mathrm{Regret}(\mathrm{wnn}) \le c_1, \ \sum_{i=1}^{n} w_{ni} = 1, \ \mathbf{w}_n \ge 0.$$

Stabilized Nearest Neighbor Classifier

Minimize CIS over the acceptable region where the regret is small:

$$\min_{\mathbf{w}_n} \quad \mathrm{CIS}(\mathrm{wnn})$$
$$\text{s.t.} \quad \mathrm{Regret}(\mathrm{wnn}) \leq c_1, \ \sum_{i=1}^{n} w_{ni} = 1, \ \mathbf{w}_n \geq 0.$$

By the asymptotic expansions, it is equivalent to

$$\min_{\mathbf{w}_n} \quad \left(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2 \qquad (5)$$
$$\text{s.t.} \quad \sum_{i=1}^{n} w_{ni} = 1; \mathbf{w}_n \geq 0.$$

- The tuning parameter λ controls the balance between regret and CIS.

Stabilized Nearest Neighbor (SNN) Classifier

Theorem

(Optimal Weight) For any fixed $\lambda > 0$, the minimizer of (5) is

$$w_{ni}^* = \begin{cases} \frac{1}{k^*}[1 + \frac{d}{2} - \frac{d}{2(k^*)^{2/d}}\alpha_i], & \text{for } i = 1, \dots, k^*; \\ 0, & \text{for } i = k^* + 1, \dots, n \end{cases}$$

where $\alpha_i = i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}}$ *and* $k^* = \lfloor \{\frac{d(d+4)}{2(d+2)}\}^{\frac{d}{d+4}} \lambda^{\frac{d}{d+4}} n^{\frac{4}{d+4}} \rfloor$.

- We define the weighted nearest neighbor classifier with weight $\mathbf{w}_n^*$ as the snn classifier.
- The snn classifier depends on λ, which can be tuned by CV.

Optimality of the SNN Classifier

We show that the proposed snn classifier achieves minimax-optimal rates in terms of both regret and CIS.

Theorem

(Optimal Rate of SNN) Under the same regularity assumptions, for any $\alpha \geq 0$ and $\gamma \in (0, 2]$, CIS of the proposed snn classifier with any fixed $\lambda > 0$ satisfies

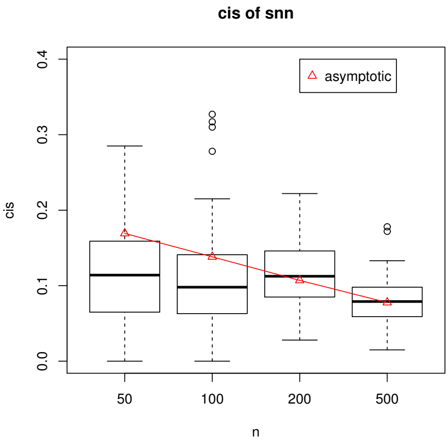
$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} Regret(snn) \leq \tilde{C} n^{-(\alpha+1)\gamma/(2\gamma+d)},$$

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} CIS(snn) \leq C n^{-\alpha\gamma/(2\gamma+d)},$$

for any $n \geq 1$ and some constants $\tilde{C}, C > 0$.

Experiment 1: Validation of Asymptotic Expansion of CIS

- Two classes: $f_1 = N(0_2, \mathbb{I}_2)$ and $f_2 = N(1_2, \mathbb{I}_2)$.

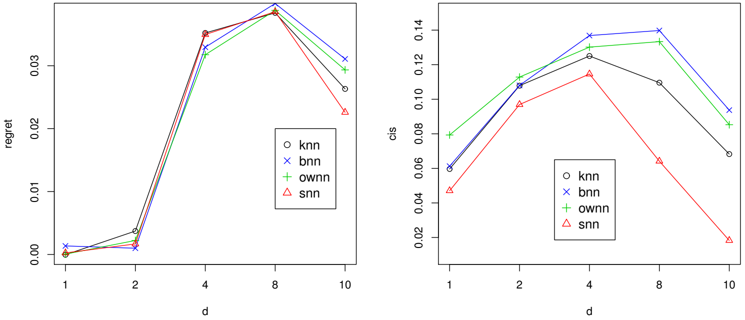


Simulations

- Compare with the knn, the bagged nearest neighbor (bnn) and the ownn classifiers.
- We tune λ in the snn classifier by minimizing $\mathrm{CIS}^2 + \mathrm{Regret}$.
- In each simulations, we fix sample size $n = 200$.
- The average misclassification error and CIS are evaluated on 1000 independently generated test data over 100 replications.

Simulation 1

Two classes are $f_1 = N(0_d, \mathbb{I}_d)$ and $f_2 = N(\mu_d, \mathbb{I}_d)$. We choose μ such that the resulting B_1 is fixed for $d = 1, 2, 4, 8$ and 10.



Slight sacrifice of accuracy may greatly reduce instability.

