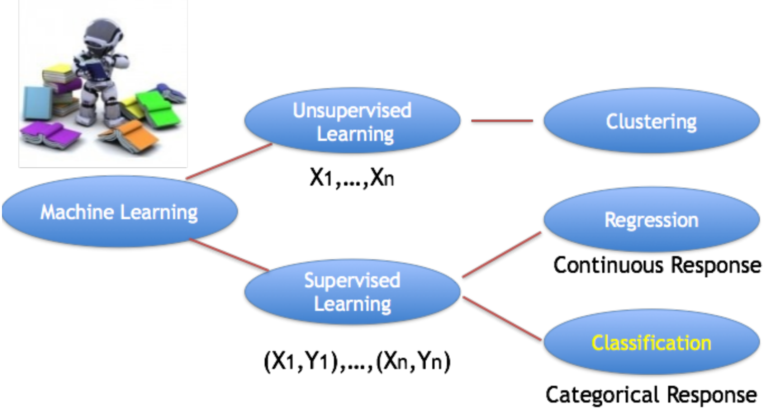
Stability of Machine Learning Algorithms

Wei Sun

Department of Statistics
Purdue University

April, 8, 2015

Machine Learning



Machine Learning

Unsupervised Learning

$X_1,\dots,X_n$

Clustering

Supervised Learning

$(X_1,Y_1),\dots,(X_n,Y_n)$

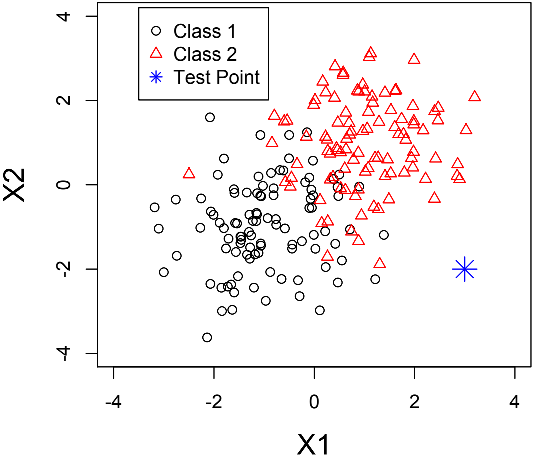
Regression

Continuous Response

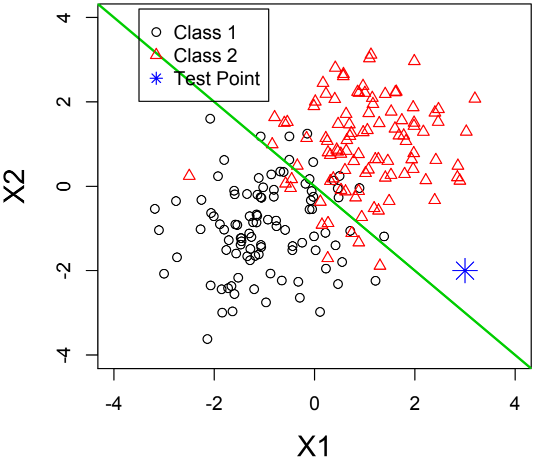
Classification

Categorical Response

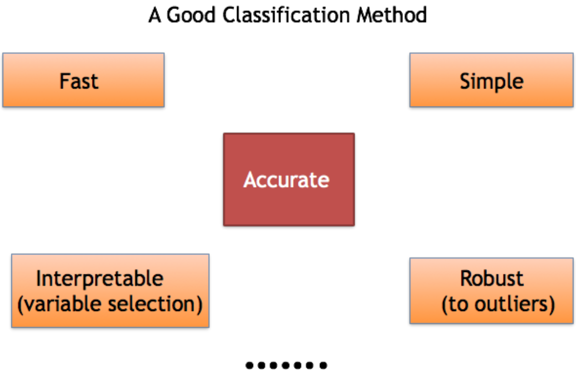
Classification



Classification

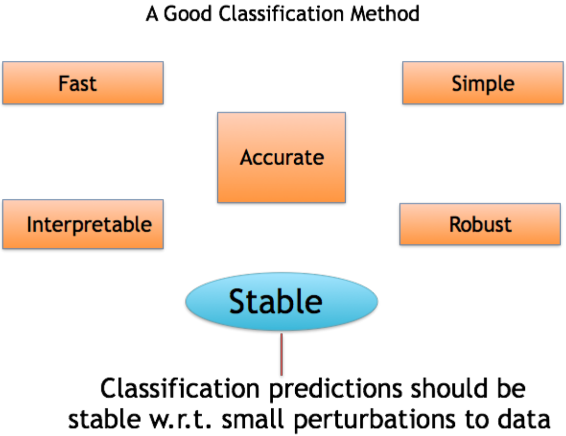


Evaluation Metrics


A Good Classification Method
Fast
Simple
Accurate
Interpretable (variable selection)
Robust (to outliers)
....

- See Zhang and Lin (2013).

Stability



A Good Classification Method

Fast

Simple

Accurate

Interpretable

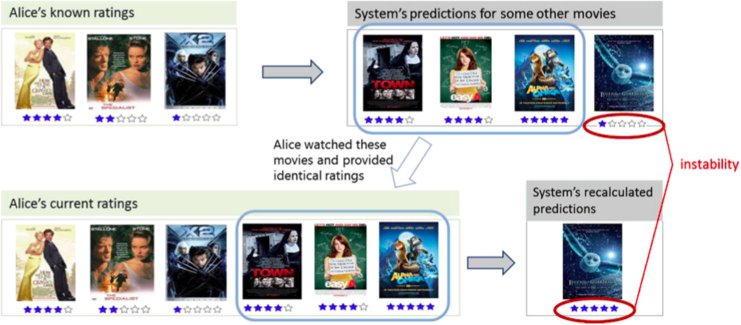
Robust

Stable

Classification predictions should be stable w.r.t. small perturbations to data

Motivation: Users' Trust

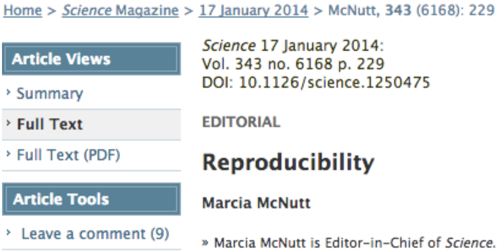
An unstable algorithm can potentially decrease users' trust in it.


Alice's known ratings
★★★★☆
★★☆☆☆
★☆☆☆☆
System's predictions for some other movies
★★★★☆
★★★★☆
★★★★★
★☆☆☆☆
Alice watched these movies and provided identical ratings
instability
Alice's current ratings
★★★★☆
★★☆☆☆
★☆☆☆☆
★★★★☆
★★★★☆
★★★★★
System's recalculated predictions
★★★★★

Source: Adomavicius and Zhang (2012), ACM Transactions on Information Systems.

Motivation: Reproducibility

- Begley and Ellis (Nature, 2012): 47/53 cancer research papers were irreproducible.
- Editor of *Science* emphasized importance of reproducibility.



