Exploiting *k*-Nearest Neighbor Information with Many Data

2017 NVIDIA DEEP LEARNING WORKSHOP

2017. 10. 31 (Tue.)

Yung-Kyun Noh

Robotics Lab., Seoul National University

Year (2015)

Help

My Registratio

Profile

Sponsorship

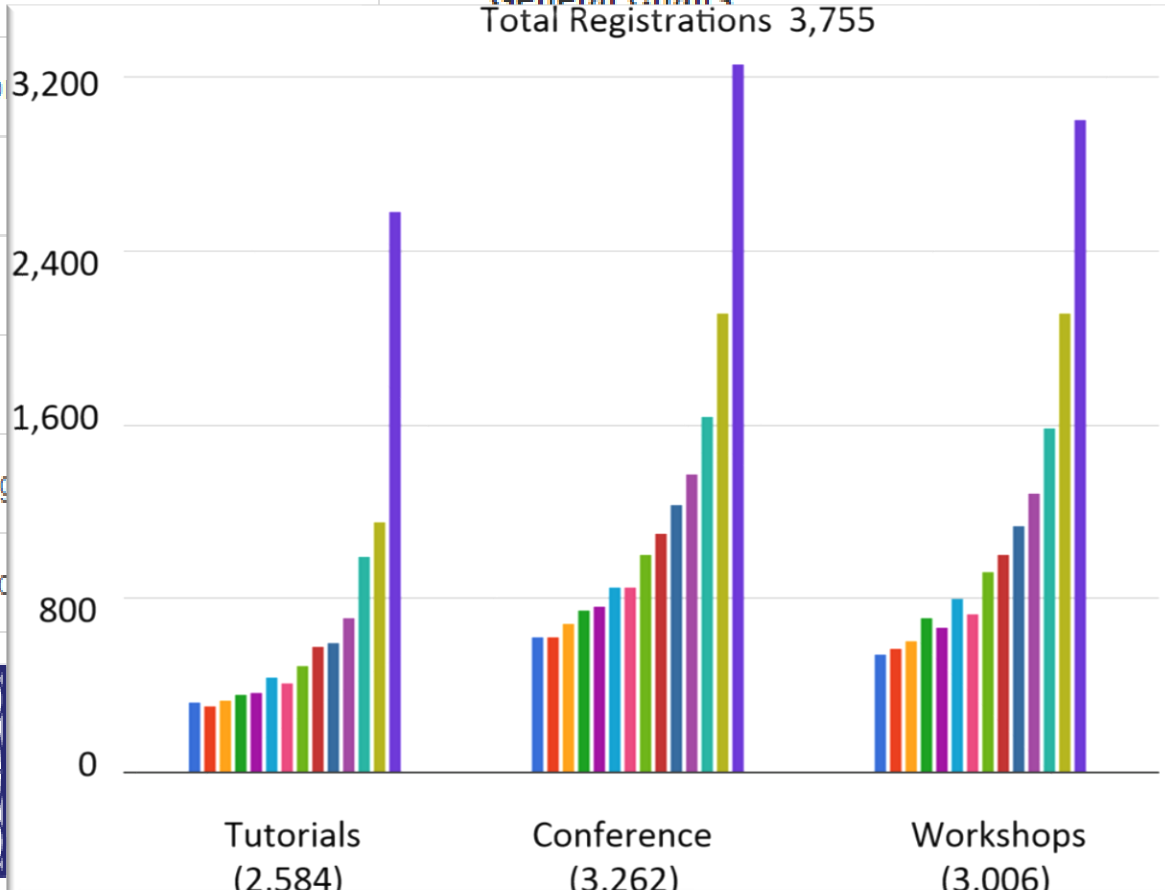
Publications

Future Meeting

Women in Mac

2015 Organizing Committee

General Chairs



o)

er Germany GmbH)

lings Chair

. Louis)

Pedro A Ortega (University of Pennsylvania)

Yung-Kyun Noh (Seoul National University)

Neural Information Processing Systems (NIPS 2015)

- Oral talks:15
- Spotlights: 37
- Accepted papers: 403
- Single session: more than 3000 participants are listening to the single presentation.
- 7pm – 12am (5hr) poster session every day

Look at the poster session how it does look ➔

From Neil Lawrence's Blog

Asian Conference on Machine Learning

ACML 2017 | Conference ▾ | Authors & Contributors ▾ | Participants ▾ | Misc ▾

The 9th Asian Conference on Machine Learning

November 15 - 17, 2017, Yonsei University, Seoul, Korea

ACML 2017

Asian Conference on Machine Learning

ACML 2017

Welcome to the 9th Asian Conference on Machine Learning (ACML 2017). The conference will take place on November 15 - 17, 2017 at Baekyang Hall of Yonsei University campus, Seoul, Korea. We invite professionals and researchers to discuss research results and ideas in machine learning. We seek original and novel research papers resulting from theory and experiment of machine learning. The conference also solicits proposals focusing on disruptive ideas and paradigms within the scope. We encourage submissions from all parts of the world, not only confined to the Asia-Pacific region.

As machine plays critical role in various fields of industry, machine learning researchers needed to gather and share new ideas and achievements at a forum. ACML has begun to take place annually over the Asian regions since 2009. This is the 9th Conference to be held in Seoul, Korea after Hamilton, New Zealand (2016), Hong Kong, China (2015), Nha Trang, Vietnam (2014), Canberra, Australia (2013), Singapore (2012), Taoyuan, Taiwan (2011), Tokyo, Japan (2010), and Nanjing, China (2009). The conference has contributed to understanding the machine leaning, bringing inspiration to scientists, and applying the technologies to industries. This conference will consist of informative and integrated programs as traditions of the previous ones.

Yonsei University, one of most prestigious universities, is about 130 years old historical campus in Korea. The University street called "Sinchon" is connected to Ewha Womans University and Hongik University as one of youth hotspots. You can walk along ‘Sinchon’s Pedestrian Friendly Street’ which is full of cafes, fashion items, and beauty goods. The district is located at the heart of Seoul with easy access to cultural and attractive sites. Seoul is ranked by Asian tourists as their favorite world city three years in a row. Come experience the history and excitement of modern Seoul.

Authors & Contributors

- Call for Papers

Speakers

The confirmed speakers are:

- Bernhard Schölkopf - *keynote*

  Professor and Director of Max Planck Institute for Intelligent Systems, Germany

- Tom Dietterich - *keynote*

Contents

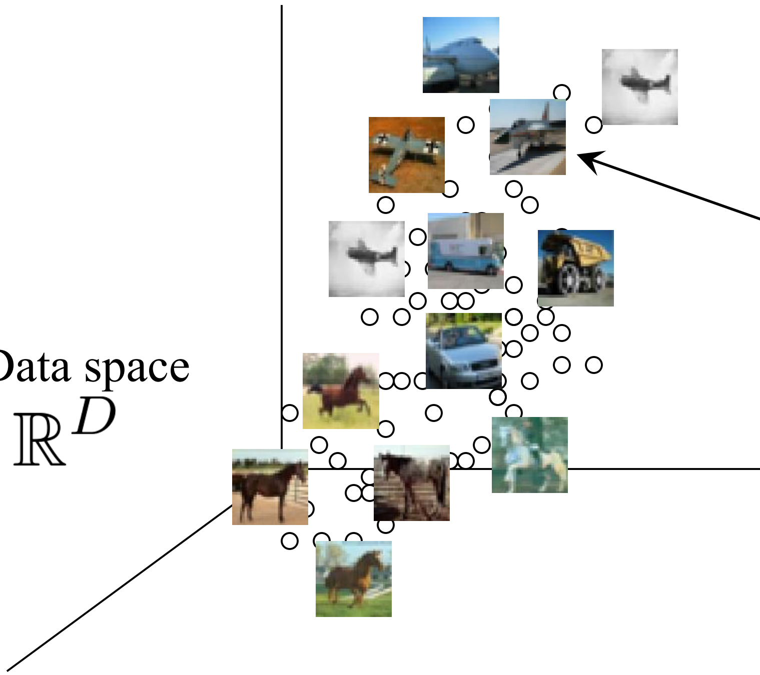
- Nonparametric methods for estimating density functions
  - Nearest neighbor methods
  - Kernel density estimation methods

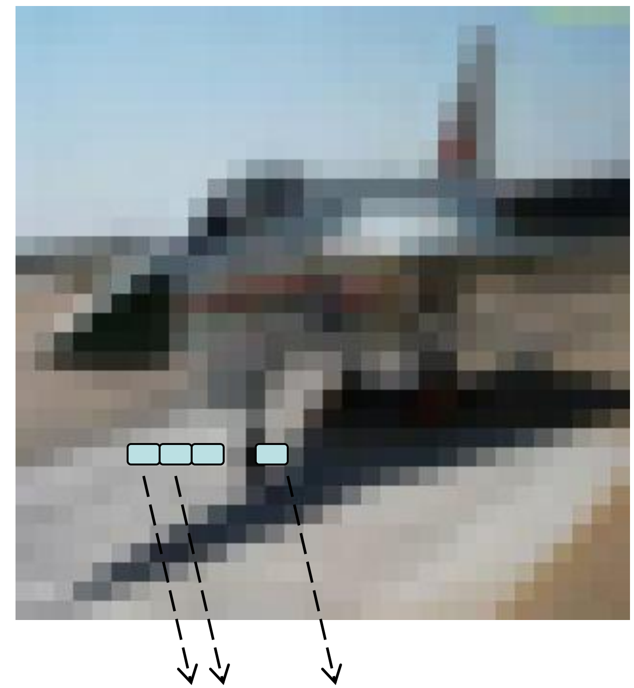
- Metric learning for nonparametric methods
  - Generative approach for metric learning

- Theoretical properties and applications

Representation of Data

Data space
$\mathbb{R}^D$



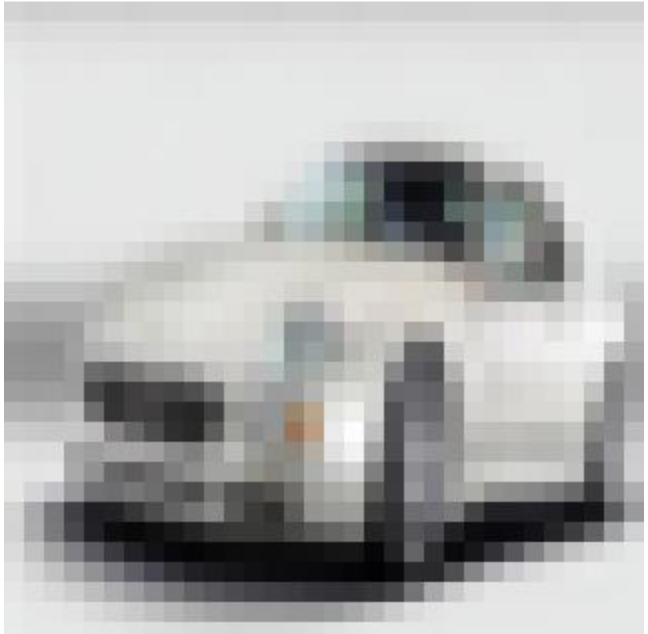


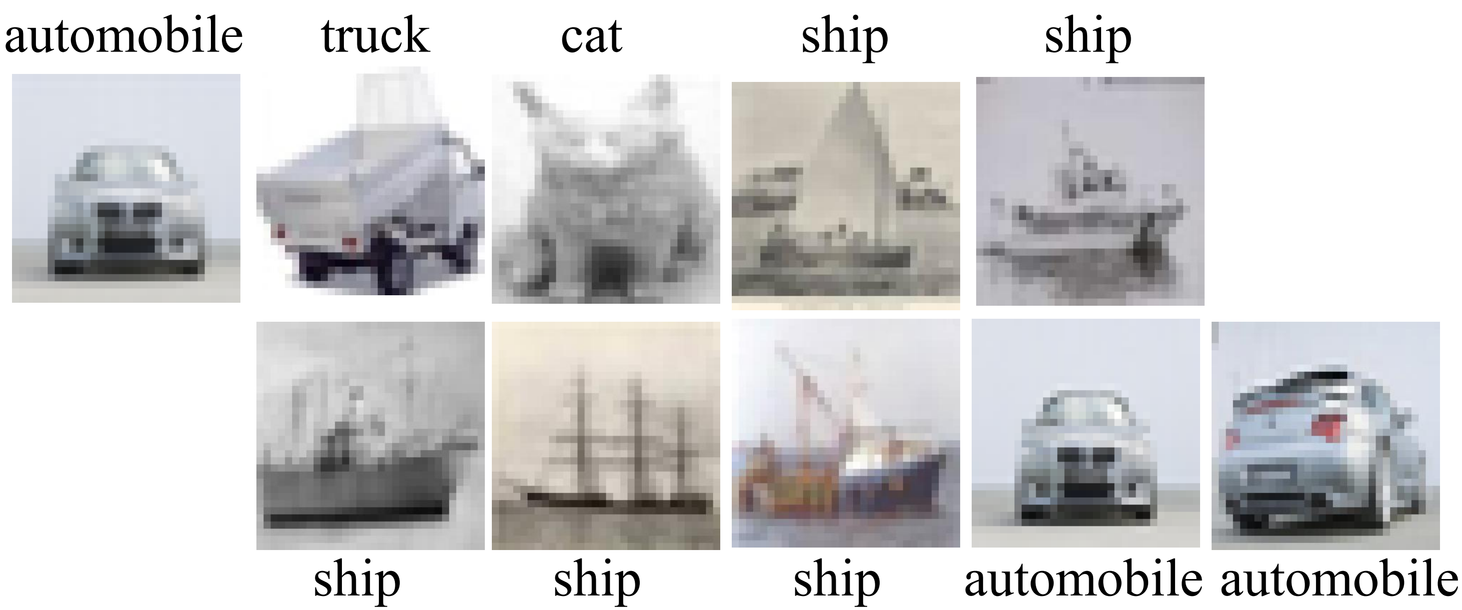
$$=\underbrace{[1, 2, 5, 10, \ldots]^{\mathrm{T}}}_{D}$$

