Representation Learning by Kernel Autoencoders

Pierre Laforgue, Stephan Clémençon, Florence d'Alché-Buc

Télécom ParisTech (Chaire Machine Learning for Big Data)

Outline

Representation Learning

Autoencoders

Kernel Autoencoders

Experiments

Conclusion & Future Work

Representation Learning (RL)

- **A representation**: a collection of features that characterize the observation

Representation Learning (RL)

- **A representation**: a collection of features that characterize the observation
- **Raw data**: redundant, non-relevant, massive
- **Ideal data**: independent, discriminative, informative

Representation Learning (RL)

- **A representation**: a collection of features that characterize the observation
- **Raw data**: redundant, non-relevant, massive
- **Ideal data**: independent, discriminative, informative
- **Feature engineering**: implies domain experts

Representation Learning (RL)

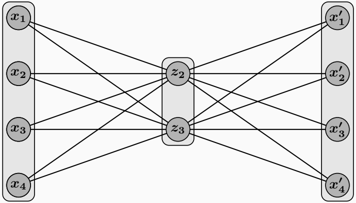
- **A representation**: a collection of features that characterize the observation
- **Raw data**: redundant, non-relevant, massive
- **Ideal data**: independent, discriminative, informative
- **Feature engineering**: implies domain experts
- **Feature/Representation learning**: automate the processus

Figure 1: Machine Learning Pipeline

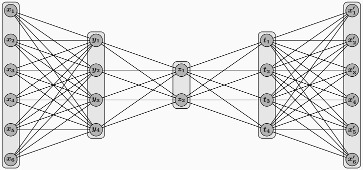
Outline

Autoencoders (AEs): Principle

- **Idea:** reconstruct the input after having compressed it
- Neural network: symmetric, hour-glass shaped
- *Self-supervised* framework



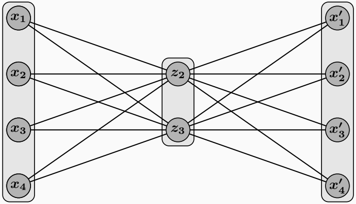
(a) 1 hidden layer AE



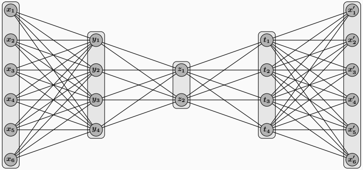
(b) 3 hidden layers AE

Autoencoders: Training

- $z = f_{\boldsymbol{W},\boldsymbol{b}}(x) = \sigma(\boldsymbol{W}x + \boldsymbol{b}) \quad x' = f_{\boldsymbol{W'},\boldsymbol{b'}}(z) = \sigma(\boldsymbol{W'}z + \boldsymbol{b'})$
- $\boldsymbol{\theta}^* = argmin_\theta \|x - x'\|^2 = argmin_\theta \left\| x - f_{\boldsymbol{W'},\boldsymbol{b'}} \circ f_{\boldsymbol{W},\boldsymbol{b}}(x) \right\|^2$
- Encoding $z = \sigma(\boldsymbol{W}^* x + \boldsymbol{b}^*)$



(c) 1 hidden layer AE



(d) 3 hidden layers AE

Outline

Representation Learning

Autoencoders

Kernel Autoencoders

Experiments

Conclusion & Future Work

Schema

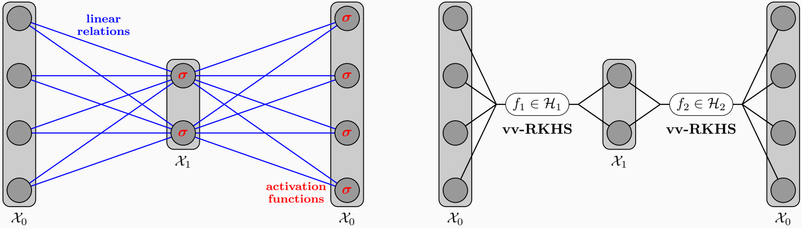


Figure 2: Standard and Kernel 2-layer Autoencoders

Formally

$$\textbf{AE :} \quad \min_{f_l \in \text{NN}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 = \mathbb{R}^d}$$

$$\textbf{KAE :} \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 \text{ Hilbert}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

Formally

$$\textbf{AE :} \quad \min_{f_l \in \mathrm{NN}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 = \mathbb{R}^d}$$

$$\textbf{KAE :} \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 \text{ Hilbert}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