Home > *Science* Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

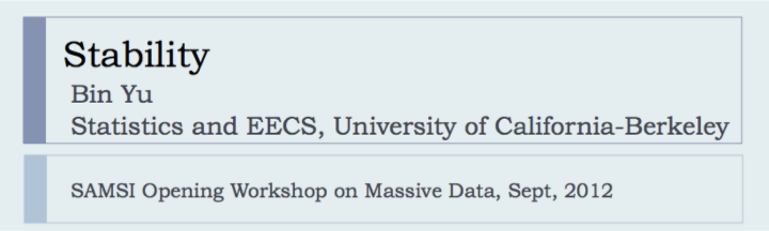
EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

Motivation: Reproducibility



Stability

Bin Yu

Statistics and EECS, University of California-Berkeley

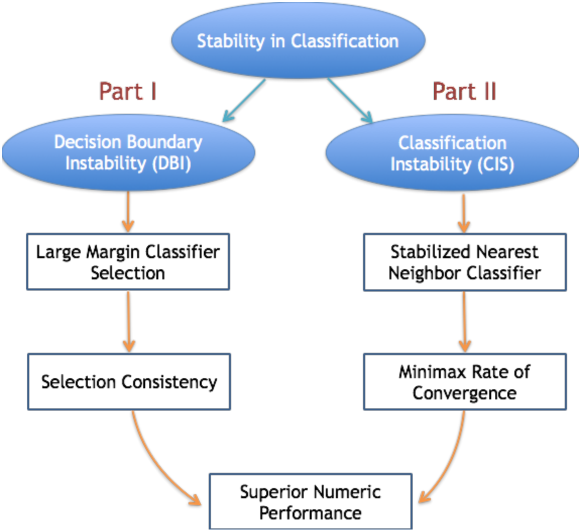
SAMSI Opening Workshop on Massive Data, Sept, 2012

- Prof. Bin Yu (Stability, Bernoulli, 2013) wrote: *reproducibility manifests itself in the stability of statistical results relative to reasonable perturbations to data...*

Stability in Statistics

- Breiman (1996) introduced stability for model selection.
- Meinshausen and Bühlmann (2010) on *stability selection*.
- Wang (2010) and Sun et al. (2012) on *clustering stability*.
- Sun et al. (2013) on *variable selection stability*.
- There has been little systematic and rigorous study of stability in classification.

Outline



Stability in Classification

Part I

Part II

Decision Boundary Instability (DBI)

Classification Instability (CIS)

Large Margin Classifier Selection

Stabilized Nearest Neighbor Classifier

Selection Consistency

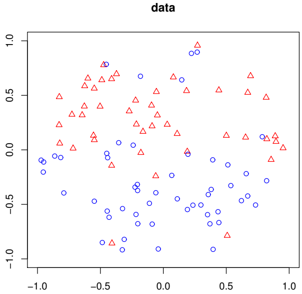
Minimax Rate of Convergence

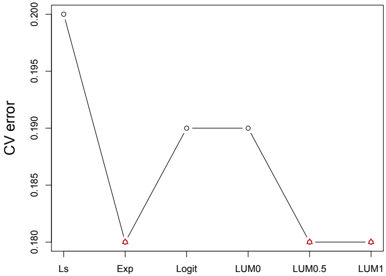
Superior Numeric Performance

Part I: DBI

Decision Boundary Instability for Classifier Selection
— Introduce statistical inference to machine learning community

Motivation

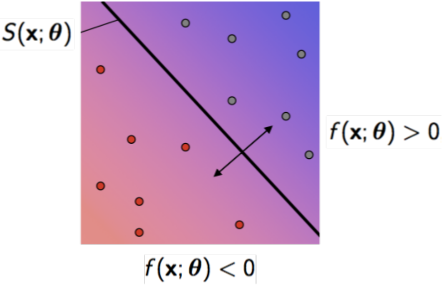




- Question 1: Is the difference of errors significant?
    - Statistical significant testing.
- Question 2: Which criterion is suitable for comparison?
    - Prediction accuracy plus stability.

Background

- Training data $\mathcal{D}_n = \{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$
  - $\mathbf{x}_i \in R^d$: input
  - $y_i \in \{1, -1\}$: binary response
- Decision function $f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{w}^T \mathbf{x} + b$ with $\boldsymbol{\theta} = (b, \mathbf{w}^T)^T$
- Decision boundary $S(\mathbf{x}; \boldsymbol{\theta}) = \{\mathbf{x} : f(\mathbf{x}; \boldsymbol{\theta}) = 0\}$

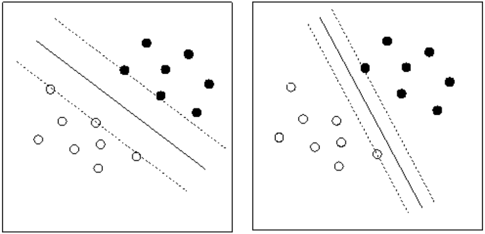

$S(\mathbf{x}; \boldsymbol{\theta})$
$f(\mathbf{x}; \boldsymbol{\theta}) > 0$
$f(\mathbf{x}; \boldsymbol{\theta}) < 0$

Large-margin Classifiers

- Large-margin classifiers solve

$$\widehat{\boldsymbol{\theta}}_L := \arg\min_{b,\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} L\Big(y_i(\mathbf{w}^T\mathbf{x}_i + b)\Big) + \frac{\lambda_n}{2}\mathbf{w}^T\mathbf{w},$$

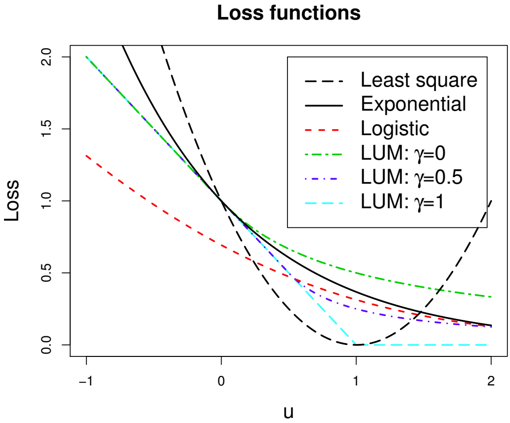
where $L(\cdot)$ is the loss function.



(a) Larger margin (b) Smaller margin

Common Loss Functions $L(u)$

- Least square, Exponential, Logistic
- Large-margin unified machines (LUM; Liu, 2011): $\gamma \in [0, 1]$
  - $\gamma = 0.5$: distance weighted discriminant (Marron et al., 2007)
  - $\gamma = 1$: SVM.

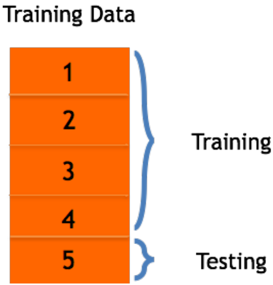


Which Loss to Use?

- Existing selection criterion is Generalization Error (GE).
- Given training data and a test sample $(\mathbf{X}_0, Y_0)$, GE of loss L is

$$D_{0L} = \mathbb{E}|Y_0 - \mathrm{sign}\{f(\mathbf{X}_0; \widehat{\boldsymbol{\theta}}_L)\}|.$$

- Estimate GE by 5-fold CV error, $\widehat{\mathcal{D}}_L = \frac{1}{5}\sum_{k=1}^{5} \widehat{D}(\widehat{\boldsymbol{\theta}}_{L(-k)})$.


Training Data

1

2

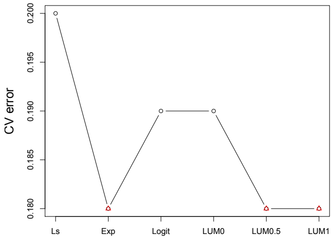
3

4

5

Training

Testing

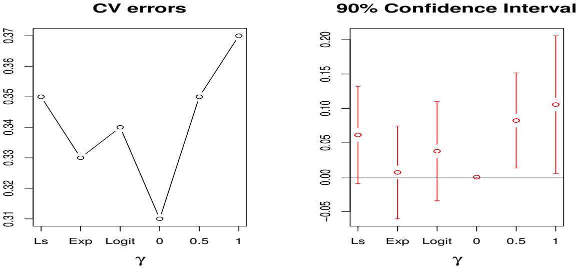


Our Proposal

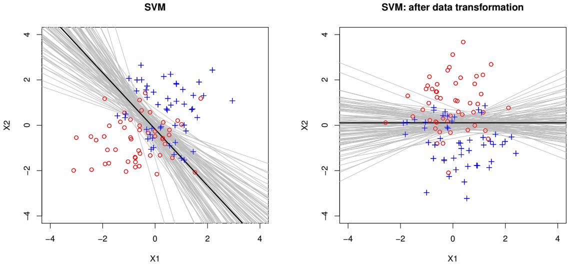
- Select the *most accurate and stable* classifier.
- Stage 1: exclude classifiers with significantly large CV errors. This is done by statistical testing.
- Stage 2: choose the optimal classifier according to minimal decision boundary instability (DBI).