- Each datum is one point in a data space

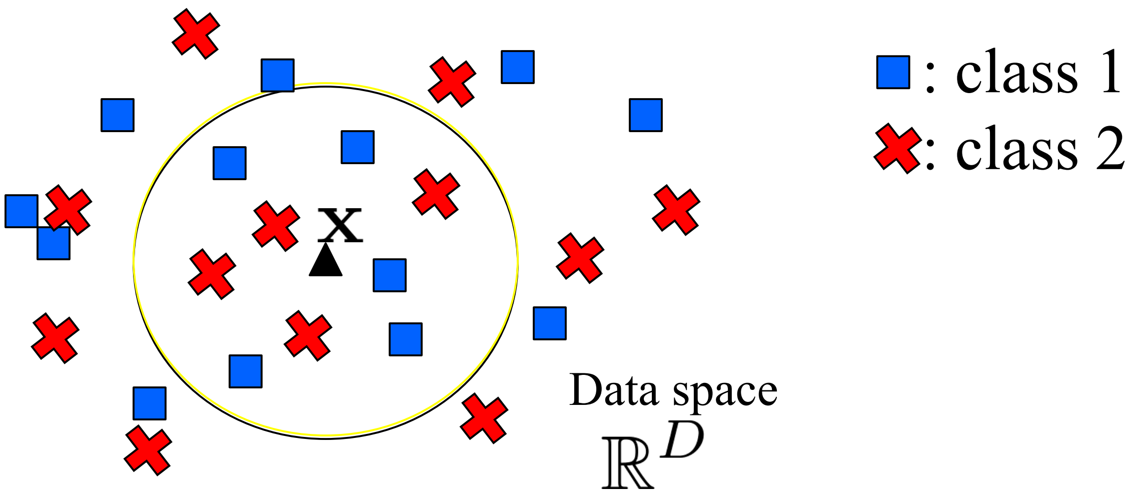
Nearest Points

Nearest Points





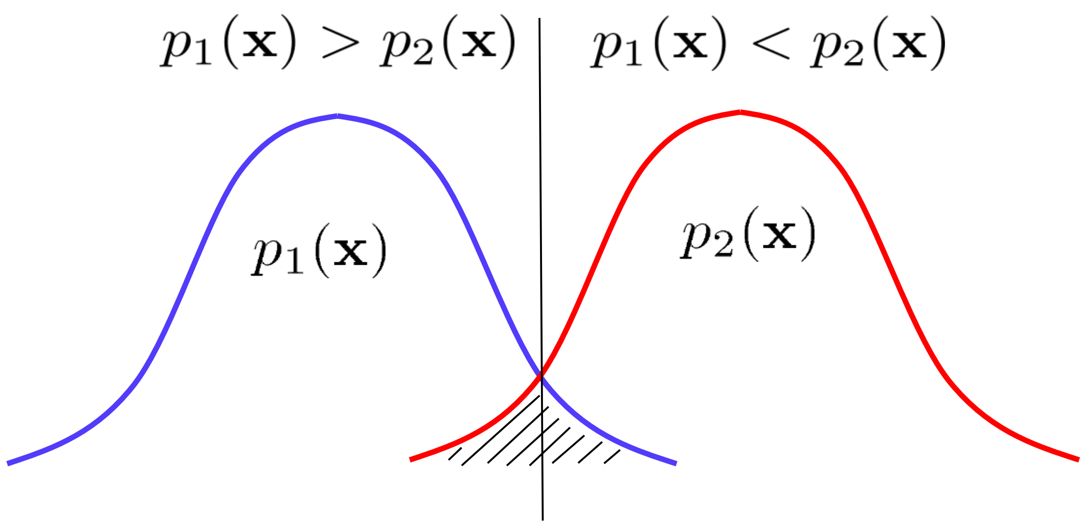
Classification with Nearest Neighbors



- Use majority voting (*k*-nearest neighbor classification)
- $k = 9$ (five ■ / four ✖)
- Classify a testing point $\mathbf{x}$ (▲) as class 1 (■).

Bayes Classification

- Bayes classification using underlying density functions: *Optimal*



➔ Error:

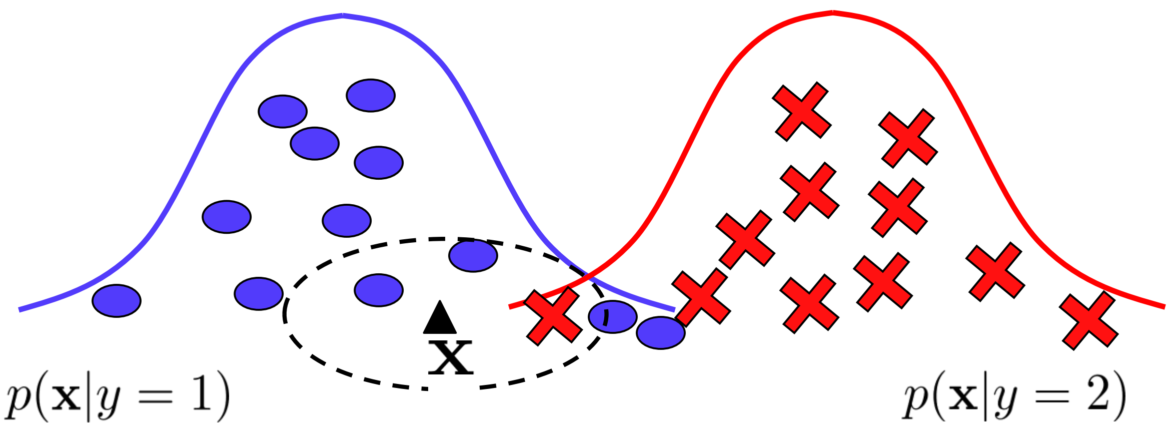
$$\frac{1}{2}\int \min[p_1, p_2]\,d\mathbf{x}$$

Bayes risk

In general,
we do not know the underlying density.

Nearest Neighbors and Bayes Classification

- Surrogate method of using underlying density functions.

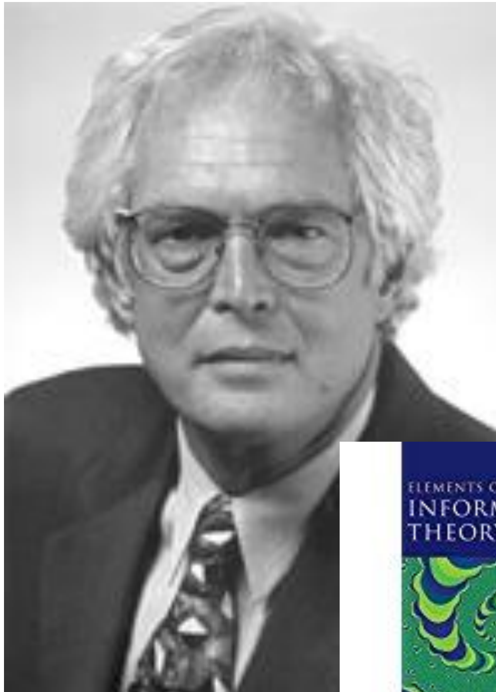


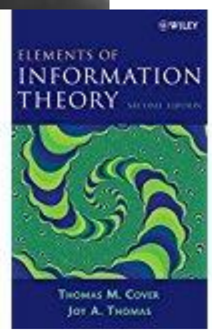
$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$

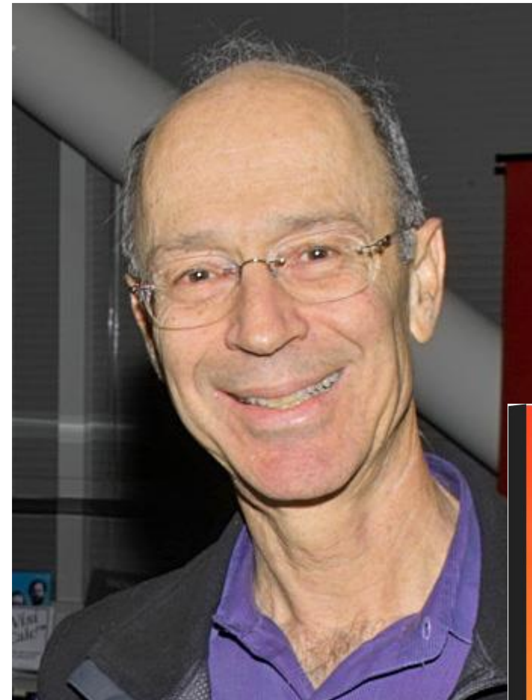
$$p_1(\mathbf{x}) \gtrless p_2(\mathbf{x})? \Longrightarrow$$

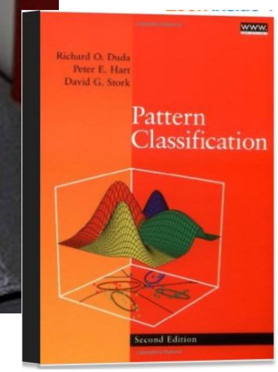
Count nearest neighbors!

$$N_1 \gtrless N_2?$$









- **Tomas M. Cover** (8/7/1938~3/26/2012)
- BS. in Physics from MIT
- Ph.D. in EE from Stanford
- Professor in EE and Statistics, Stanford

- **Peter E. Hart** (Bone c. 1940s)
- MS., Ph.D. from Stanford
- A strong advocate of artificial intelligence in industry
- Currently Group Senior Vice President at the Ricoh Company, Ltd.

This Week's Citation Classic

CC/NUMBER 13
MARCH 29, 1982

Cover T M & Hart P E. Nearest neighbor pattern classification.
IEEE Trans. Inform. Theory **IT-13**:21-7, 1967.
[Dept. Electrical Engineering, Stanford Univ., Stanford, and Stanford Res. Inst., Menlo Park, CA]

The nearest neighbor decision rule assigns to an unclassified sample the classification of the nearest of a set of previously classified samples. This paper proves that the probability of error of the nearest neighbor rule is bounded above by twice the Bayes minimum probability of error. In this sense, it may be said that half the classification information in an infinite sample set is contained in the nearest neighbor. [The *SCI*® indicates that this paper has been cited over 190 times since 1967.]

Thomas M. Cover
Departments of Statistics and
Electrical Engineering
Stanford University
Stanford, CA 94305

March 5, 1982

"Early in 1966 when I first began teaching at Stanford, a student, Peter Hart, walked into my office with an interesting problem. He said that Charles Cole and he were using a pattern classification scheme which, for lack of a better word, they described as the nearest neighbor procedure. This scheme assigned to an as yet unclassified observation the classification of the nearest neighbor. Were there any good theoretical properties of this procedure? Of course the motivation for such a classification rule must go back to prehistoric times. The idea is that 'things that look alike must be alike.'