1. Feasible optimization $\rightarrow$ novel RL algorithm

Formally

$$\textbf{AE}: \quad \min_{f_l \in \text{NN}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 = \mathbb{R}^d}$$

$$\textbf{KAE}: \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 \text{ Hilbert}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

1. Feasible optimization $\rightarrow$ novel RL algorithm
2. $\mathcal{X}_0$ Hilbert non necessarily Euclidean (not only $\mathbb{R}^d$)

Formally

$$\textbf{AE :} \quad \min_{f_l \in \text{NN}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 = \mathbb{R}^d}$$

$$\textbf{KAE :} \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 \text{ Hilbert}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

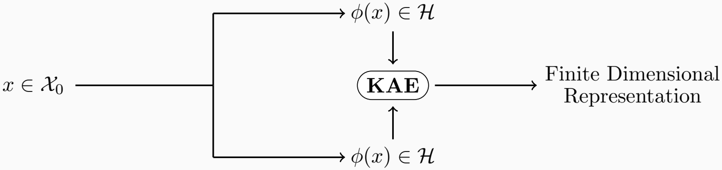
1. Feasible optimization $\rightarrow$ novel RL algorithm
2. $\mathcal{X}_0$ Hilbert non necessarily Euclidean (not only $\mathbb{R}^d$)
3. Interesting Hilbert: (kernel) feature space

Autoencoding any data

$$\textbf{AE :} \quad \min_{f_l \in \text{NN}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 = \mathbb{R}^d}$$

$$\textbf{KAE :} \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|x_i - f_L \circ \ldots \circ f_1(x_i)\|^2_{\mathcal{X}_0 \text{ Hilbert}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

$$\textbf{K}^2\textbf{AE:} \quad \min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i))\|^2_{\mathcal{H}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$



Outline

Concentric Circles

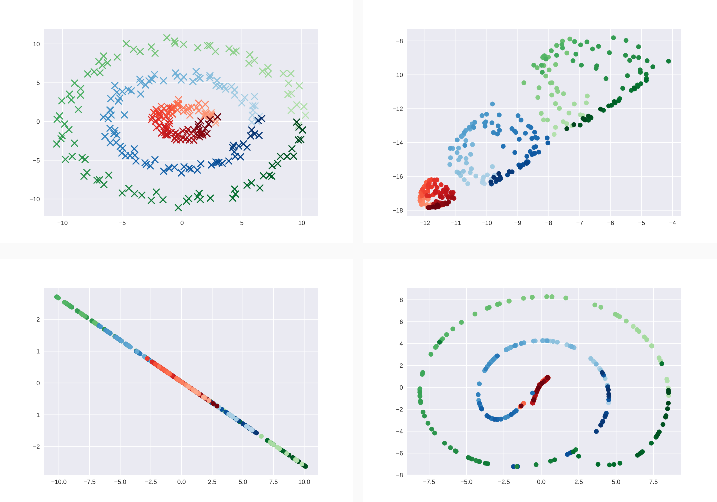


Figure 3: KAE performance on concentric circles

Molecular data

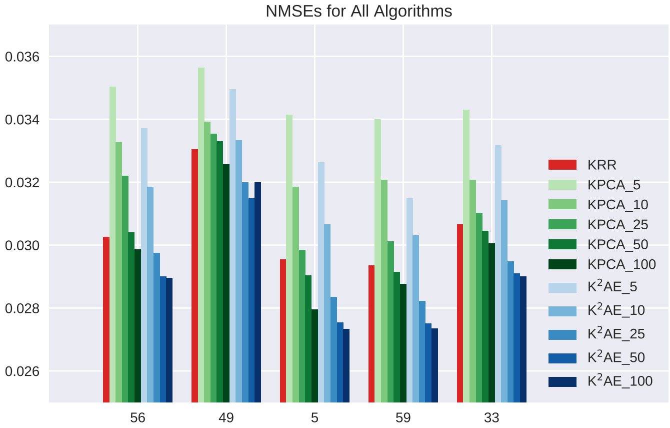


Figure 4: Performance of the different strategies on 5 cancers

Outline

Conclusion & Future Work

- Flexible tool for Representation Learning
- Advantages from AEs and Kernel Methods
- Extension of standard AEs to any type of data

- Parallel with Kernel PCA
- Combine with a supervised criterion

Preprint available at: http://arxiv.org/abs/1805.11028