Stage 1: Statistical Testing

- Goal: Make statistical inference on $\Delta_{L_1,L_2} = D_{0L_2} - D_{0L_1}$.
- Estimator $\widehat{\Delta}_{L_1,L_2} = \widehat{\mathcal{D}}_{L_2} - \widehat{\mathcal{D}}_{L_1}$
- We show $n^{1/2}\{\widehat{\Delta}_{L_1,L_2} - \Delta_{L_1,L_2}\}$ is asymptotically normal.
- Mimic its dist. via perturbation-resampling (Jiang et al., 2008)
- Construct the confidence interval for Δ_{L_1,L_2}



Stage 2: Decision Boundary Instability (DBI)



Definition

The DBI of decision boundary $S(\mathbf{x}; \widehat{\theta}_L)$ *is defined to be*

$$DBI\left(S(\mathbf{X}; \widehat{\theta}_L)\right) = E\left[Var\left(S_d | \mathbf{X}^{\dagger}_{(-d)}\right)\right],$$

where S_d *is the* d*th dimension of transformed decision boundary and* $\mathbf{X}^{\dagger}_{(-d)}$ *is the first* $d-1$ *axes after transformation.*

Relationship of DBI with Other Variability Measures

- Variance of the CV error:
  - Different focus.
  - separable: small variation in CV error; large variation in DB
- Stability in Bousquet and Elisseeff (2002): maximal difference of decision functions from original and leave-one-out data.
  - Their stability suffers from the transformation variant issue.
  - Our DBI is transformation invariant.
- In experiments, DBI leads to superior numerical performance.

Selection consistency for LUM L_γ

- Selected classifier index $\widehat{\gamma}_0$; True optimal classifier index γ_0.
- We show that the selected optimal classifier has achieved the minimal GE and minimal DBI asymptotically.

Theorem

Under regularity assumptions,

$$\left|\widehat{\mathcal{D}}_{\widehat{\gamma}_0} - D_{0\gamma_0}\right| = O_P(n^{-1/2}).$$

Theorem

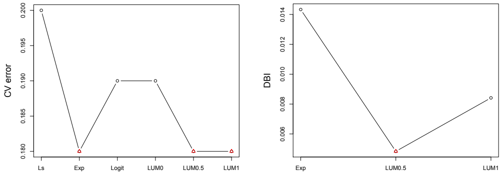
Under regularity assumptions, as $N \to \infty$,

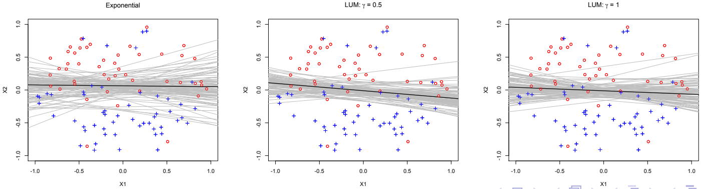
$$\left|\widehat{DBI}\Big(S(\mathbf{X}; \widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})\Big) - DBI\Big(S(\mathbf{X}; \widehat{\boldsymbol{\theta}}_{\gamma_0})\Big)\right| = o_P(n^{-1}).$$

Recall that N is the number of perturbation resamplings.

Simulation: Illustration

- Generate two predictors uniformly over $\{x_1^2 + x_2^2 \leq 1\}$.
- $y = 1$ when $x_2 \geq 0$ and -1 otherwise.
- Generate 100 samples and contaminate data by randomly flipping the labels of 15% (Sim 1) or 25% (Sim 2) samples.





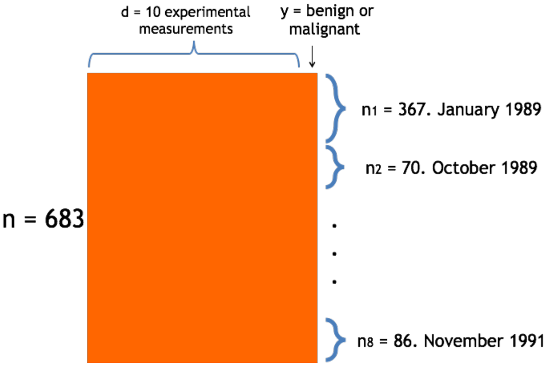
Simulation: Results

- Compare three methods:
  - cv+varcv: minimal variance of the K-CV error in Stage 2
  - cv+be: minimal stability of Bousquet and Elisseeff in Stage 2
  - cv+dbi: our method
- Test size 1000.
- Compare test error and test DBI over 100 replications.

Simulations		cv+varcv	cv+be	cv+dbi
Sim 1	Error	$0.191_{0.002}$	$0.194_{0.002}$	$\mathbf{0.190}_{0.002}$
	DBI	$0.139_{0.043}$	$0.135_{0.019}$	$\mathbf{0.081}_{0.002}$
Sim 2	Error	$0.296_{0.002}$	$0.303_{0.003}$	$\mathbf{0.295}_{0.002}$
	DBI	$0.291_{0.044}$	$0.318_{0.036}$	$\mathbf{0.229}_{0.012}$

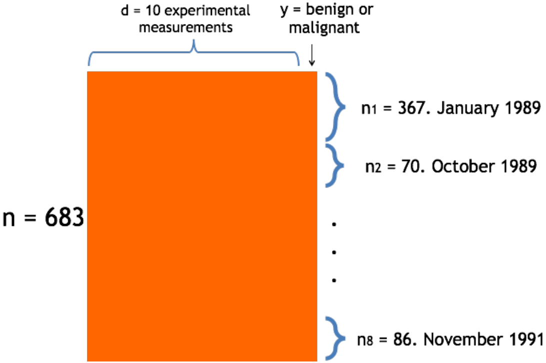
Real Example

- Breast cancer data set (Wolberg and Mangasarian, 1992)



Real Example

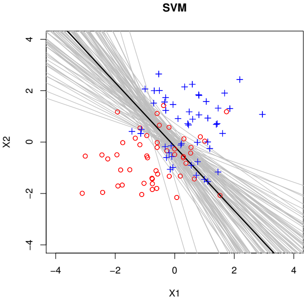
- Breast cancer data set (Wolberg and Mangasarian, 1992)



	cv+varcv	cv+be	cv+dbi
Error	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$
DBI	$0.388_{0.066}$	$0.152_{0.028}$	$\mathbf{0.124}_{0.023}$

Summary of Part I

- New concept of decision boundary instability (DBI).
- Incorporate it to select the *most accurate and stable* classifier

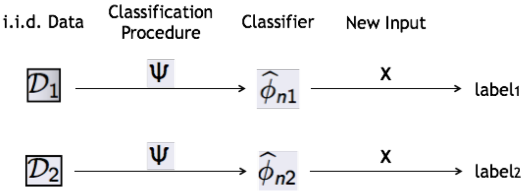


- DBI has two limitations:
  - Only apply to linear classifiers
  - Estimation of DBI is complicated

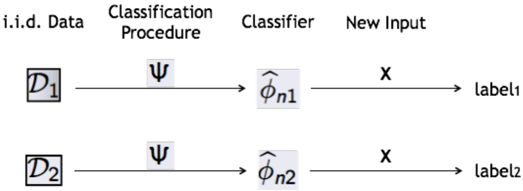
Part II

Classification Instability (CIS) and
Stabilized Nearest Neighbor Classifier
— Push the frontier of statistical theory in machine learning

Classification Instability (CIS)



Classification Instability (CIS)



Definition

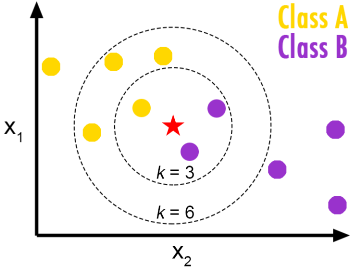
Define instability of a classification procedure Ψ *as*

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

A classification procedure is reliable if the classifiers trained from two homogeneous training data sets yield similar predictions.

