Autoencoding any Data through Kernel Autoencoders

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

Télécom ParisTech (Chaire Machine Learning for Big Data)

Outline

Example: Type II diabetes occurrence prediction

A representation: a collection of features that characterizes the observation

Example: Type II diabetes occurrence prediction

A representation: a collection of features that characterizes the observation

- **Representation 1:** (PL, 42, dark brown, **green**, 175 cm, ...)
  $\rightarrow$ Diabetes occurrence prediction complex (impossible)

Example: Type II diabetes occurrence prediction

A representation: a collection of features that characterizes the observation

- **Representation 1:** (PL, 42, dark brown, **green**, 175 cm, ...)
  $\rightarrow$ Diabetes occurrence prediction complex (impossible)
- **Representation 2:** (175 cm, 62 kg, 25 years old, ♂, ...)
  $\rightarrow$ Diabetes occurrence prediction possible

Example: Type II diabetes occurrence prediction

A representation: a collection of features that characterizes the observation

- **Representation 1:** (PL, 42, dark brown, **green**, 175 cm, ...)
  $\rightarrow$ Diabetes occurrence prediction complex (impossible)
- **Representation 2:** (175 cm, 62 kg, 25 years old, ♂, ...)
  $\rightarrow$ Diabetes occurrence prediction possible
- **Representation 3:** (BMI=20.24, family background, ...)
  $\rightarrow$ Diabetes occurrence prediction facilitated

Representation Learning

- **Feature engineering**: implies domain experts
- **Feature learning** / **Representation learning**: automate the processus

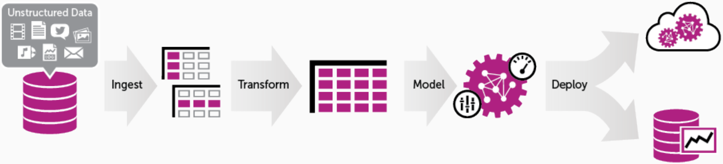
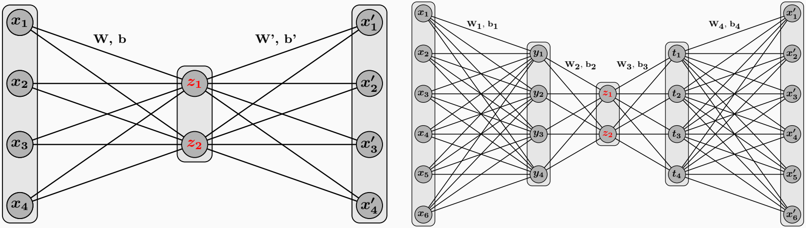


Figure 1: Machine Learning Pipeline

Outline

Autoencoders (AEs): Principle

- **Idea:** compress and reconstruct input by a Neural Net (NN)
- Elementary mapping: $f : [0,1]^d \to [0,1]^p$ such that
$$f(x) = \sigma(Wx + b),\ W \in \mathbb{R}^{p\times d}, b \in \mathbb{R}^p$$
- **NN:** composition of such mappings. $y = f_L \circ \ldots \circ f_1(x)$
- **AE:** output x' must match input x