"The problem seemed extremely inviting from a theoretical point of view. We began meeting for two or three hours every afternoon in an attempt to find some distribution-free properties of this classification rule. By distribution-free, I mean properties that are true regardless of the underlying joint distribution of the categories and observations. Obviously, we could not hope to prove that a procedure always has, for example, a zero probability of error, because there are many cases where the observations yield no information about the underlying category. In those problems, the goal would be more modest. Apparently, the proper goal would be to relate the probability of error of this procedure to the minimal probability of error given complete statistical information, namely, the Bayes risk.

"After some effort, we were able to prove that the nearest neighbor risk is always less than the Bayes risk plus one-sixth (if I remember correctly). This was the sort of result we were looking for, but it seemed quite unnatural. Also, it was not a very ambitious bound when the Bayes risk is near zero. We would have preferred to relate risks by a factor rather than by an additive constant. Soon thereafter we found what we were looking for. The nearest neighbor risk is less than twice the Bayes risk for all reasonable distributions and for any number of categories. Thus ancient man was proved right—'things that look alike are alike'—with a probability of error that is no worse than twice the probability of error of the most sophisticated modern day statistician using the same information. Moreover, we were soon able to prove that our bound was the best possible. So the search was over.

"The simplicity of the bound and the sweeping generality of the statement, combined with the obvious simplicity of the nearest neighbor rule itself, have caused this result to be used by others, thus accounting for the high number of citations. Since the properties of the nearest neighbor rule can be easily remembered, the bound yields a benchmark for other more sophisticated data analysis procedures, which sometimes actually perform worse than the nearest neighbor rule. This is probably due to the fact that more ambitious rules have too many parameters for the data set.

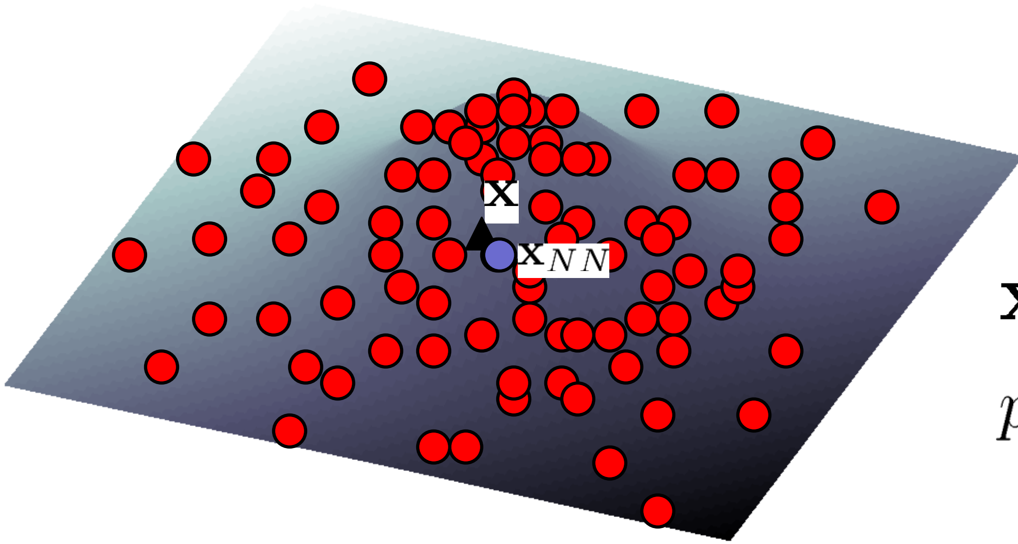
"It should be mentioned that we had to exclude a certain technical set of joint distributions from the proof of our theorem. The attendant measure-theoretic difficulties in eliminating the so-called singular distributions almost delayed the publication of our paper. It was wise that we did not hold up publication, because the theorem was not proved in total generality until ten years later in Charles Stone's 1977 paper in the *Annals of Statistics*.[1] The result remains the same, but now it applies to all possible probability distributions."

1. Stone C J. Consistent nonparametric regression. *Ann. Statist.* 5:595-645, 1977.

- Early in 1966 when I first began teaching at Stanford, a student, Peter Hart, walked into my office with an interesting problem.
- Charles Cole and he were using a pattern classification scheme which, for lack of a better word, they described as the nearest neighbor procedure.
- The proper goal would be to relate the probability of error of this procedure to the minimal probability of error … namely, the Bayes risk.

ROBOTICS LABORATORY

Nearest Neighbors and Bayes Risk



$\mathbf{x}_{NN} \to \mathbf{x}$, uniformly!

$p(\mathbf{x}) \sim p(\mathbf{x}_{NN})$

- *1*-NN error

$$\epsilon_{(k=1)} = \int \frac{p_1(\mathbf{x})p_2(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} d\mathbf{x} \quad \leq 2E_{Bayes}(1 - E_{Bayes})$$

- *k*-NN error

$$E_{Bayes} = \underline{\epsilon_{(k=\infty)} \leq \ldots \leq \epsilon_{(k=2)} \leq \epsilon_{(k=1)}} \leq 2E_{Bayes}(1 - E_{Bayes})$$

[T. Cover and P. Hart, 1967]

Bias in the Expected Error

- Assumption:

  A nearest neighbor appears at nonzero $d_N > 0$.

$$E_{NN} \cong \int \frac{p_1(\mathbf{x})p_2(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} d\mathbf{x} \quad \cdots ①$$

$$+ \frac{1}{4D} \int \frac{\mathbb{E}_{d_N}[d_N^2|\mathbf{x}]}{(p_1+p_2)^2} \boxed{[p_1^2 \underline{\nabla^2 p_2} + p_2^2 \underline{\nabla^2 p_1} - p_1 p_2 (\underline{\nabla^2 p_1} + \underline{\nabla^2 p_2})]} d\mathbf{x} \quad \cdots ②$$

Metric-dependent terms

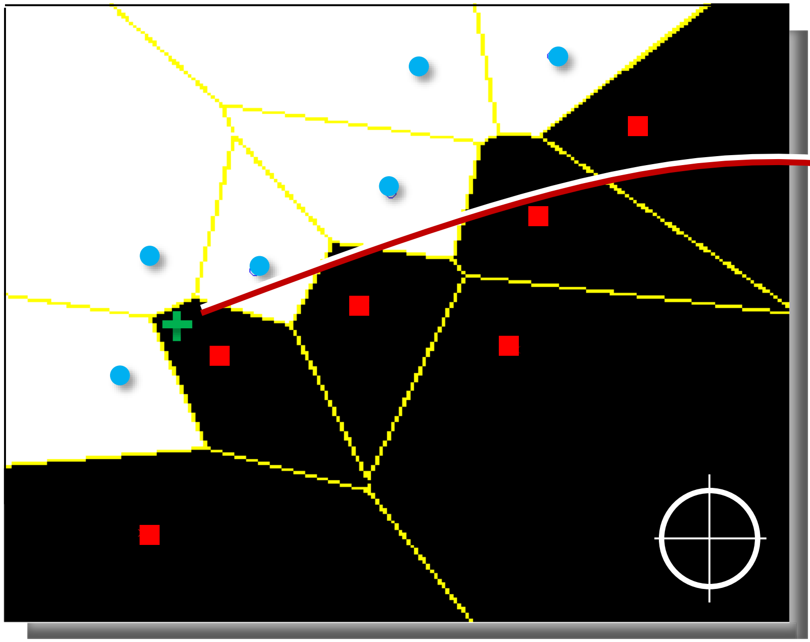
①: Asymptotic *NN* Error

②: Residual due to *Finite Sampling* .

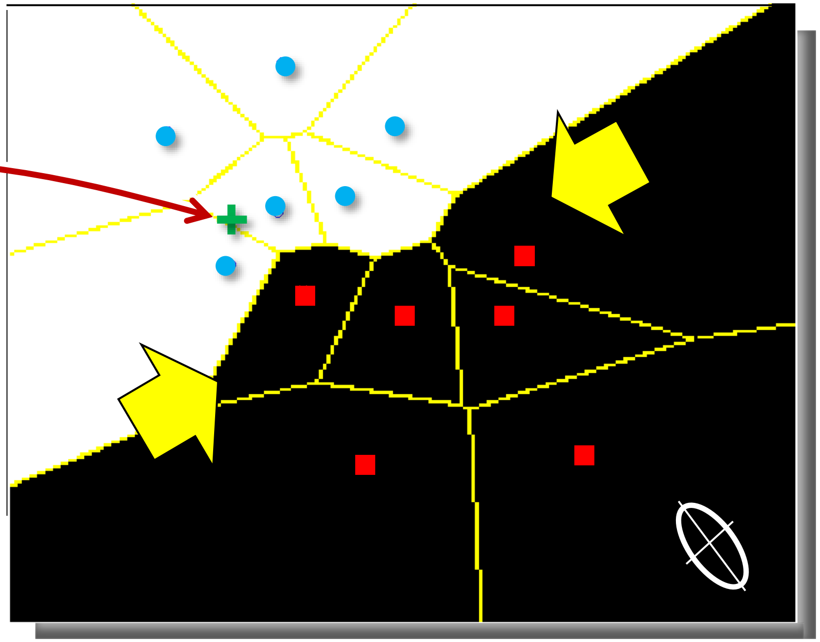
R. R. Snapp et al. (1998) Asymptotic expansions of the *k* nearest neighbor risk, *The Annals of Statistics*
Y.-K. Noh et al. (2010) Generative local metric learning for nearest neighbor classification, *NIPS*