Nearest Neighbor Classifiers

- knn classifier



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

Regret of WNN

Regret = Classification error - Bayes risk

Theorem

(Samworth, 2012) Under regularity assumptions,

$$Regret(wnn) = B_1 \sum_{i=1}^{n} w_{ni}^2 + B_2 \Big(n^{-\frac{2}{d}} \sum_{i=1}^{n} \alpha_i w_{ni} \Big)^2, \qquad (2)$$

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

- Minimizing (2) w.r.t. w_{ni} leads to OWNN.
- OWNN is not reliable if its prediction vary much given a small perturbation to the samples.

CIS of WNN

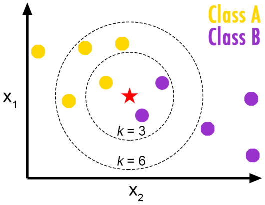
Theorem

Under regularity assumptions, we show that

$$CIS(wnn) = B_3 \left(\sum_{i=1}^{n} w_{ni}^2 \right)^{1/2},$$

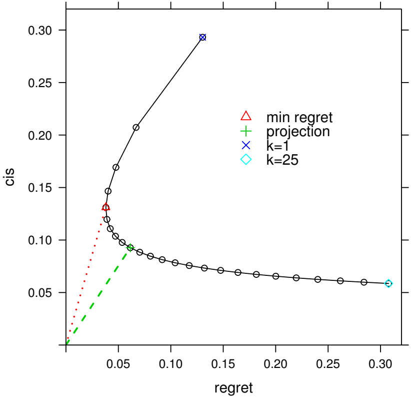
where $B_3 = 4B_1/\sqrt{\pi}$ is a positive constant.

- The CIS of a knn classifier is $B_3/\sqrt{k}$ asymptotically.



Regret and CIS of KNN

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



Stabilized Nearest Neighbor (SNN) Classifier

- Minimize CIS over the region where 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

$$\min_{\mathbf{w}_n} \quad \left(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2 \qquad (3)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_{ni} = 1; \mathbf{w}_n \geq 0.$$

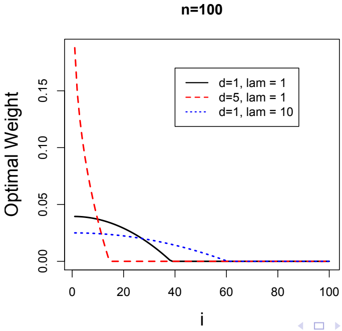
Stabilized Nearest Neighbor (SNN) Classifier

- The optimal weight of (3) is

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

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

- We define the WNN with weight $\mathbf{w}_n^*$ as SNN.



Theoretical Properties

- A sharp convergence rate of CIS for a general plug-in classification procedure.
- As a special plug-in procedure, the proposed SNN achieves minimax optimal convergence rate in regret (Audibert and Tsybakov, 2007) and the sharp convergence rate in CIS.
- Asymptotic comparisons among kNN, BNN, OWNN and SNN.

Upper Bound of CIS

- We focus on the plug-in classifier: it estimates $\eta(x) := \mathbb{P}(Y = 1|X = x)$ and then predicts x as

$$\widehat{\phi}_n(x) = \begin{cases} 1, & \text{if } \widehat{\eta}_n(x) \geq 1/2; \\ 0, & \text{otherwise} \end{cases}$$

- We say distribution P satisfies the *margin condition* if there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for any $\epsilon > 0$,

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

The larger α, the easier the classification problem

Upper Bound of CIS

Theorem

(Upper Bound) Let $\mathcal{P}$ be a set of p.d. on $\mathcal{R} \times \{1, 2\}$ 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,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{\mathcal{D}}\Big(|\widehat{\eta}_n(x) - \eta(x)| \geq \delta\Big) \leq C_1 \exp(-C_2 a_n \delta^2). \qquad (4)$$

Then CIS of the plug-in procedure Ψ corresponding to $\widehat{\eta}_n$ satisfies

$$\sup_{P \in \mathcal{P}} CIS(\Psi) \leq C a_n^{-\alpha/2},$$

for any $n \geq 1$ and some constant C.

Upper Bound of CIS

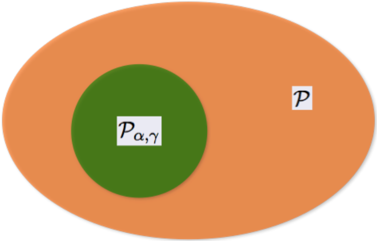
$$\sup_{P\in\mathcal{P}} \mathbb{P}_{\mathcal{D}}\Big(|\widehat{\eta}_n(x)-\eta(x)|\geq\delta\Big)\leq C_1\exp(-C_2a_n\delta^2).$$

- Above condition 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 SNN 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 can not be improved.

Lower Bound of CIS

Definition

For $\alpha \geq 0$, $\gamma > 0$, let $\mathcal{P}_{\alpha,\gamma}$ be the class of p.d. P on $\mathcal{R} \times \{1,2\}$ 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 distribution P_X satisfies the strong density assumption.



Lower Bound of CIS

Theorem

(Lower Bound) Let α, γ be positive constants satisfying $\alpha\gamma \le d$. Assume $\mathcal{P}_{\alpha,\gamma}$ satisfies (4) with $a_n = n^{2\gamma/(2\gamma+d)}$. Then there exists a constant $C' > 0$ such that for any $n \ge 1$, we have

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

- $n^{-\alpha\gamma/(2\gamma+d)}$ is a sharp convergence rate of CIS for general plug-in classification procedure.
- When $\alpha\gamma = d$, the rate is approaching n^{-1} as d increases.
- Our SNN can achieve this sharp convergence rate in CIS.

Optimality of SNN

The proposed SNN procedure achieves minimax optimal convergence rate in regret and the sharp convergence rate in CIS.

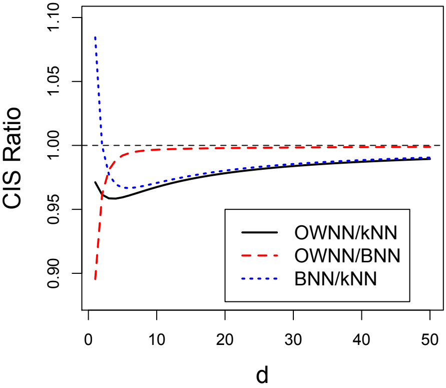
Theorem

Under regularity assumptions, our SNN procedure with any fixed $\lambda > 0$ satisfies, for any $n \geq 1$,

$$\begin{aligned} \sup_{P \in \mathcal{P}_{\alpha,\gamma}} Regret(SNN) &\leq \tilde{C} n^{-(\alpha+1)\gamma/(2\gamma+d)}, \\ \sup_{P \in \mathcal{P}_{\alpha,\gamma}} CIS(SNN) &\leq C n^{-\alpha\gamma/(2\gamma+d)}. \end{aligned}$$

Asymptotic Comparisons of kNN, BNN, and OWNN

- *k* nearest neighbor (*k*NN, Cover and Hart, 1967)
- Bagged nearest neighbor (BNN, Hall and Samworth, 2005)
- OWNN (Samworth, 2012)



Asymptotic Comparisons OWNN and SNN

Corollary

Under regularity conditions, we have, as $n \to \infty$,

$$\frac{Regret(SNN)}{Regret(OWNN)} \longrightarrow \left\{\frac{B_1}{\lambda B_2}\right\}^{d/(d+4)}\left\{\frac{4+d\lambda B_2/B_1}{4+d}\right\},$$

$$\frac{CIS(SNN)}{CIS(OWNN)} \longrightarrow \left\{\frac{B_1}{\lambda B_2}\right\}^{d/(2(d+4))},$$

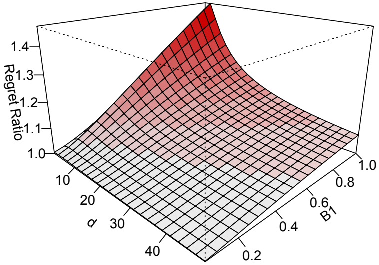
where B_1 and B_2 are constants.