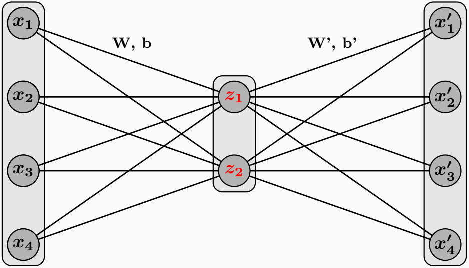


(a) 1 hidden layer AE

(b) 3 hidden layers AE

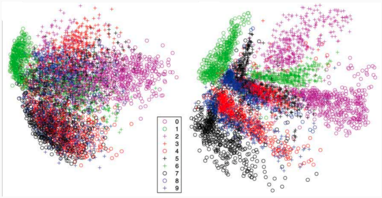
Autoencoders: Training

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

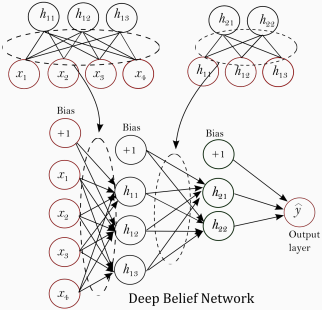


Autoencoders: Uses

- Data compression (PCA) [*Bourlard 1988, Hinton 2006*]
- Pre-training of neural networks [*Bengio & al. 2007*]
- Denoising [*Vincent, Larochelle & al. 2010*]



(c) PCA / AE

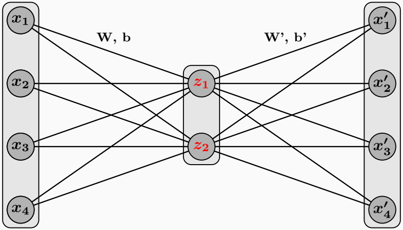


(d) Pre-training by AE

Autoencoders: Summary

❶ $$\min_{f_l \in \textcolor{red}{\mathrm{NN}_{\mathrm{em}}}} \quad \frac{1}{n}\sum_{i=1}^{n} \left\| x_i - f_L \circ \ldots \circ f_1(x_i) \right\|^2$$

❷ $x_i \in [0,1]^d$ or $x_i \in \mathbb{R}^d$



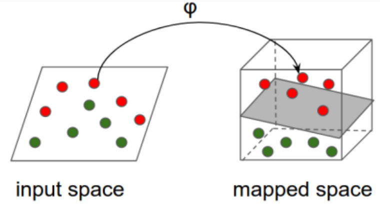
Outline

Kernel Methods: Definitions

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- $\forall (x, x') \in \mathcal{X} \times \mathcal{X}, \quad k(x, x') = k(x', x)$ (symmetry)
- $\sum_{i,j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j = \alpha^T K \alpha \geq 0$ (positiveness)

- $\exists \mathcal{H}_k$ Hilbert, $\varphi : \mathcal{X} \to \mathcal{H}_k, \quad k(x, x') = \left\langle \varphi(x), \varphi(x') \right\rangle_{\mathcal{H}_k}$
- $\mathcal{H}_k = \overline{Span\,\{\varphi(x),\ x \in \mathcal{X}\}}$ (RKHS)

Kernel Methods: Kernelization of the Ridge Regression



$X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$

- $\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 + 2n\lambda\|\beta\|^2$
- $\min_{\beta \in \mathbb{R}^p} \sum_i (y_i - \langle x_i, \beta \rangle_{\mathbb{R}^p})^2 + 2n\lambda\|\beta\|^2_{\mathbb{R}^p}$

- $\min_{\omega \in \mathcal{H}_k} \sum_i (y_i - \langle \varphi(x_i), \omega \rangle_{\mathcal{H}_k})^2 + 2n\lambda\|\omega\|^2_{\mathcal{H}_k}$ $\quad \omega^* = \sum_j \varphi(x_j)\alpha_j^*$
- $\min_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + 2n\lambda\alpha^T K\alpha$

Kernel Methods: Summary

❶ $$\min_{f_l \in \textcolor{red}{\mathrm{NN}_{\mathrm{em}}}} \quad \frac{1}{n}\sum_{i=1}^{n} \left\| x_i - f_L \circ \ldots \circ f_1(x_i) \right\|^2$$