Metric Dependency of Nearest Neighbors

- Different metric changes class belongings



Classified as red

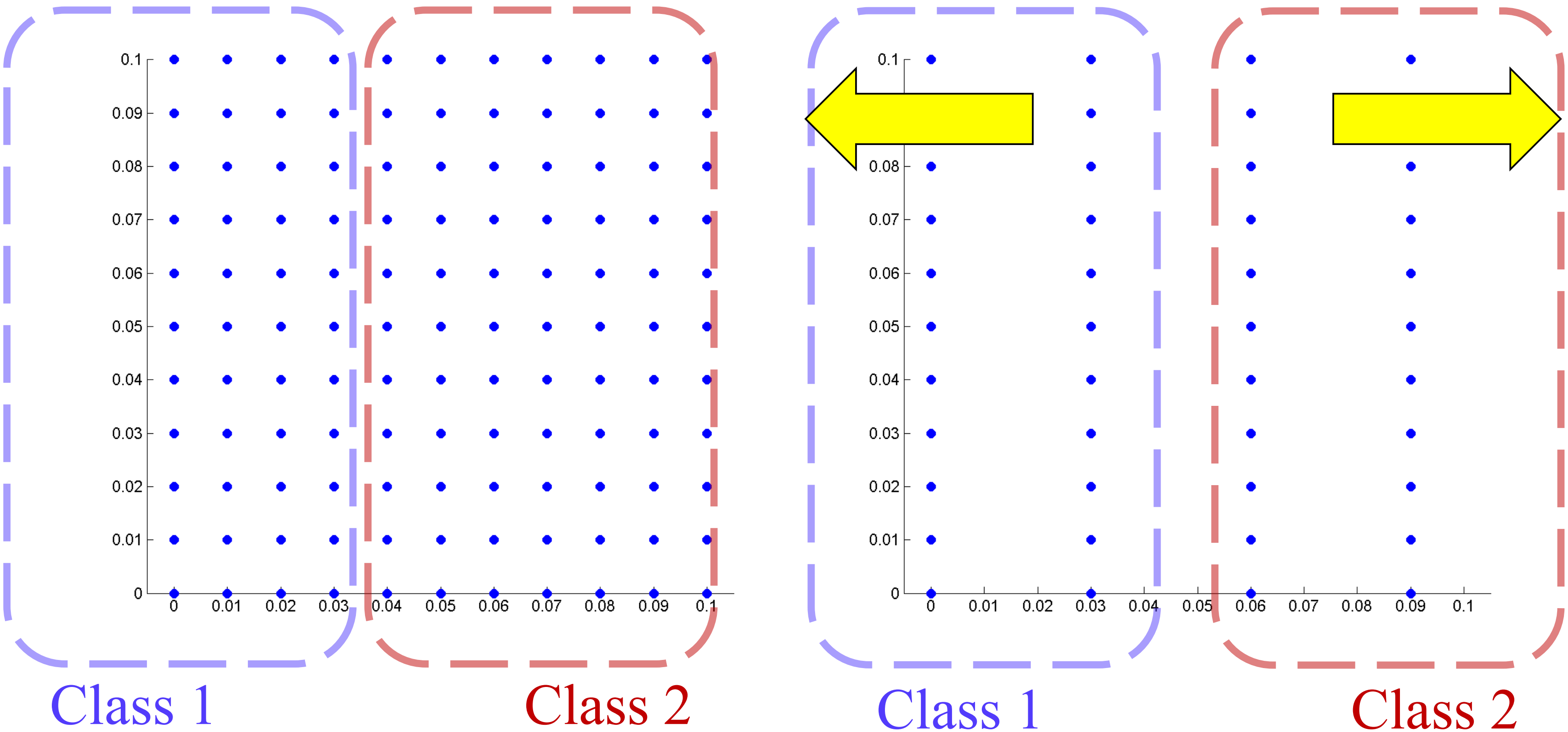


Classified as blue

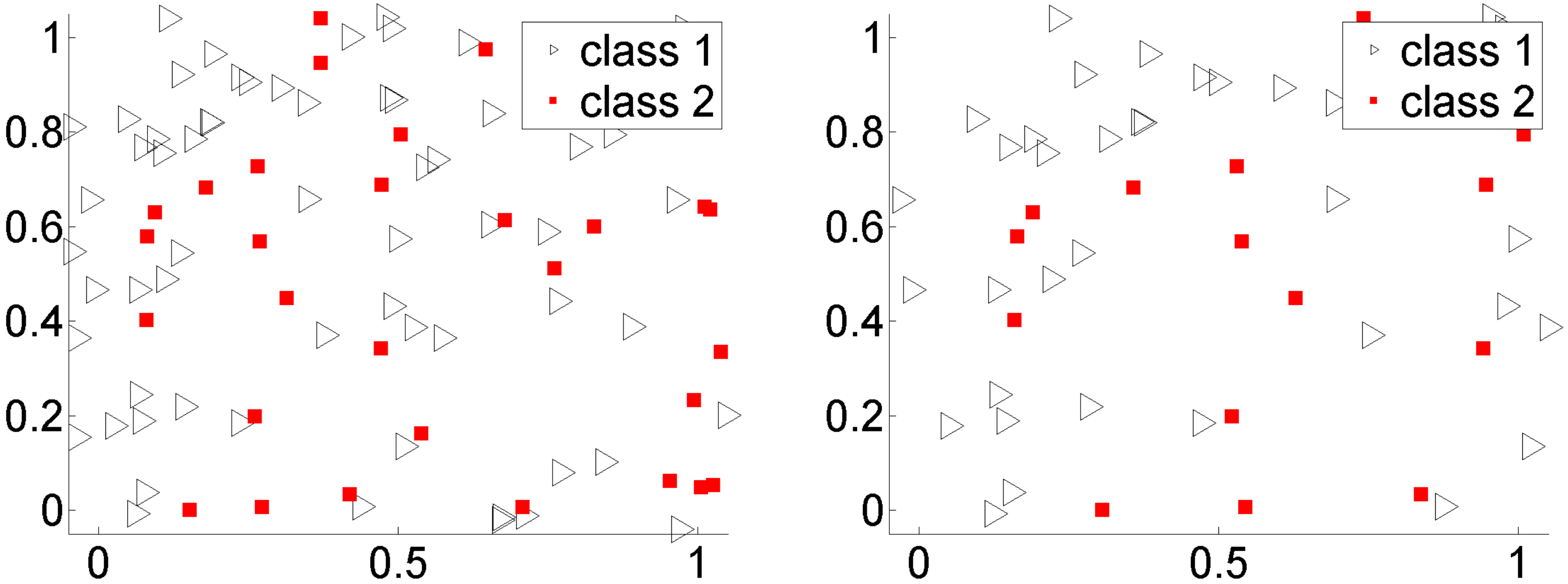
Mahalanobis-type distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T A (\mathbf{x}_i - \mathbf{x}_j)}, \quad A \succ 0$$

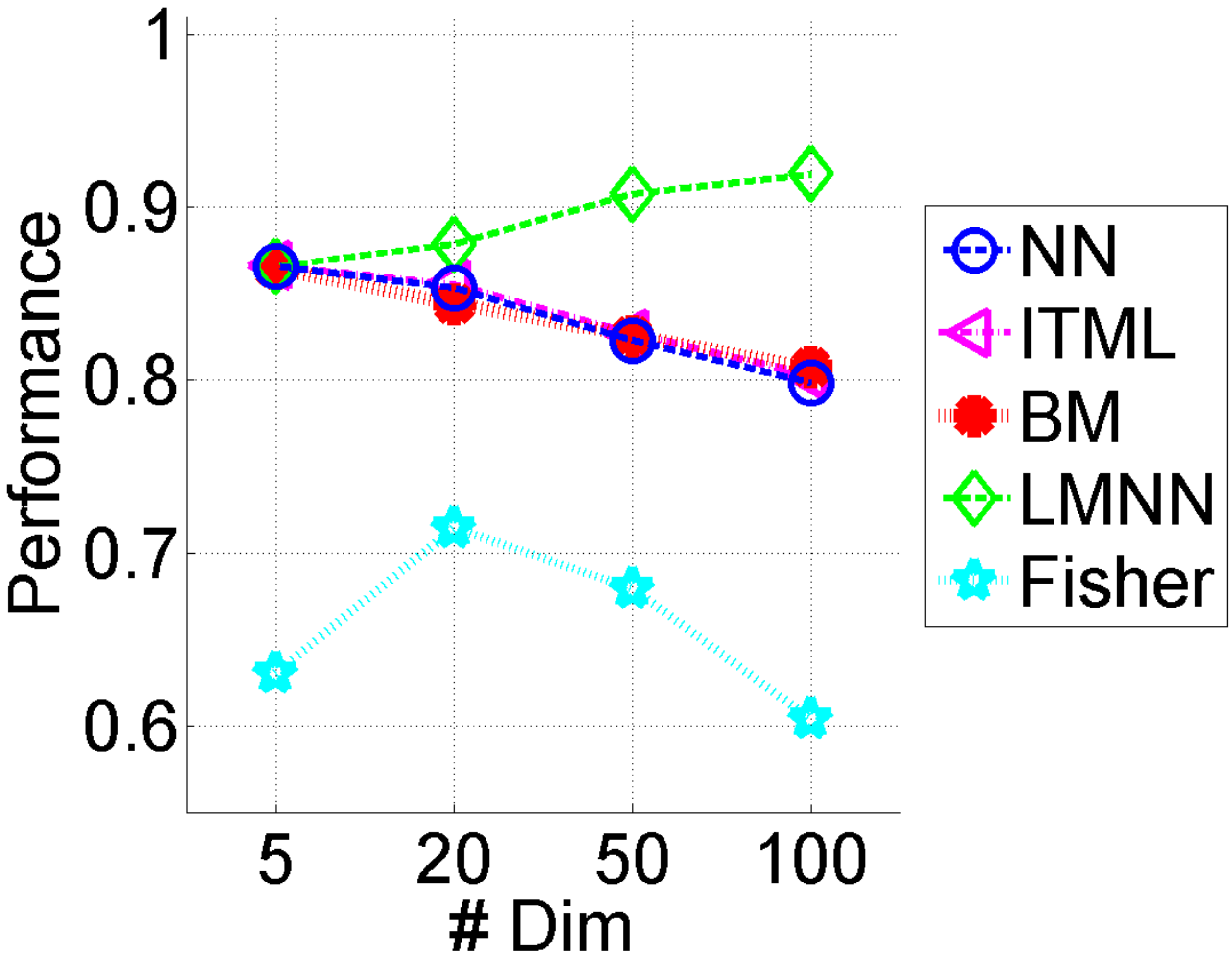
Conventional Idea of Metric Learning



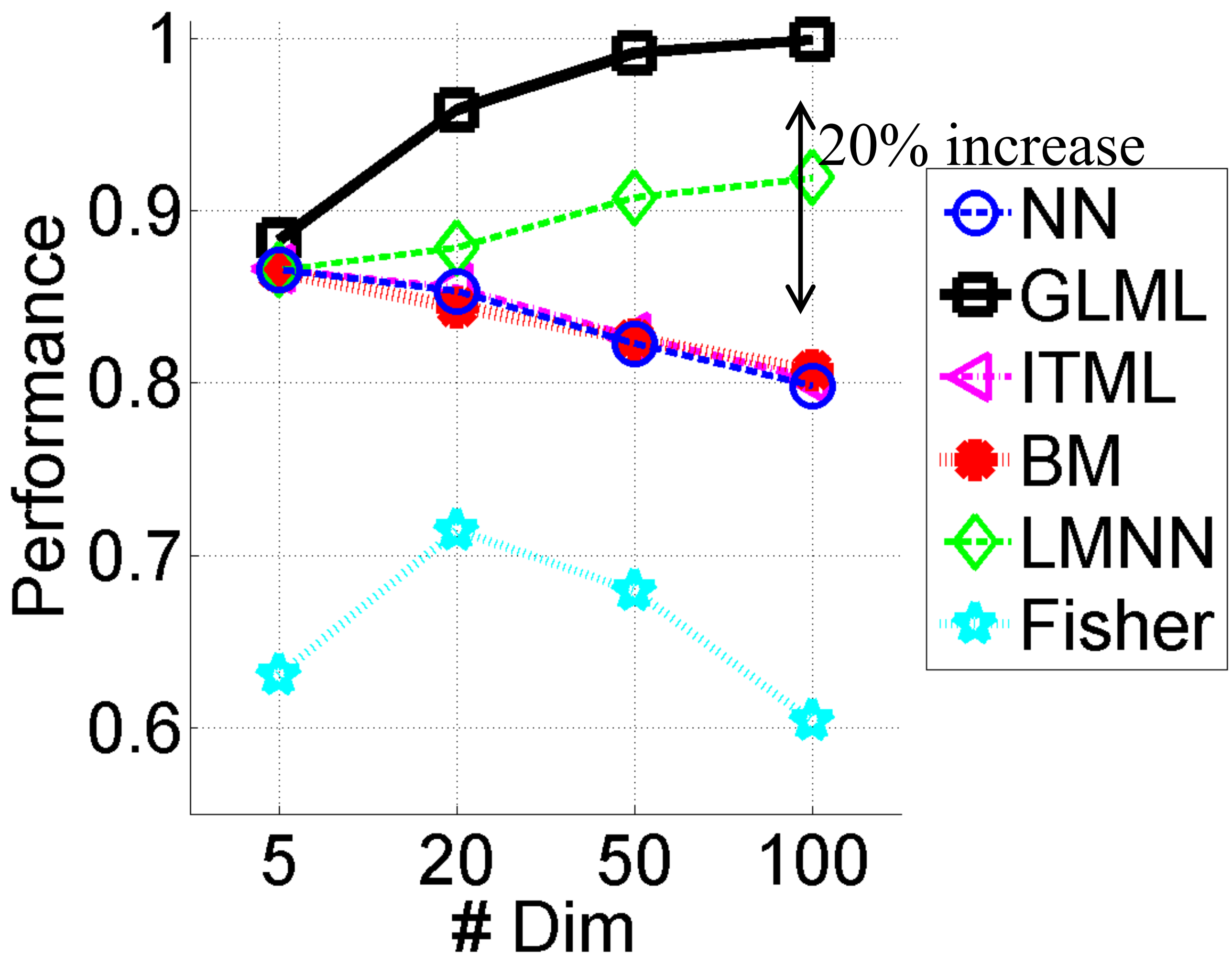
Many Data Situation with Overlap



Conventional Metric Learning

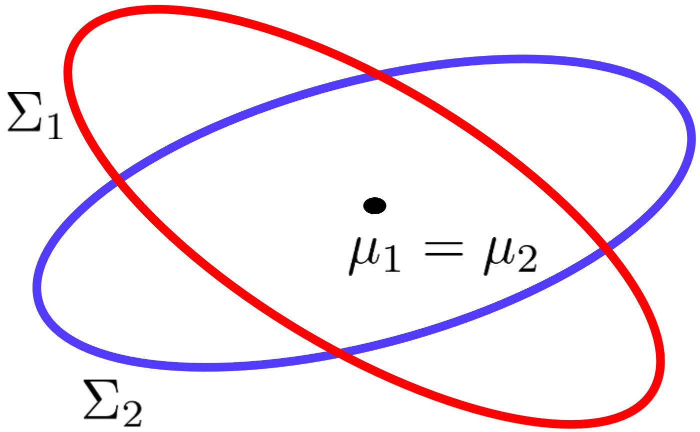


Generative Local Metric Learning (GLML)