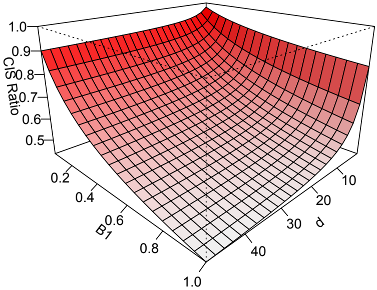
- When we fix $\lambda = (B_1 + B_3^2)/B_2$, we have

$$\frac{\mathrm{Regret(SNN)}}{\mathrm{Regret(OWNN)}} \longrightarrow \left\{\frac{1}{1+16B_1/\pi}\right\}^{d/(d+4)}\left\{\frac{4+d(1+16B_1/\pi)}{4+d}\right\},$$

$$\frac{\mathrm{CIS(SNN)}}{\mathrm{CIS(OWNN)}} \longrightarrow \left\{\frac{1}{1+16B_1/\pi}\right\}^{d/(2(d+4))}.$$

Asymptotic Comparisons OWNN and SNN





Asymptotic Comparisons OWNN and SNN

- Define $\text{Relative Gain} = |\Delta\text{CIS}/\Delta\text{Regret}|$
- $\Delta\text{CIS} = [\text{CIS(snn)} - \text{CIS(ownn)}]/\text{CIS(ownn)}$
- $\Delta\text{Regret} = [\text{Regret(snn)} - \text{Regret(ownn)}]/\text{Regret(ownn)}$

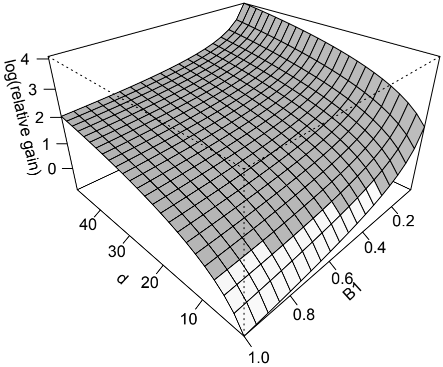


Figure : Grey color: value > 0; white color: value < 0.

Simulation: Illustration

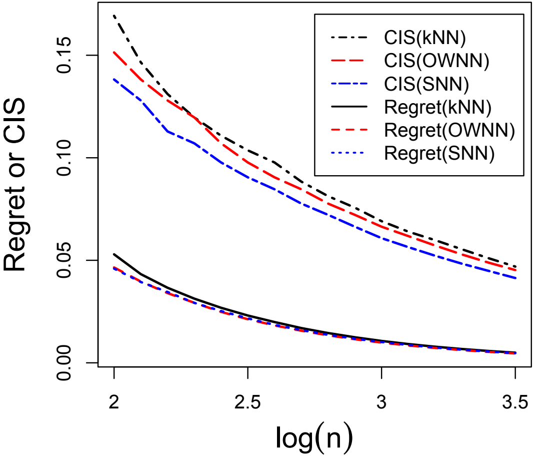
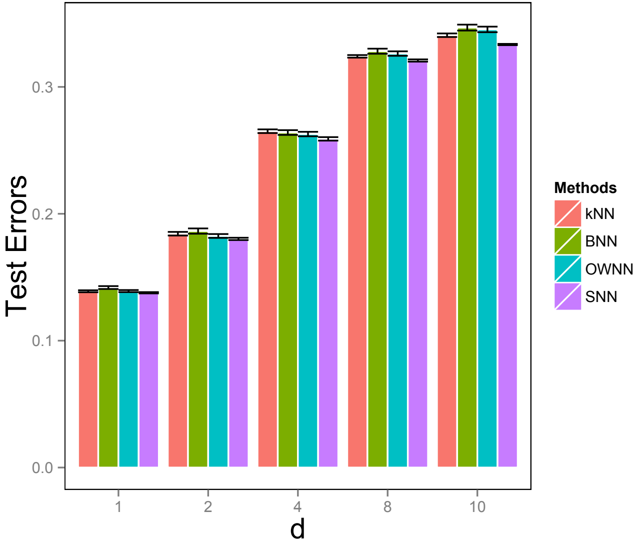


Figure : Regret and CIS of *k*NN, OWNN, SNN in a bivariate normal example.

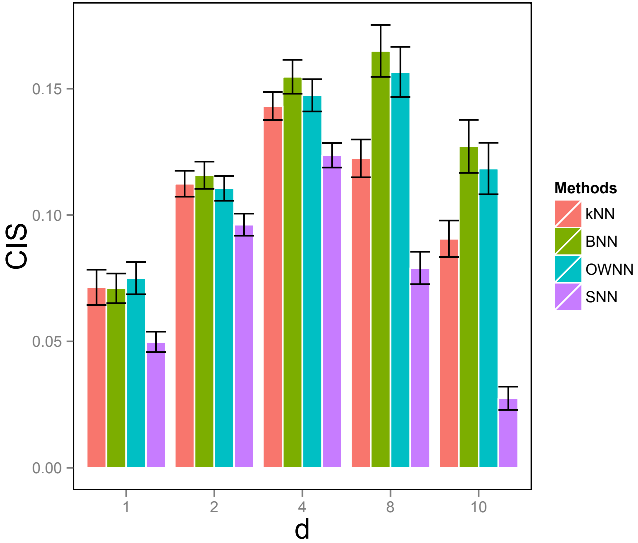
Simulation

- Two classes are $f_1 = N(0_d, \mathbb{I}_d)$ and $f_2 = N(\mu_d, \mathbb{I}_d)$.
- Choose μ s.t. $B_1 = 0.1$ for $d = 1, 2, 4, 8$ and 10.
- Training size $n = 200$.
- Test size 1000.
- Test error and test CIS are evaluated over 100 replications.
- Compare kNN, BNN, OWNN, and our SNN

Simulation: Test Error

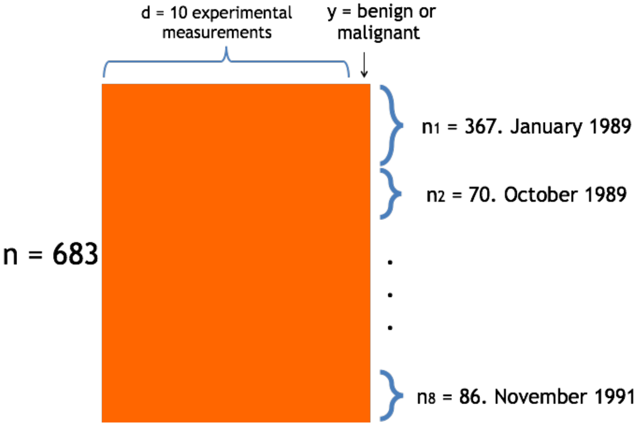


Simulation: Test CIS



Real Examples

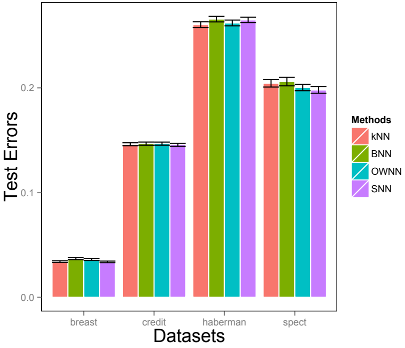
- Breast cancer data set (Wolberg and Mangasarian, 1992).

