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
2015 Organizing Committee

Total Registrations 3,755

- Tutorials (2,584)
- Conference (3,262)
- Workshops (3,006)

Pedro A. Ortega (University of Pennsylvania)

Yung-Kyun Noh (Seoul National University)

- 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
The 9th Asian Conference on Machine Learning

November 15 - 17, Yonsei University, Seoul, Korea

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 learning, 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

- Each datum is one point in a data space

\[ \mathbb{R}^D \]

\[ = [1, 2, 5, 10, \ldots]^T \]
Nearest Points
Nearest Points

automobile  truck  cat  ship  ship

ship  ship  ship  automobile  automobile
Classification with Nearest Neighbors

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

- Bayes classification using underlying density functions: *Optimal*

\[
\begin{align*}
\text{If } p_1(x) > p_2(x) & \quad \text{then classify as } p_1(x) \\quad \text{Error:} \\
\text{If } p_1(x) < p_2(x) & \quad \text{then classify as } p_2(x) \\
\end{align*}
\]

\[\frac{1}{2} \int \min[p_1, p_2] \, dx \]

Bayes risk

In general, we do not know the underlying density.
Nearest Neighbors and Bayes Classification

- Surrogate method of using underlying density functions.

\[ D = \{ x_i, y_i \}_{i=1}^{N} \sim p_1(x), p_2(x) \]

\[ p_1(x) \geq p_2(x) ? \quad \rightarrow \quad \text{Count nearest neighbors!} \]

\[ N_1 \geq N_2 ? \]
• **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.
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.
Nearest Neighbors and Bayes Risk

\[ \mathbf{x}_{NN} \rightarrow \mathbf{x}, \text{ uniformly!} \]
\[ p(\mathbf{x}) \sim p(\mathbf{x}_{NN}) \]

- **1-NN error**
  \[ \epsilon_{(k=1)} = \int \frac{p_1(\mathbf{x})p_2(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} d\mathbf{x} \leq 2E_{Bayes}(1 - E_{Bayes}) \]

- **k-NN error**
  \[ E_{Bayes} = \epsilon_{(k=\infty)} \leq \ldots \leq \epsilon_{(k=2)} \leq \epsilon_{(k=1)} \leq 2E_{Bayes}(1 - E_{Bayes}) \]
  [T. Cover and P. Hart, 1967]
Bias in the Expected Error

• Assumption:

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

$$E_{NN} \approx \int \frac{p_1(x)p_2(x)}{p_1(x) + p_2(x)} \, dx$$

$$+ \frac{1}{4D} \int \frac{\mathbb{E}_{d_N}[d_N^2|x]}{(p_1 + p_2)^2} \left[ p_2^2 \nabla^2 p_2 + p_1^2 \nabla^2 p_1 - p_1p_2 (\nabla^2 p_1 + \nabla^2 p_2) \right] \, dx \quad \cdots \quad (\text{①})$$

Metric-dependent terms

①: Asymptotic $NN$ Error

②: Residual due to *Finite Sampling*.


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

Mahalanobis-type distance:

\[ d(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)}, \quad A \succ 0 \]
Conventional Idea of Metric Learning
Many Data Situation with Overlap
Conventional Metric Learning

Performance

# Dim

NN
ITML
BM
LMNN
Fisher
Generative Local Metric Learning (GLML)

Performance vs. # Dim

20% increase

- NN
- GLML
- ITML
- BM
- LMNN
- Fisher
Bayes Classification with True Model

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

\[ p_1(x) : \mathcal{N}(\mu_1, \Sigma_1) \]
\[ p_2(x) : \mathcal{N}(\mu_2, \Sigma_2) \]
Bayes Classification with True Model

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

\[ p_1(x) : \mathcal{N}(\mu_1, \Sigma_1) \]
\[ p_2(x) : \mathcal{N}(\mu_2, \Sigma_2) \]
Bayes Classification with True Model

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

\[ p_1(x) : \mathcal{N}(\mu_1, \Sigma_1) \]
\[ p_2(x) : \mathcal{N}(\mu_2, \Sigma_2) \]
$k$-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
Manifold Embedding (Isomap)
Isomap with LMNN Metric
Isomap with GLM Metric
Nadaraya-Watson Estimator

\[
\hat{y}_N(x) = \frac{\sum_{i=1}^{N} K(x_i, x)y_i}{\sum_{j=1}^{N} K(x_j, x)}
\]

\[D = \{x_i, y_i\}_{i=1}^{N}\]

\[x_i \in \mathbb{R}^D\]

\[K(x_i, x) = K\left(\frac{|x_i - x|}{h}\right)\]

\[= \frac{1}{\sqrt{2\pi}^D h^D} \exp\left(-\frac{1}{2h^2}|x_i - x|^2\right)\]

\[y_i \in \{0, 1\} \rightarrow \text{Classification}\]

\[y_i \in \mathbb{R} \rightarrow \text{Regression}\]
Kernel regression (Nadaraya-Watson regression) with metric learning

\[ D = \{x_i, y_i\}_{i=1}^N \]

\[ K(x_i, x) = K \left( \frac{|x_i - x|}{h} \right) \]

\[ = \frac{1}{\sqrt{2\pi}^D h^D} \exp \left( -\frac{1}{2h^2} ||x_i - x||^2 \right) \]

\[ \hat{y}_N(x) = \frac{\sum_{i=1}^N K(x_i, x)y_i}{\sum_{i=1}^N K(x_i, x)} \]

\[ \hat{y}_5(x) = \frac{K(x_1, x)}{\sum_{i=1}^5 K(x_i, x)} y_1 + \frac{K(x_2, x)}{\sum_{i=1}^5 K(x_i, x)} y_2 + \frac{K(x_3, x)}{\sum_{i=1}^5 K(x_i, x)} y_3 + \frac{K(x_4, x)}{\sum_{i=1}^5 K(x_i, x)} y_4 + \frac{K(x_5, x)}{\sum_{i=1}^5 K(x_i, x)} y_5 \]
Kernel regression (Nadaraya-Watson regression) with metric learning

\[ D = \{ x_i, y_i \}_{i=1}^{N} \]

\[ K(x_i, x; A) = K \left( \frac{||x_i - x||_A}{h} \right) \]

\[ = \frac{1}{\sqrt{2\pi}^D h^D} \exp \left( -\frac{1}{2h^2} (x_i - x)^T A (x_i - x) \right) \]

\[ \hat{y}_N(x) = \frac{\sum_{i=1}^{N} K(x_i, x; A) y_i}{\sum_{i=1}^{N} K(x_i, x; A)} \]
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 $\text{bias}$ 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