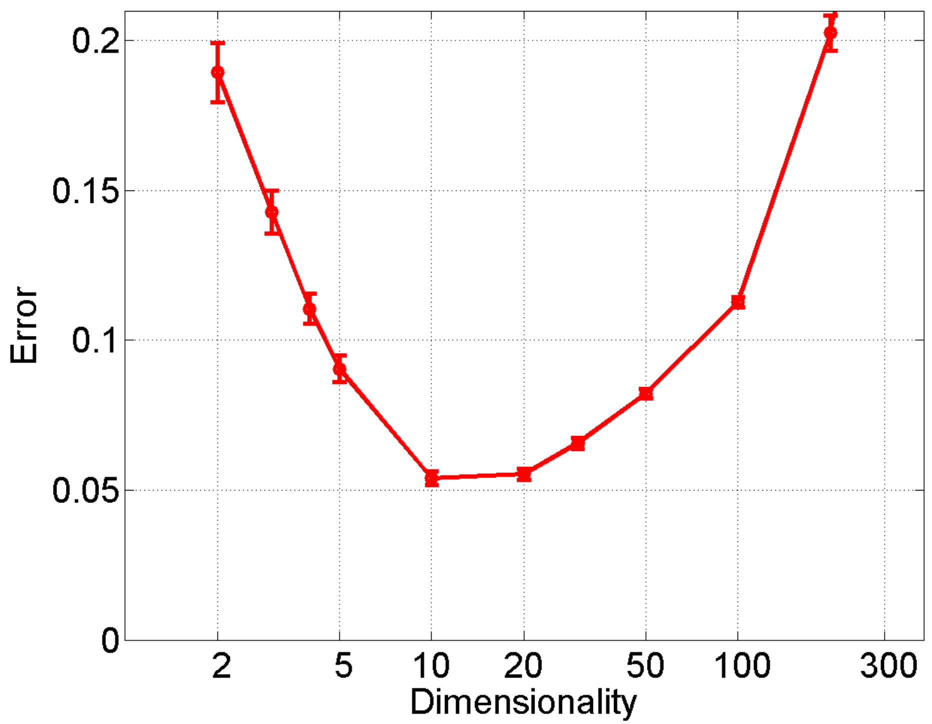


Bayes Classification with True Model

- Two Gaussians
  - same means, random covariance matrices
  - Number of data: 20 per class

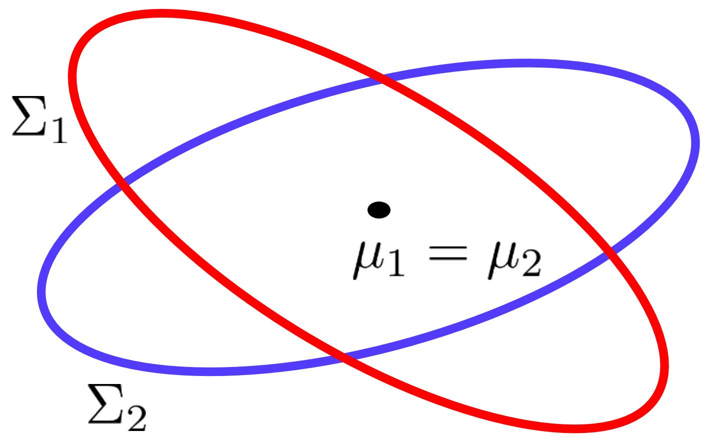


$$p_1(\mathbf{x}):\ \mathcal{N}(\mu_1, \Sigma_1)$$
$$p_2(\mathbf{x}):\ \mathcal{N}(\mu_2, \Sigma_2)$$

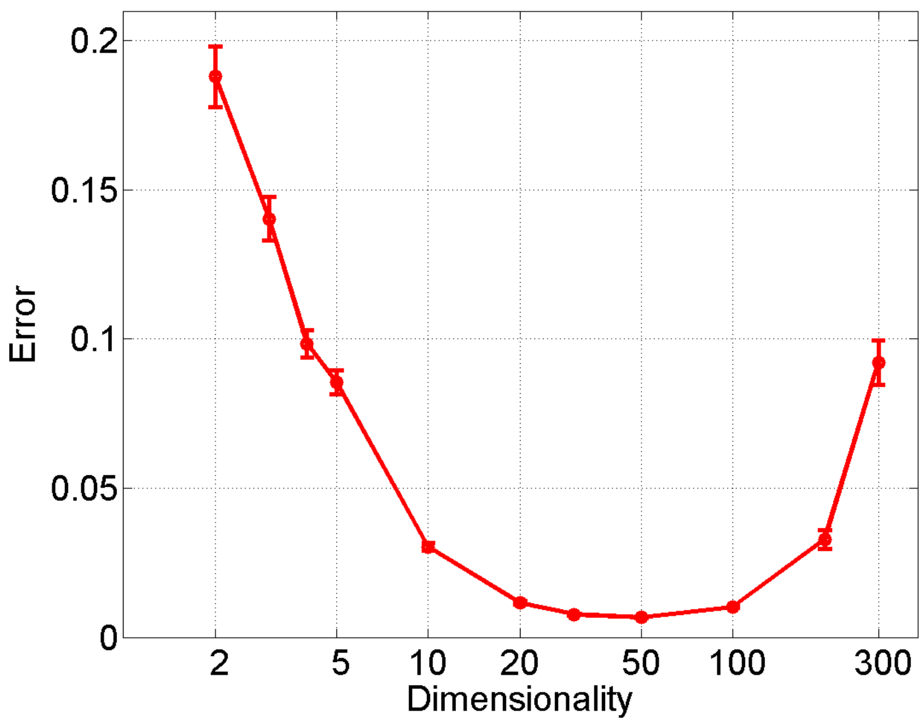


Bayes Classification with True Model

- Two Gaussians
  - same means, random covariance matrices
  - Number of data: 50 per class

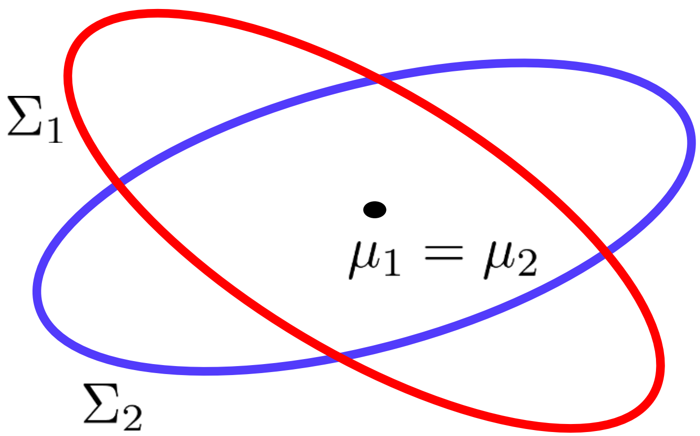


$$p_1(\mathbf{x}):\ \mathcal{N}(\mu_1, \Sigma_1)$$
$$p_2(\mathbf{x}):\ \mathcal{N}(\mu_2, \Sigma_2)$$

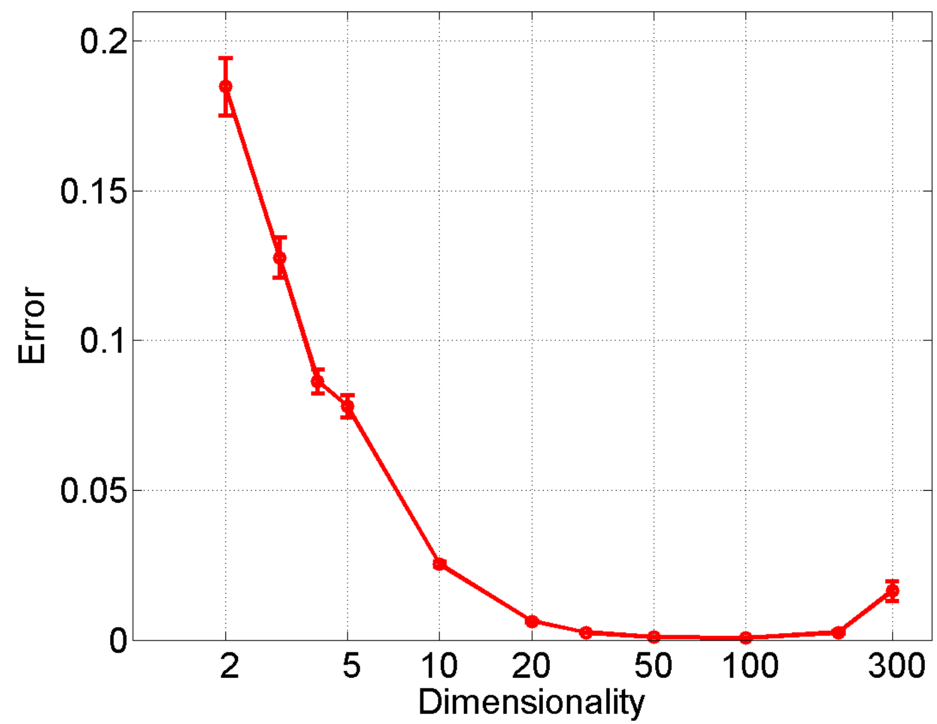


Bayes Classification with True Model

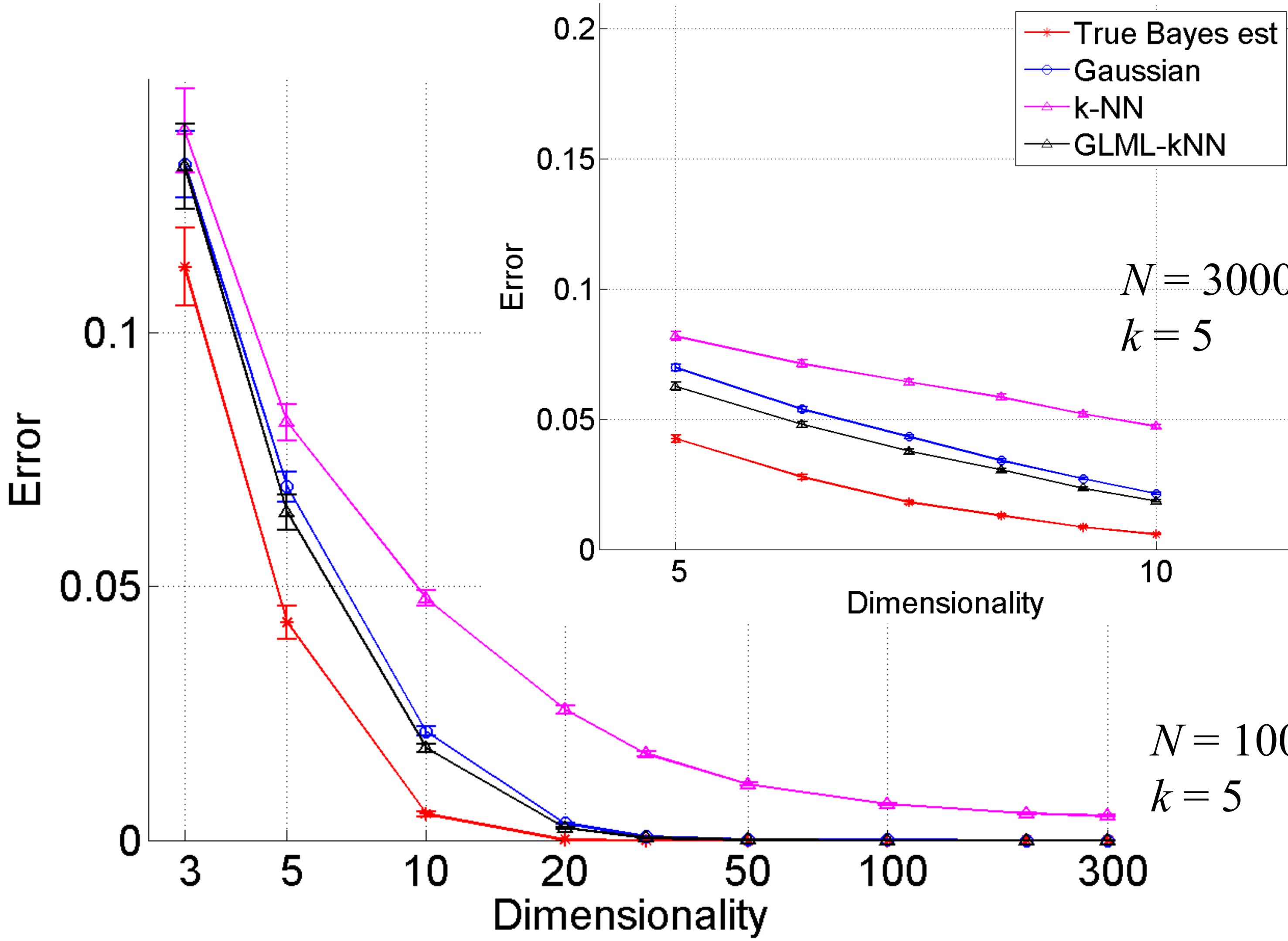
- Two Gaussians
  - same means, random covariance matrices
  - Number of data: 100 per class



$$p_1(\mathbf{x}): \ \mathcal{N}(\mu_1, \Sigma_1)$$
$$p_2(\mathbf{x}): \ \mathcal{N}(\mu_2, \Sigma_2)$$