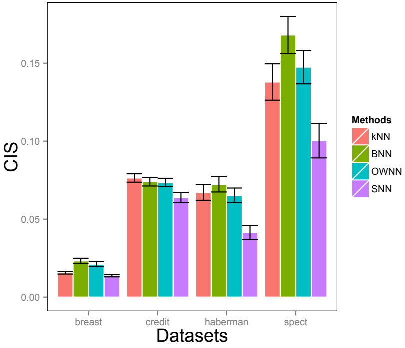


Real Examples: Continue

- Credit approval data set (*credit*): $n = 690$ credit card applications and $d = 14$ user attributes.
- Haberman's survival data set (*haberman*): $n = 306$ cases and $d = 3$ patient attributes.
- SPECT heart data set (*spect*): diagnosing of cardiac SPECT images. $n = 267$ images with $d = 22$ binary features.
- All data are available on UCI Machine Learning Website.

Real Examples: Results



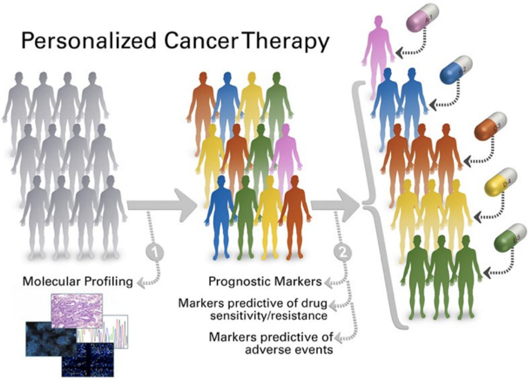


Summary of Part II

- Classification Instability (CIS)
- Sharp rate of CIS for general plug-in classifiers
- SNN: minimax optimal rate in regret and sharp rate in CIS.
- Superior numeric performance in both accuracy and stability.

Future Work on Stability

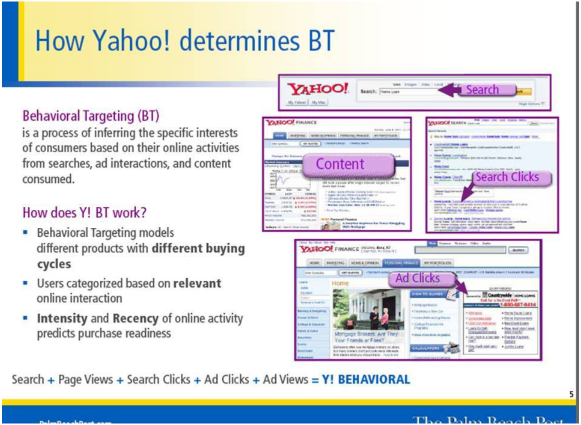
- Personalized medicine via outcome weighted learning (JASA, Zhao et al., 2012)


Personalized Cancer Therapy
1
2
Molecular Profiling
Prognostic Markers
Markers predictive of drug sensitivity/resistance
Markers predictive of adverse events
A1
B2
C3
D4
E5

- Incorporate CIS into outcome weighted learning.

Future Work on Stability

- Behavioral targeting: deliver consistent and relevant ads!



How Yahoo! determines BT

Behavioral Targeting (BT)
is a process of inferring the specific interests of consumers based on their online activities from searches, ad interactions, and content consumed.

How does Y! BT work?

- Behavioral Targeting models different products with **different buying cycles**
- Users categorized based on **relevant** online interaction
- **Intensity** and **Recency** of online activity predicts purchase readiness

Search + Page Views + Search Clicks + Ad Clicks + Ad Views = **Y! BEHAVIORAL**

5

PalmBeachPost.com

The Palm Beach Post

Wei Sun
Department of Statistics
Purdue University
sun244@purdue.edu

Backup

Backup slides start from here!

Assumptions for Asymptotic Normality of K-CV Error

(L1) The probability distribution function of $\mathbf{X}$ and the conditional probability $p(\mathbf{x})$ are both continuously differentiable.

(L2) The parameter $\boldsymbol{\theta}_{0L}$ is bounded and unique.

(L3) The map $\boldsymbol{\theta} \mapsto L(yf(\mathbf{x}; \boldsymbol{\theta}))$ is convex.

(L4) The map $\boldsymbol{\theta} \mapsto L(yf(\mathbf{x}; \boldsymbol{\theta}))$ is differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_{0L}$ a.s.. Furthermore, $G(\boldsymbol{\theta}_{0L})$ is element-wisely bounded, where $G(\boldsymbol{\theta}_{0L}) = E\left[\nabla_{\boldsymbol{\theta}} L(Yf(\mathbf{X}; \boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}} L(Yf(\mathbf{X}; \boldsymbol{\theta}))^T\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}$.

(L5) The surrogate risk $\mathcal{R}_L(\boldsymbol{\theta})$ is bounded and twice differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_{0L}$ with the positive definite Hessian matrix $H(\boldsymbol{\theta}_{0L}) = \nabla^2_{\boldsymbol{\theta}} \mathcal{R}_L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}$.

- Assumption (L1) ensures that the GE is continuously differentiable with respect to $\boldsymbol{\theta}$. Assumption (L3) ensures that the uniform convergence theorem for convex functions can be applied. Assumptions (L4) and (L5) are required to obtain the local quadratic approximation to the surrogate risk function around $\boldsymbol{\theta}_{0L}$.

Asymptotic Normality of K-CV Error

Theorem

Suppose Assumptions (L1)–(L5) hold and $\lambda_n = o(n^{-1/2})$. Then for any fixed K,

$$\mathcal{W}_L = \sqrt{n}\Big(\widehat{\mathcal{D}}_L - D_{0L}\Big) \xrightarrow{d} N\Big(0, E(\psi_1^2)\Big) \quad \text{as } n \to \infty.$$

Theorem

Suppose the above assumptions hold, we have, as $n \to \infty$,

$$\mathcal{W}_{\Delta_{12}} = n^{1/2}(\widehat{\Delta}_{12} - \Delta_{12}) \xrightarrow{d} N\Big(0, Var(\psi_{12} - \psi_{11})\Big).$$

DBI and Its Estimator

- DBI is approximately

$$DBI\left(S(\mathbf{X};\widehat{\boldsymbol{\theta}}_L)\right) \approx (w^{\dagger}_{L,d})^{-2} E\left[\tilde{\mathbf{X}}^{\dagger T}_{(-d)} \left(n^{-1}\Sigma^{\dagger}_{0L,(-d)}\right) \tilde{\mathbf{X}}^{\dagger}_{(-d)}\right],$$

  where $w^{\dagger}_{L,d}$ is the last entry of the transformed coefficient $\boldsymbol{\theta}^{\dagger}_{0L}$, and $n^{-1}\Sigma^{\dagger}_{0L,(-d)}$ is the asymptotic variance of the first d dimensions of $\widehat{\boldsymbol{\theta}}^{\dagger}_{L}$.