❷ $x_i \in [0,1]^d$ or $x_i \in \mathbb{R}^d$

Kernelization of a problem: $x \longleftrightarrow \varphi(x)$

❸ is computable as long as dot products only are involved

❹ allows to deal with non-vectorial data

Outline

Representation Learning

Autoencoders

Kernel Methods

Kernel Autoencoders

Experiments

Conclusion & Future Work

Schema

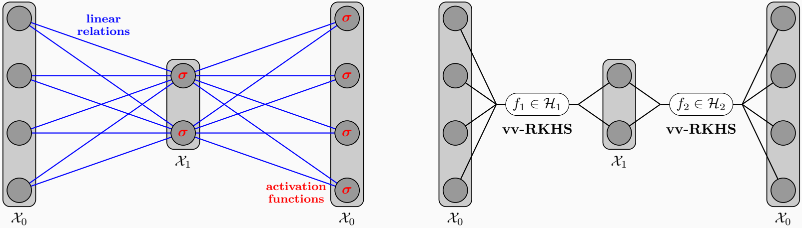


Figure 2: Standard and Kernel 2-layer Autoencoders

Formally

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

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

Formally

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

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

1. Novel algorithm of Representation Learning

Formally

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

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

1. Novel algorithm of Representation Learning
2. $\mathcal{X}_0$ Hilbert non necessarily Euclidean (not only $\mathbb{R}^d$)

Formally

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

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

1. Novel algorithm of Representation Learning
2. $\mathcal{X}_0$ Hilbert non necessarily Euclidean (not only $\mathbb{R}^d$)
3. Interesting Hilbert: (kernel) feature space

Autoencoding any data

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

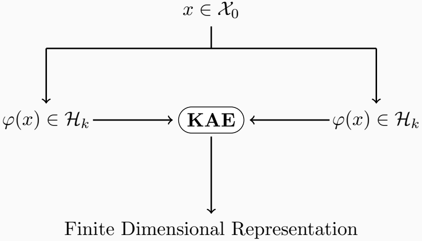


Figure 3: Autoencoding on any $\mathcal{X}_0$

Outline

Concentric Circles

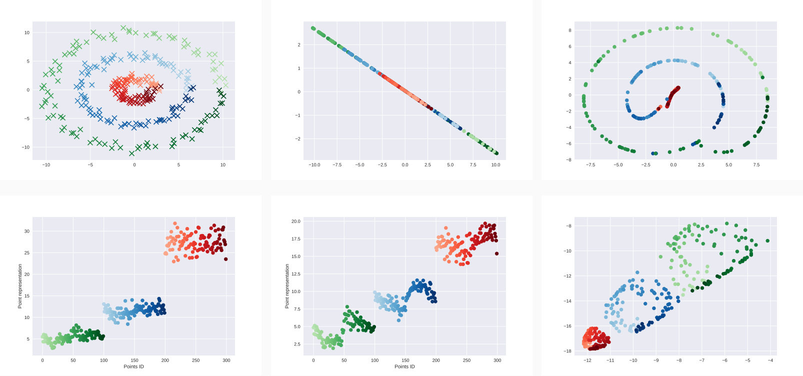


Figure 4: KAE performance on concentric circles

Molecular data (graph)

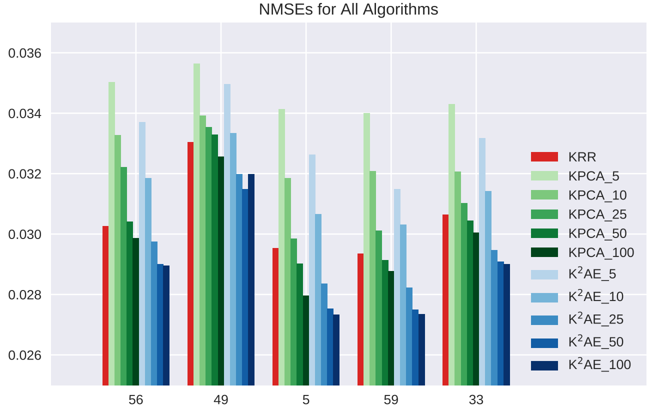


Figure 5: Performance of the different strategies on 8 cancer

Outline

Representation Learning

Autoencoders

Kernel Methods

Kernel Autoencoders

Experiments

Conclusion & Future Work

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
- Consider another loss / optimization strategy

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