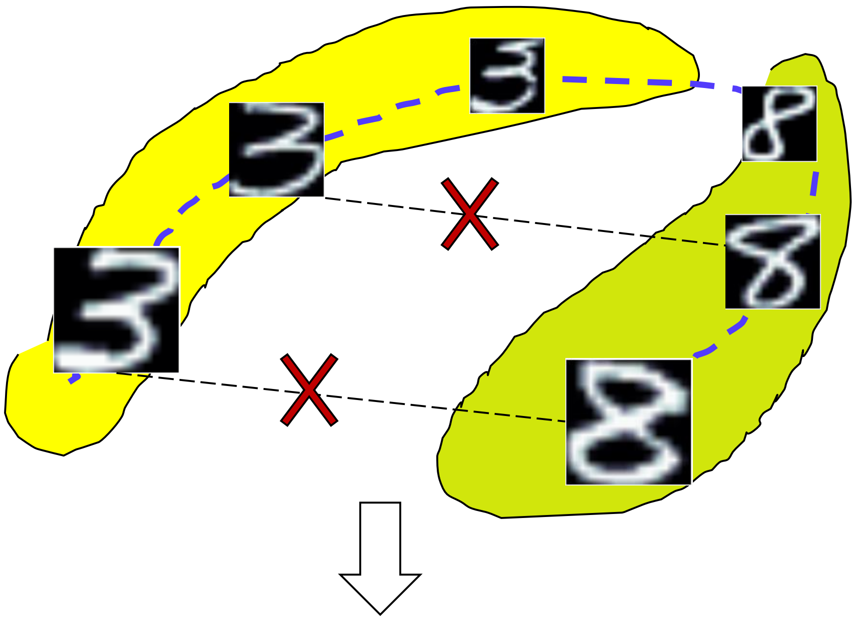
k-NN Beats True Model With Metric Learning!



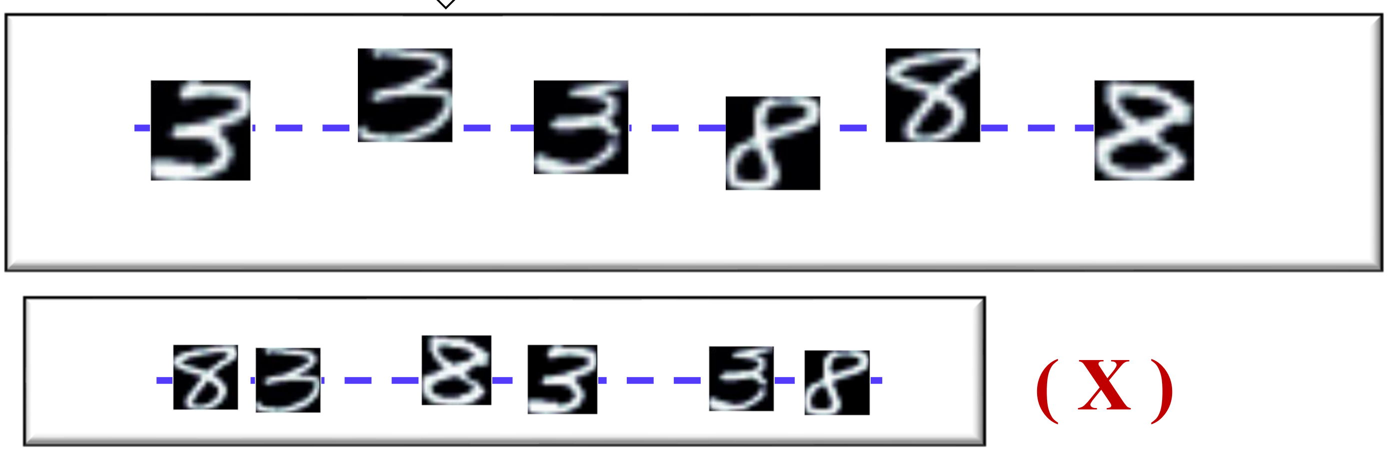
$N = 3000$/class
$k = 5$

$N = 1000$/class
$k = 5$

Manifold Embedding (Isomap)

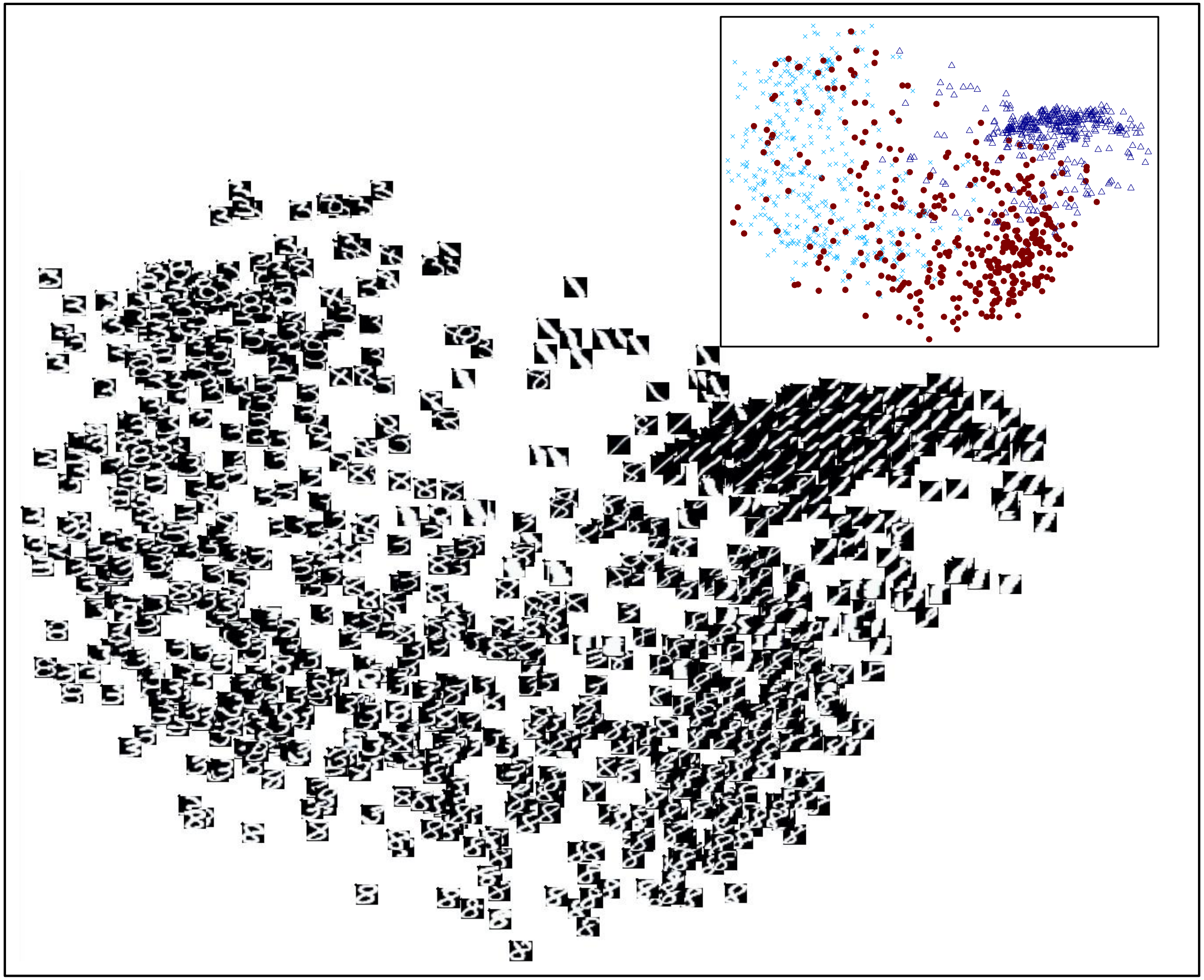


Use Dijkstra algorithm to calculate the manifold distance from nearest neighbor distance
➔ MDS using manifold distance

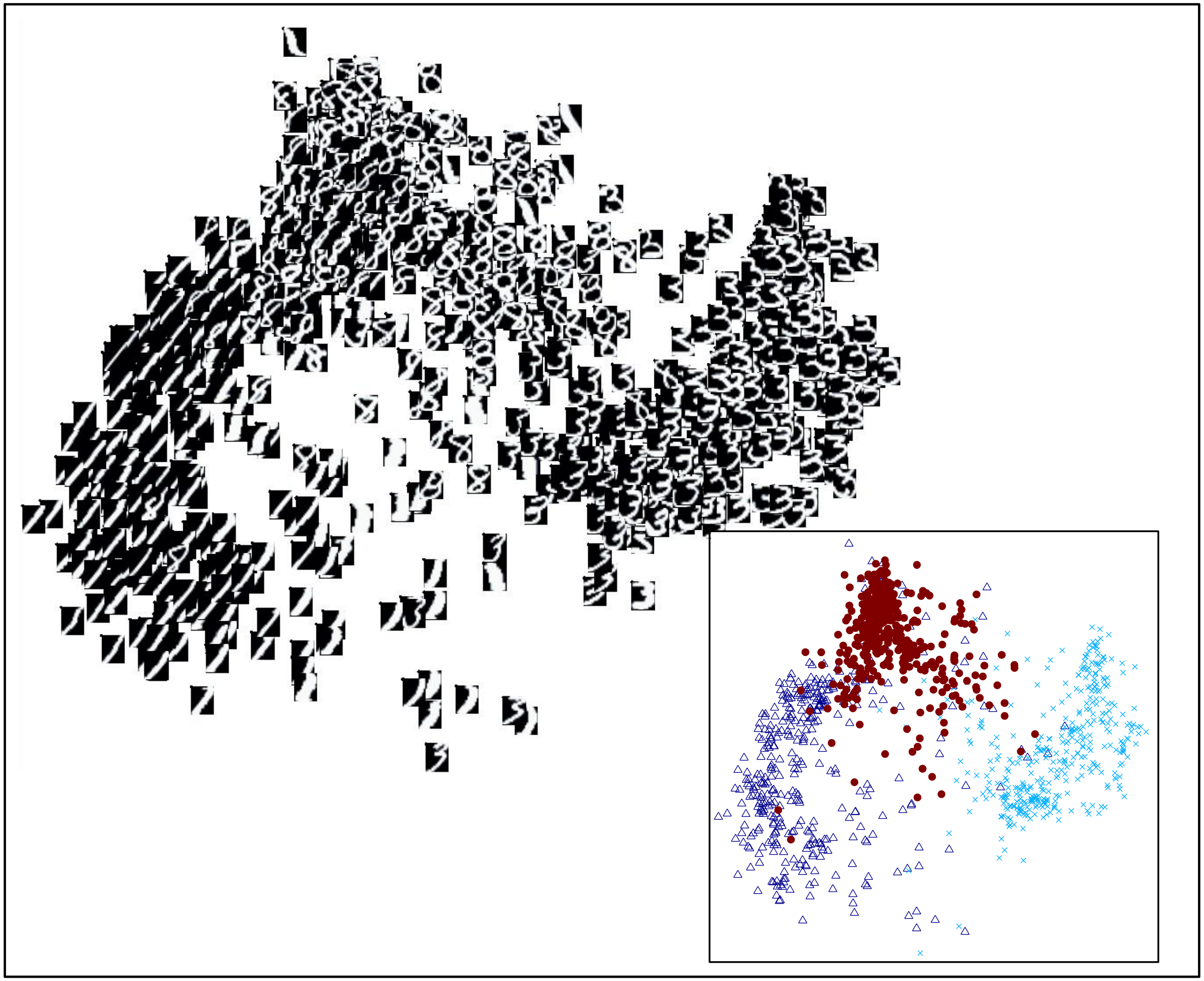


(X)