- We propose the following DBI estimate:

$$\widehat{DBI}\left(S(\mathbf{X};\widehat{\boldsymbol{\theta}}_L)\right) = \frac{\sum_{i=1}^{n} \widetilde{\mathbf{x}}_{i(-d)}^{\dagger T} \widehat{\Sigma}^{\dagger}_{L,(-d)} \widetilde{\mathbf{x}}_{i(-d)}^{\dagger}}{(n\widehat{w}^{\dagger}_{L,d})^2},$$

  where $\widehat{w}^{\dagger}_{L,d}$ is the last entry of $\widehat{\boldsymbol{\theta}}^{\dagger}_{L}$, and $\widehat{\Sigma}^{\dagger}_{L,(-d)}$ is obtained by removing the last row and last column of $\widehat{\Sigma}^{\dagger}_{L}$.

Selection consistency

- Focus on LUM: L_γ
- $\gamma_0^* = \arg\min_{\gamma\in[0,1]} D_{0\gamma}$
- $\widehat{\gamma}_0^* = \arg\min_{\gamma\in[0,1]} \widehat{\mathcal{D}}_\gamma$.
- Λ_0: the population set of potentially good classifiers
- $\widehat{\Lambda}_0$: the empirical set of potentially good classifiers

$$\begin{aligned} \Lambda_0 &= \left\{\gamma \in [0,1] : D_{0\gamma} = D_{0\gamma_0^*}\right\} \\ \widehat{\Lambda}_0 &= \left\{\gamma \in [0,1] : \widehat{\mathcal{D}}_\gamma \leq \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} + n^{-1/2}\phi_{\gamma,\widehat{\gamma}_0^*;\alpha/2}\right\} \end{aligned}$$

- The true optimal index:

$$\gamma_0 = \arg\min_{\gamma\in\Lambda_0} DBI\Big(S(\mathbf{X};\widehat{\boldsymbol{\theta}}_\gamma)\Big)$$

- The empirical optimal index:

$$\widehat{\gamma}_0 = \arg\min_{\gamma\in\widehat{\Lambda}_0} \widehat{DBI}\Big(S(\mathbf{X};\widehat{\boldsymbol{\theta}}_\gamma)\Big)$$

Assumptions of Selection consistency

(L1) The probability distribution function of $\mathbf{X}$ and the conditional probability $p(\mathbf{x})$ are both continuously differentiable.

(A1) $\mathrm{Var}(\mathbf{X}|Y) \in \mathbb{R}^{d\times d}$ is a positive definite matrix for $Y \in \{1, -1\}$. This guarantees the uniqueness of the true minimizer $\boldsymbol{\theta}_{0\gamma}$.

(B1) The smallest eigenvalue of the true Hessian matrix $\lambda_{\min}(H(\boldsymbol{\theta}_{0\gamma})) \geq c_1$, and the largest eigenvalue of the true Hessian matrix $\lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma})) \leq c_2$, where the positive constants c_1, c_2 do not depend on γ.

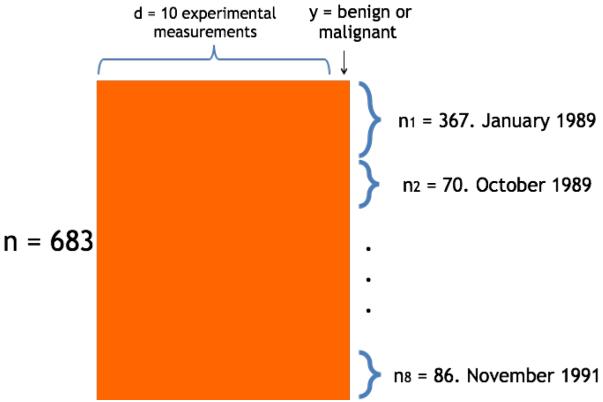
More Simulations of DBI Project

- **Simulation 3**: The setting was the same as Simulation 1 except that we contaminated the data by randomly flipping the labels of 80% of the instances whose $|x_2| \geq 0.7$.
- **Simulation 4**: Two predictors were uniformly generated over $\{(x_1, x_2) : |x_1| + |x_2| \leq 2\}$. Conditionally on $X_1 = x_1$ and $X_2 = x_2$, the class label y took 1 with probability $e^{3(x_1+x_2)}/(1 + e^{3(x_1+x_2)})$ and -1 otherwise.

Simulations		cv+varcv	cv+be	cv+dbi
Sim 3	Error	$0.218_{0.006}$	$0.234_{0.006}$	$\mathbf{0.209}_{0.004}$
	DBI	$0.124_{0.008}$	$0.291_{0.037}$	$\mathbf{0.107}_{0.003}$
Sim 4	Error	$0.120_{0.001}$	$0.121_{0.001}$	$\mathbf{0.119}_{0.001}$
	DBI	$0.884_{0.207}$	$0.414_{0.106}$	$\mathbf{0.235}_{0.038}$

Real Examples

- Breast cancer data set (Wolberg and Mangasarian, 1992)



- Liver disorders data: 345 samples with 6 blood test features.

Real Examples: Results

Data		cv+varcv	cv+be	cv+dbi
Breast	Error	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$
	DBI	$0.388_{0.066}$	$0.152_{0.028}$	$\mathbf{0.124}_{0.023}$
Liver	Error	$0.331_{0.006}$	$0.335_{0.006}$	$\mathbf{0.327}_{0.006}$
	DBI	$0.140_{0.013}$	$0.157_{0.024}$	$\mathbf{0.113}_{0.012}$

Assumptions for CIS Expansion

For a smooth function g, denote $\dot{g}(x)$ as its gradient vector at x.

(A1) The set $\mathcal{R} \subset \mathbb{R}^d$ is a compact d-dimensional manifold with boundary $\partial\mathcal{R}$.

(A2) The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is nonempty. There exists an open subset U_0 of $\mathbb{R}^d$ which contains $\mathcal{S}$ such that: (i) η is continuous on $U \backslash U_0$ with U an open set containing $\mathcal{R}$; (ii) the restriction of the conditional distributions of X, P_1 and P_2, to U_0 are absolutely continuous with respect to Lebesgue measure, with twice continuously differentiable Randon-Nikodym derivatives f_1 and f_2.

(A3) There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^{\rho} d\bar{P}(x) < \infty$. Moreover, for sufficiently small $\delta > 0$, $\inf_{x \in \mathcal{R}} \bar{P}(B_{\delta}(x))/(a_d \delta^d) \geq C_3 > 0$, where $a_d = \pi^{d/2}/\Gamma(1 + d/2)$, $\Gamma(\cdot)$ is gamma function, and C_3 is a constant independent of δ.

(A4) For all $x \in \mathcal{S}$, we have $\dot{\eta}(x) \neq 0$, and for all $x \in \mathcal{S} \cap \partial\mathcal{R}$, we have $\dot{\partial}\eta(x) \neq 0$, where $\partial\eta$ is the restriction of η to $\partial\mathcal{R}$.

- Assumptions (A1)–(A4) have also been employed to show the asymptotic expansion of the regret of the kNN classifier (Hall, 2008). The condition $\dot{\eta}(x) \neq 0$ in (A4) is equivalent to the margin condition with $\alpha = 1$; see (2.1) in Samworth (2012). These assumptions ensure that $\bar{f}(x_0)$ and $\dot{\eta}(x_0)$ are bounded away from zero and infinity on $\mathcal{S}$.