Simulation 2

$f_1 \sim \frac{1}{2}N(0_d, \mathbb{I}_d) + \frac{1}{2}N(3_d, 2\mathbb{I}_d)$ and $f_2 \sim \frac{1}{2}N(\frac{3}{2}_d, \mathbb{I}_d) + \frac{1}{2}N(\frac{9}{2}_d, 2\mathbb{I}_d)$.
Δ refers to percentage of change of snn compared with ownn.

d	π_0		knn	bnn	ownn	snn	Δ
Sim 2							
2	1/2	Bayes 26.83					
		Error	$30.13_{0.167}$	$29.85_{0.162}$	$\mathbf{29.75}_{0.176}$	$30.14_{0.174}$	1.31%
		CIS	$31.80_{0.973}$	$30.48_{0.873}$	$30.06_{0.833}$	$\mathbf{17.82}_{0.76}$	-40.72%
2	1/3	Bayes 22.76					
		Error	$23.79_{0.111}$	$23.85_{0.131}$	$\mathbf{23.68}_{0.113}$	$23.91_{0.075}$	0.97%
		CIS	$14.93_{0.517}$	$13.99_{0.508}$	$14.99_{0.503}$	$\mathbf{6.90}_{0.394}$	-53.97%
5	1/2	Bayes 11.61					
		Error	$16.50_{0.132}$	$16.00_{0.142}$	$15.91_{0.131}$	$\mathbf{15.51}_{0.118}$	-2.51%
		CIS	$17.02_{0.414}$	$16.19_{0.391}$	$16.15_{0.449}$	$\mathbf{14.43}_{0.332}$	-10.65%
5	1/3	Bayes 10.58					
		Error	$15.14_{0.115}$	$15.00_{0.101}$	$\mathbf{14.88}_{0.102}$	$15.01_{0.110}$	0.87%
		CIS	$11.57_{0.332}$	$12.52_{0.324}$	$11.99_{0.324}$	$\mathbf{10.57}_{0.276}$	-11.84%

Slight sacrifice of accuracy may greatly reduce instability.

Real Examples from UCI Machine Learning Repository

Data	n	d		knn	bnn	ownn	snn	Δ
haberman	306	3						
			Error	$\mathbf{26.08}_{0.281}$	$26.60_{0.268}$	$26.30_{0.275}$	$26.56_{0.260}$	0.99%
			CIS	$5.39_{0.485}$	$6.03_{0.526}$	$5.25_{0.476}$	$\mathbf{3.92}_{0.450}$	-25.33%
liver	345	6						
			Error	$38.76_{0.356}$	$38.61_{0.488}$	$\mathbf{37.50}_{0.360}$	$38.27_{0.399}$	2.05%
			CIS	$37.95_{1.472}$	$39.86_{1.322}$	$39.38_{1.384}$	$\mathbf{33.20}_{1.731}$	-15.69%
appendicitis	106	7						
			Error	$15.36_{0.477}$	$17.91_{0.786}$	$15.92_{0.533}$	$\mathbf{15.19}_{0.493}$	-4.59%
			CIS	$10.43_{0.686}$	$18.43_{1.250}$	$14.36_{0.918}$	$\mathbf{9.38}_{0.709}$	-34.68%
pima	768	8						
			Error	$26.08_{0.212}$	$25.92_{0.198}$	$\mathbf{25.83}_{0.192}$	$26.04_{0.205}$	0.81%
			CIS	$13.95_{0.431}$	$14.36_{0.465}$	$14.11_{0.462}$	$\mathbf{12.64}_{0.405}$	-10.42%
stalog	270	13						
			Error	$17.44_{0.236}$	$17.64_{0.297}$	$17.37_{0.245}$	$\mathbf{16.97}_{0.238}$	-2.30%
			CIS	$13.39_{0.821}$	$12.72_{0.678}$	$11.94_{0.614}$	$\mathbf{11.28}_{0.477}$	-5.53%
credit	690	14						
			Error	$14.55_{0.144}$	$14.63_{0.144}$	$14.60_{0.146}$	$\mathbf{14.54}_{0.144}$	-0.41%
			CIS	$7.52_{0.256}$	$6.85_{0.271}$	$6.77_{0.267}$	$\mathbf{6.41}_{0.253}$	-5.32%
spect	267	22						
			Error	$20.66_{0.330}$	$20.41_{0.402}$	$20.34_{0.310}$	$\mathbf{20.25}_{0.298}$	-0.44%
			CIS	$11.06_{1.114}$	$12.90_{1.228}$	$11.09_{1.013}$	$\mathbf{6.86}_{0.987}$	-38.14%

Slight sacrifice of accuracy may greatly reduce instability.

Take home messages

- We introduced a general measure of classification instability CIS and established its minimax rate for general plug-in classifiers.
- We proposed a novel stabilized nearest neighbor classifier to achieve this optimal rate.

Acknowledgement

Guang Cheng (Purdue)

Xingye Qiao (Binghamton)