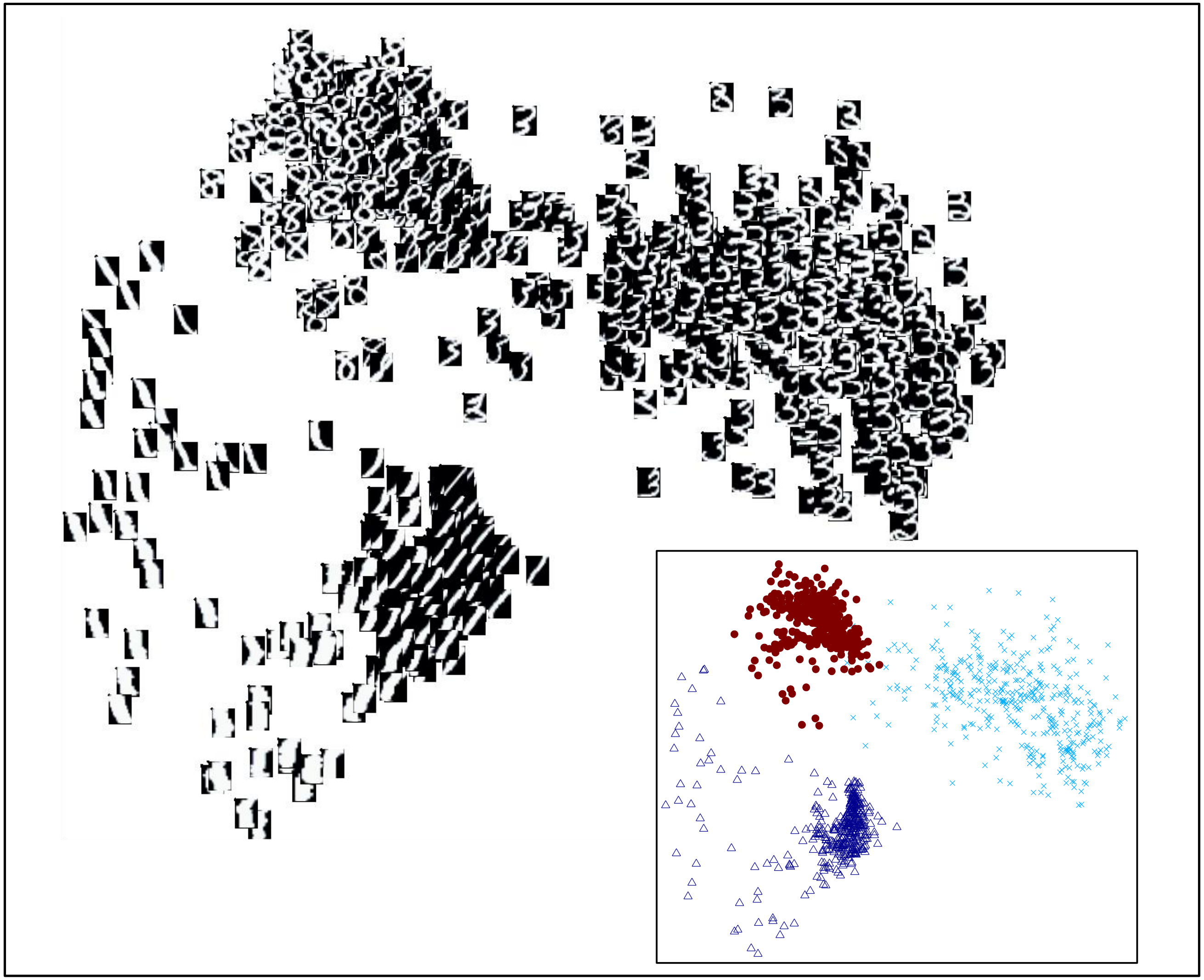
Manifold Embedding (Isomap)



Isomap with LMNN Metric



Isomap with GLM Metric



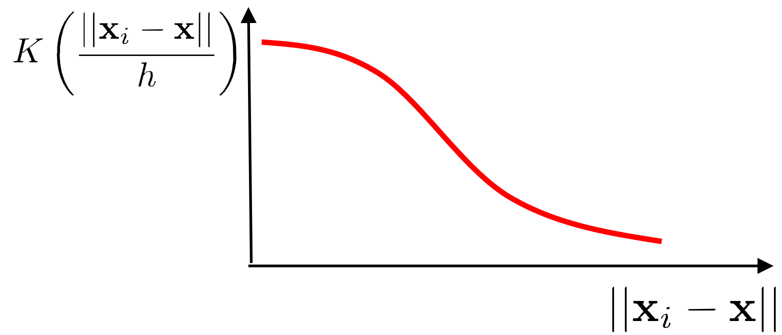
Nadaraya-Watson Estimator

$$\widehat{y}_N(\mathbf{x}) = \frac{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{j=1}^{N} K(\mathbf{x}_j, \mathbf{x})}$$

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

$$\mathbf{x}_i \in \mathbb{R}^D$$

$$K(\mathbf{x}_i, \mathbf{x}) = K\left(\frac{||\mathbf{x}_i - \mathbf{x}||}{h}\right) = \frac{1}{\sqrt{2\pi}^D h^D} \exp\left(-\frac{1}{2h^2}||\mathbf{x}_i - \mathbf{x}||^2\right)$$



$y_i \in \{0, 1\}$ → Classification

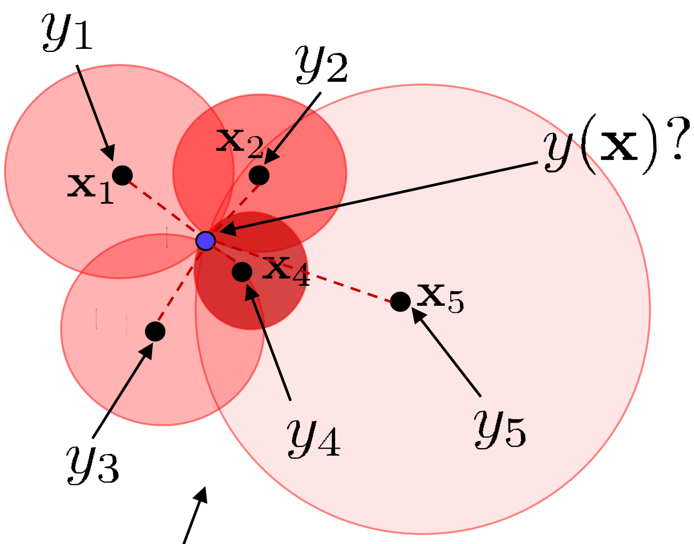
$y_i \in \mathbb{R}$ → Regression

Kernel regression (Nadaraya-Watson regression) with metric learning

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

$$K(\mathbf{x}_i, \mathbf{x}) = K\left(\frac{||\mathbf{x}_i - \mathbf{x}||}{h}\right) = \frac{1}{\sqrt{2\pi}^D h^D} \exp\left(-\frac{1}{2h^2}||\mathbf{x}_i - \mathbf{x}||^2\right)$$

$$\widehat{y}_N(\mathbf{x}) = \frac{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x})}$$



$$\mathbf{x}_i, \mathbf{x} \in \mathbb{R}^D$$

$$\widehat{y}_5(\mathbf{x}) = \frac{K(\mathbf{x}_1, \mathbf{x})}{\sum_{i=1}^{5} K(\mathbf{x}_i, \mathbf{x})} y_1 + \frac{K(\mathbf{x}_2, \mathbf{x})}{\sum_{i=1}^{5} K(\mathbf{x}_i, \mathbf{x})} y_2 + \frac{K(\mathbf{x}_3, \mathbf{x})}{\sum_{i=1}^{5} K(\mathbf{x}_i, \mathbf{x})} y_3 + \frac{K(\mathbf{x}_4, \mathbf{x})}{\sum_{i=1}^{5} K(\mathbf{x}_i, \mathbf{x})} y_4 + \frac{K(\mathbf{x}_5, \mathbf{x})}{\sum_{i=1}^{5} K(\mathbf{x}_i, \mathbf{x})} y_5$$

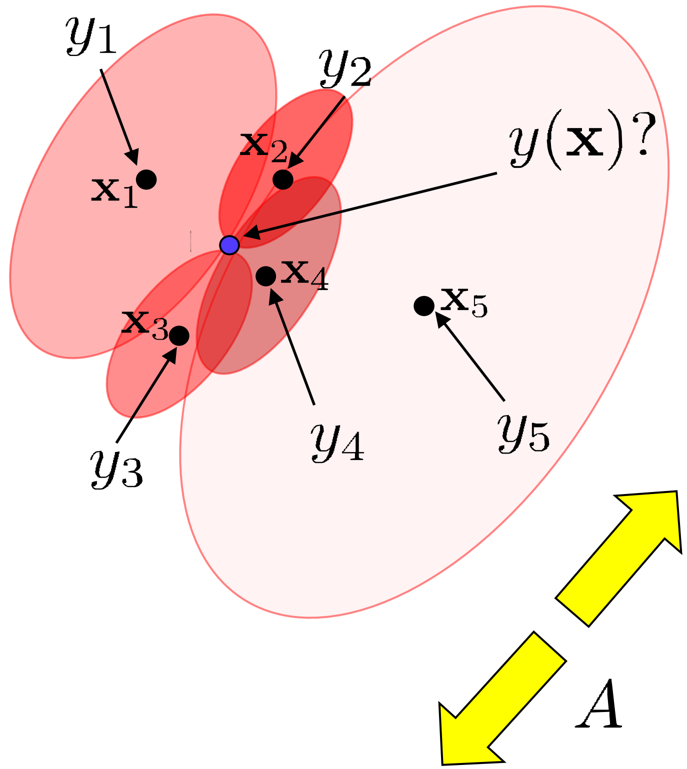
Kernel regression (Nadaraya-Watson regression) with metric learning

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$$

$$K(\mathbf{x}_i, \mathbf{x}; A) = K\left(\frac{||\mathbf{x}_i - \mathbf{x}||_A}{h}\right)$$

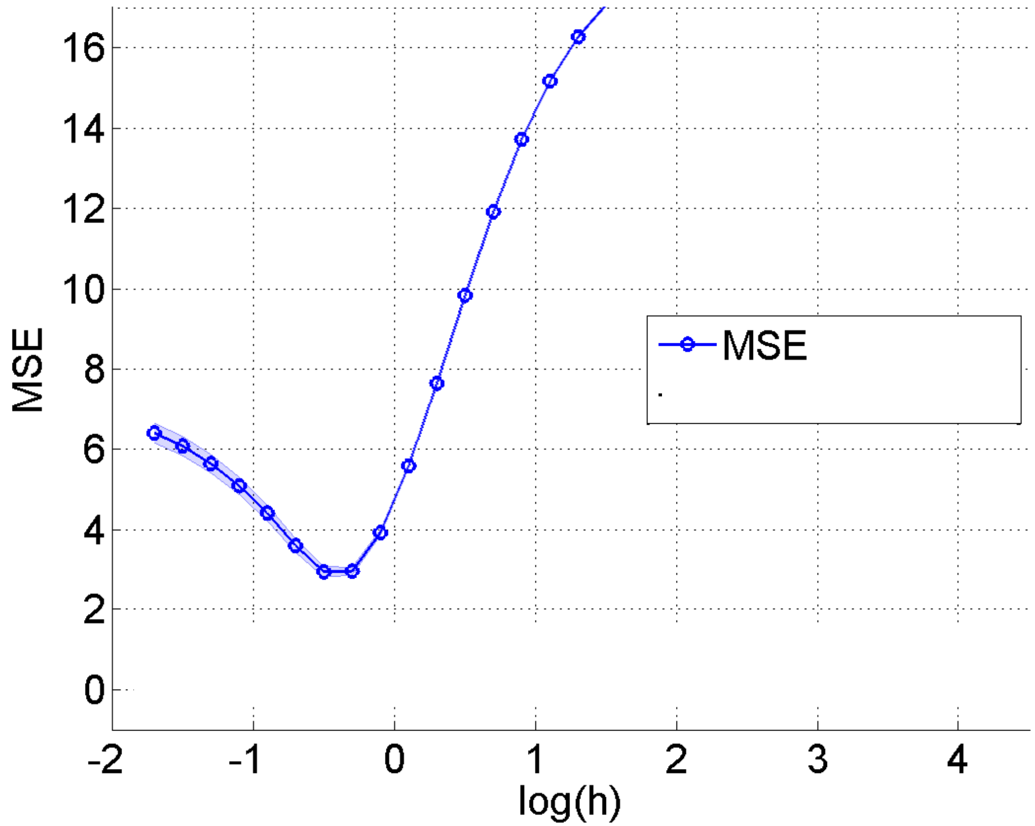
$$= \frac{1}{\sqrt{2\pi}^D h^D} \exp\left(-\frac{1}{2h^2}(\mathbf{x}_i - \mathbf{x})^\top A(\mathbf{x}_i - \mathbf{x})\right)$$

$$\widehat{y}_N(\mathbf{x}) = \frac{\sum_{i=1}^N K(\mathbf{x}_i, \mathbf{x}; A) y_i}{\sum_{i=1}^N K(\mathbf{x}_i, \mathbf{x}; A)}$$