Definitions of B_1 and B_2

- For a smooth function g: $\mathbb{R}^d \to \mathbb{R}$, let $g_j(x)$ its jth partial derivative at x, $\ddot{g}(x)$ the Hessian matrix at x, and $g_{jk}(x)$ the (j,k)th element of $\ddot{g}(x)$. Let $c_{j,d} = \int_{v:\|v\|\leq 1} v_j^2 dv$. Define

$$a(x) = \sum_{j=1}^{d} \frac{c_{j,d}\{\eta_j(x)\bar{f}_j(x) + 1/2\eta_{jj}(x)\bar{f}(x)\}}{a_d^{1+2/d}\bar{f}(x)^{1+2/d}}.$$

- Define two distribution-related constants

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x)}{4\|\dot{\eta}(x)\|} d\mathrm{Vol}^{d-1}(x), \quad B_2 = \int_{\mathcal{S}} \frac{\bar{f}(x)}{\|\dot{\eta}(x)\|} a(x)^2 d\mathrm{Vol}^{d-1}(x),$$

  where Vol^{d-1} is the natural $(d-1)$-dimensional volume measure that $\mathcal{S}$ inherits.

- Under Assumptions (A1)-(A4), B_1 and B_2 are finite with $B_1 > 0$ and $B_2 \geq 0$.

Assumptions for Sharp Rate of CIS

- A distribution function P satisfies the *margin condition* if there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for any $\epsilon > 0$,
$$\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \epsilon) \leq C_0\epsilon^{\alpha}.$$
The parameter α characterizes the behavior of the regression function η near $1/2$, and a larger α implies a lower noise level and hence an easier classification scenario.
- The second condition is on the smoothness of $\eta(x)$. Specifically, we assume that η belongs to a *Hölder class of functions* $\Sigma(\gamma, L, \mathbb{R}^d)$ (for some fixed $L, \gamma > 0$) containing the functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor\gamma\rfloor$ times continuously differentiable and satisfy, for any $x, x' \in \mathbb{R}^d$, $|g(x') - g_x(x')| \leq L\|x - x'\|^{\gamma}$, where $\lfloor\gamma\rfloor$ is the largest integer not greater than γ, g_x is the Taylor polynomial series of degree $\lfloor\gamma\rfloor$ at x, and $\|\cdot\|$ is the Euclidean norm.
- Our last condition assumes that the marginal distribution $\bar{P}$ satisfies the *strong density assumption* which satisfies that for a compact set $\mathcal{R} \subset \mathbb{R}^d$ and constants $c_0, r_0 > 0$, $\bar{P}$ is supported on a compact (c_0, r_0)-regular set $A \subset \mathcal{R}$ satisfying $\nu_d(A \cap B_r(x)) \geq c_0\nu_d(B_r(x))$ for all $r \in [0, r_0]$ and all $x \in A$, where ν_d denotes the d-dimensional Lebesgue measure and $B_r(x)$ is a closed Euclidean ball in $\mathbb{R}^d$ centered at x and of radius $r > 0$. Moreover, for all $x \in A$, the Lebesgue density $\bar{f}$ of $\bar{P}$ satisfies $\bar{f}_{\min} \leq \bar{f}(x) \leq \bar{f}_{\max}$ for some $0 < \bar{f}_{\min} < \bar{f}_{\max}$, and $\bar{f}(x) = 0$ otherwise. In addition, $\bar{f} \in \Sigma(\gamma - 1, L, A)$.

Tuning for SNN

- *Step 1*. Randomly partition $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ into five subsets I_i, $i = 1, \cdots, 5$.
- *Step 2*. For $i = 1$, let I_1 be the test set and I_2, I_3, I_4 and I_5 be training sets. Obtain predicted labels from $\widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j)$ and $\widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j)$ respectively for each $X_j \in I_1$. Estimate the CIS and risk of the classifier with parameter λ by

$$\begin{aligned} \widehat{\mathrm{CIS}}_i(\lambda) &= \frac{1}{|I_1|} \sum_{(X_j, Y_j) \in I_1} \mathbb{1}\, \widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j) \neq \widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j), \\ \widehat{\mathrm{Risk}}_i(\lambda) &= \frac{1}{2|I_1|} \sum_{(X_j, Y_j) \in I_1} \left\{ \mathbb{1}\, \widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j) \neq Y_j + \mathbb{1}\, \widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j) \neq Y_j \right\}. \end{aligned}$$

- *Step 3*. Repeat *Step 2* for $i = 2, \ldots, 5$ and estimate the CIS and risk, with I_i being the test set and the rest being the training sets. Finally, the estimated CIS and risk are,

$$\widehat{\mathrm{CIS}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \widehat{\mathrm{CIS}}_i(\lambda), \quad \widehat{\mathrm{Risk}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \widehat{\mathrm{Risk}}_i(\lambda).$$

- *Step 4*. Perform *Step 2* and *Step 3* for each $\lambda_k \in \Lambda$. Denote the set of tuning parameters with top accuracy as

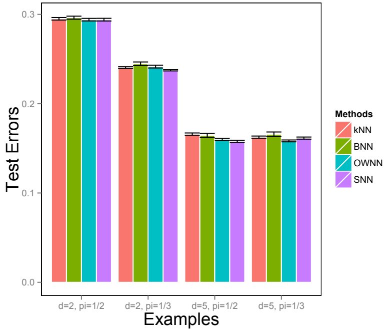
$$\mathcal{A} := \{\lambda : \widehat{\mathrm{Risk}}(\lambda) \text{ is less than the 10th percentile of } \widehat{\mathrm{Risk}}(\lambda_k),\ k = 1, \ldots, K\}.$$

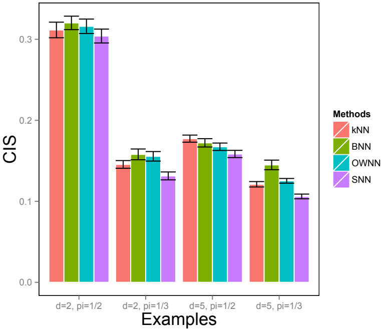
- *Step 5*. Output the optimal tuning parameter $\widehat{\lambda}$ as

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \mathcal{A}} \widehat{\mathrm{CIS}}(\lambda).$$

More Simulations in CIS Project

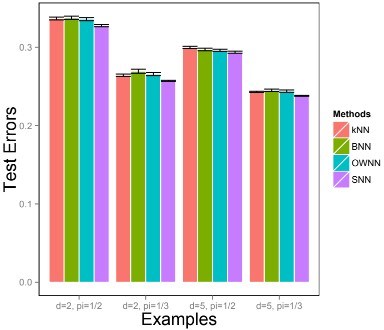
- In Simulation 2, the training data set were generated by setting $n = 200$, $d = 2$ or 5,
  $P_1 \sim 0.5N(0_d, \mathbb{I}_d) + 0.5N(3_d, 2\mathbb{I}_d)$,
  $P_2 \sim 0.5N(1.5_d, \mathbb{I}_d) + 0.5N(4.5_d, 2\mathbb{I}_d)$, and $\pi_1 = 1/2$ or $1/3$.

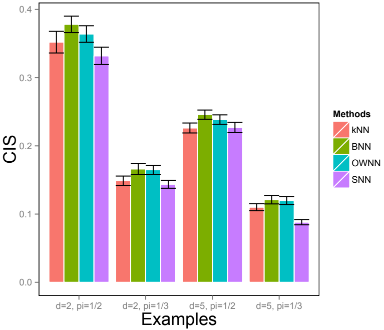




More Simulations in CIS Project

- Simulation 3 has the same setting as Simulation 2, except that $P_1 \sim 0.5N(0_d, \Sigma) + 0.5N(3_d, 2\Sigma)$ and $P_2 \sim 0.5N(1.5_d, \Sigma) + 0.5N(4.5_d, 2\Sigma)$, where Σ is the Toeplitz matrix whose jth entry of the first row is 0.6^{j-1}.


Test Errors
d=2, pi=1/2
d=2, pi=1/3
d=5, pi=1/2
d=5, pi=1/3
Examples
Methods
kNN
BNN
OWNN
SNN


CIS
d=2, pi=1/2
d=2, pi=1/3
d=5, pi=1/2
d=5, pi=1/3
Examples
Methods
kNN
BNN
OWNN
SNN