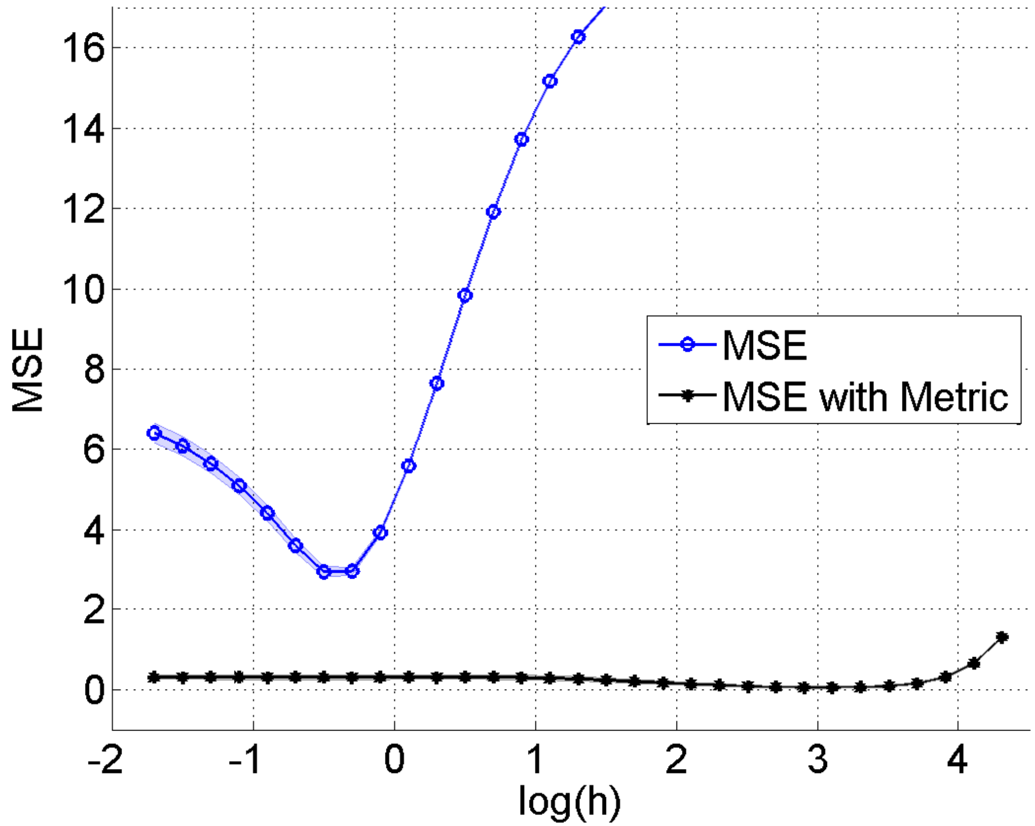
For x & y Jointly Gaussian

- Learned metric is not sensitive to the bandwidth

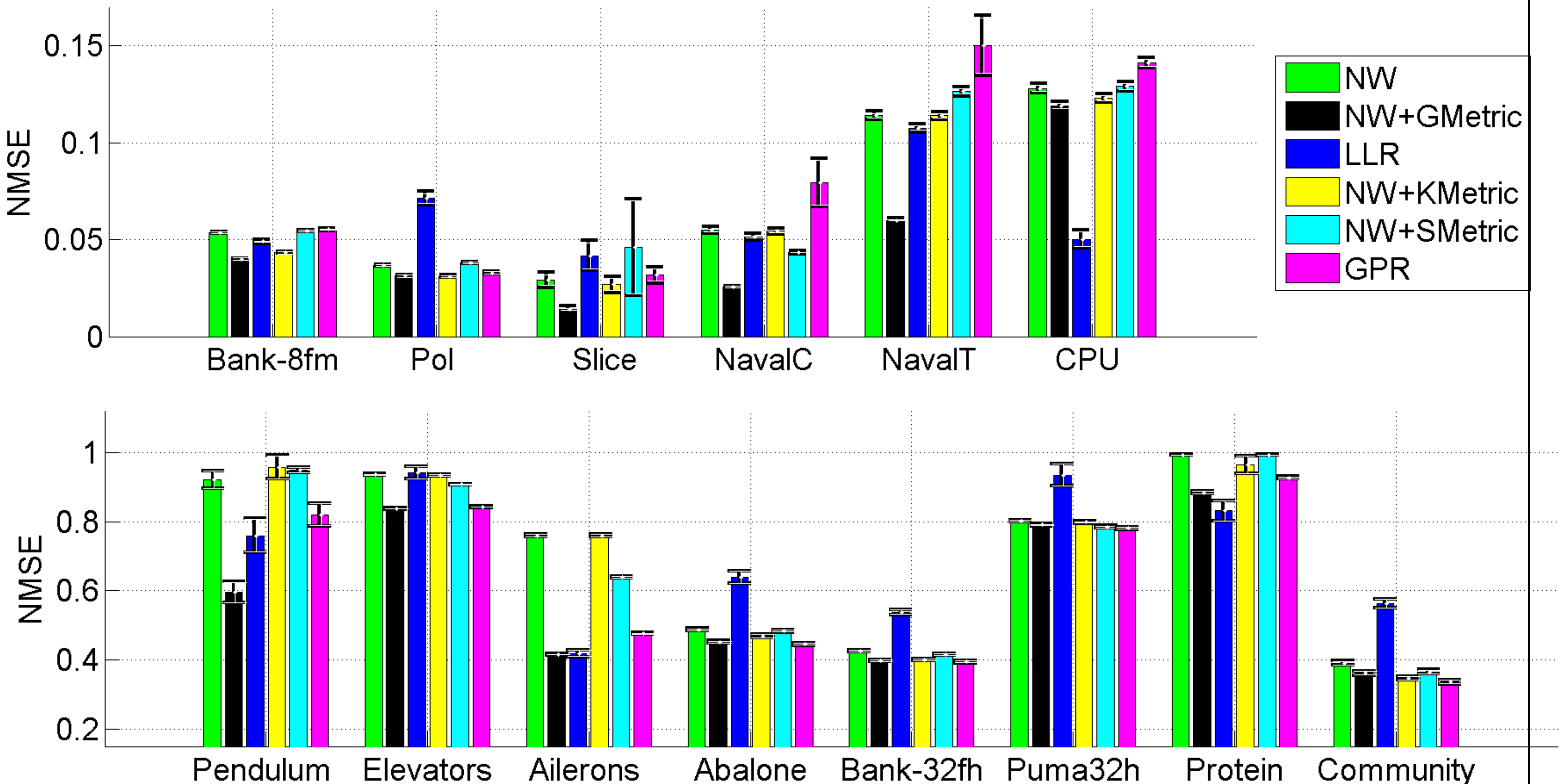


For x & y Jointly Gaussian

- Learned metric is not sensitive to the bandwidth



Benchmark Data

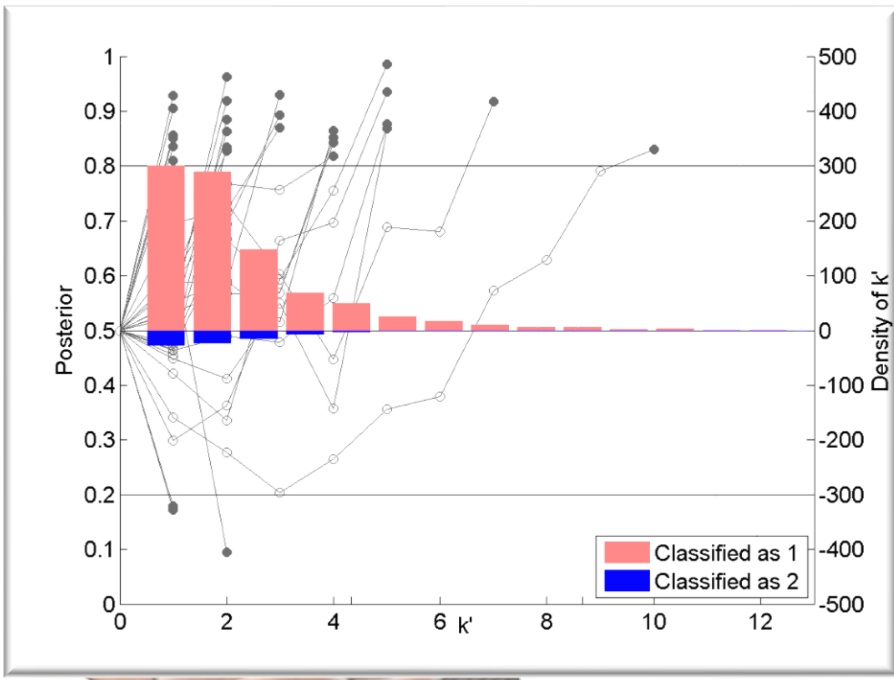


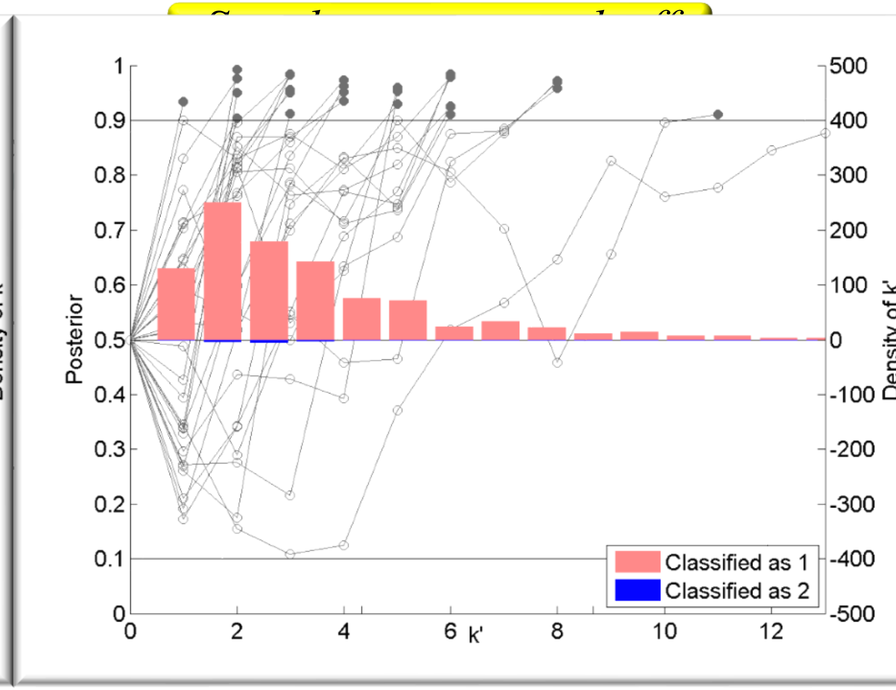
Two Theoretical Properties for Gaussians

- *The existence of a symmetric positive definite matrix A that eliminates the first term of the bias.*
- *With optimal bandwidth h minimizing the leading order terms, the minimum mean square error is the square of **bias** in infinitely high-dimensional space.*

Diffusion Decision Model

- Choosing between two alternatives under time pressure with uncertain information.





Confidence level 0.8

Confidence level 0.9

Summary

- Nearest neighbor methods and asymptotic property
- Naradaya-Watson regression with metric learning
- Diffusion decision making and nearest neighbor methods

THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr