Autoencoding any Data through Kernel Autoencoders

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

Télécom ParisTech (Chaire Machine Learning for Big Data)

Outline

Representation Learning

Autoencoders

Kernel Methods

Kernel Autoencoders

Experiments

Conclusion & Future Work

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

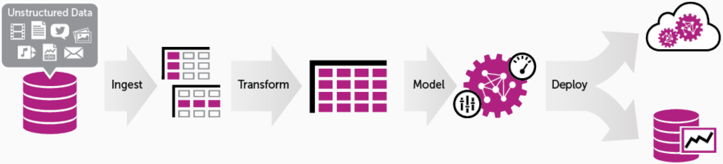
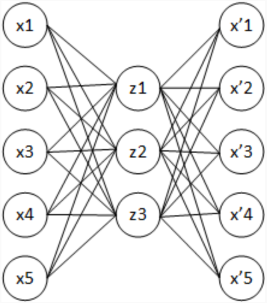


Figure 1: Machine Learning Pipeline

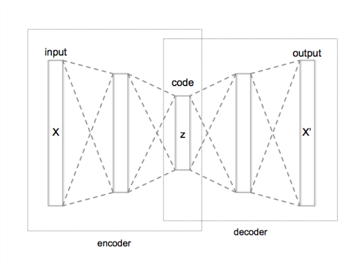
Outline

Autoencoders (AEs): Principle

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



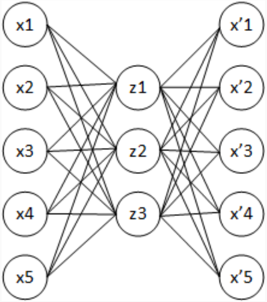
(a) 1 hidden layer AE

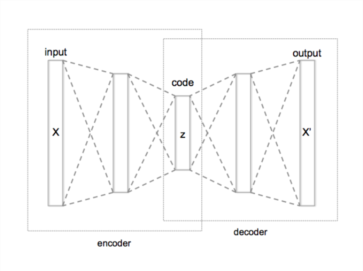


(b) 3 hidden layers AE

Autoencoders: Training

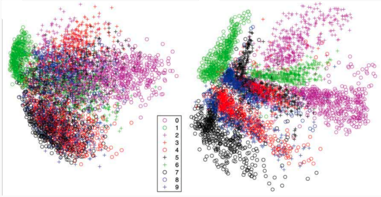
- $z = f_{\boldsymbol{W},\boldsymbol{b}}(x) = \sigma(\boldsymbol{W}x + \boldsymbol{b}) \quad x' = f_{\boldsymbol{W'},\boldsymbol{b'}}(z) = \sigma(\boldsymbol{W'}z + \boldsymbol{b'})$
- $\boldsymbol{\theta}^* = argmin_\theta \|x - x'\|^2 = argmin_\theta \left\| x - f_{\boldsymbol{W'},\boldsymbol{b'}}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



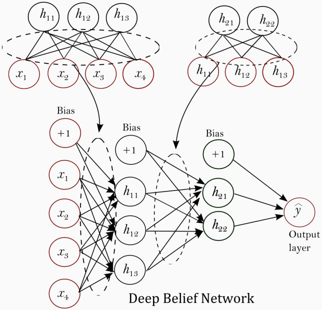


Autoencoders: Uses

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



(e) PCA / AE



(f) Pre-training by AE

Outline

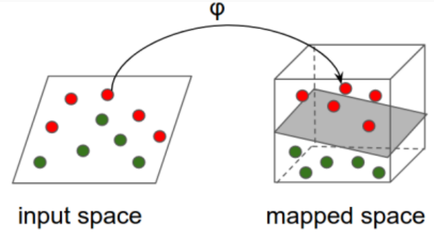
Definitions: positive definite (scalar) kernel

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- $\forall (x, x') \in \mathcal{X} \times \mathcal{X}, \quad k(x, x') = k(x', x)$ (symmetry)
- $\sum_{i,j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j = \alpha^T K \alpha \geq 0$ (positiveness)
- $\exists \phi, \mathcal{H}_k$ Hilbert, $\quad k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k} \quad \phi(x) = k(\cdot, x)$
- $\mathcal{H}_k = \overline{Span\{k(\cdot, x),\ x \in \mathcal{X}\}}$ (RKHS)

- $f^* \in \underset{f \in \mathcal{H}_k}{\operatorname{argmin}}\ V(f(x_1), \ldots, f(x_n), \|f\|), \qquad f^* = \sum_{i=1}^{n} k(\cdot, x_i) \alpha_i$

  $V = \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}_k}^2$ (representer theorem)

Kernelization of the Ridge regression



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

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

Kernelization of other problems

Kernel K-Means

- $c_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} x_i \in \mathbb{R}^p$ $\qquad \tilde{c}_k = \frac{1}{|\widetilde{\mathcal{C}}_k|} \sum_{i \in \widetilde{\mathcal{C}}_k} \phi(x_i) \in \mathcal{H}_k$
- $\min_{k \leq K} \|x_i - c_k\|^2_{\mathbb{R}^p}$ $\qquad \min_{k \leq K} \|\phi(x_i) - \tilde{c}_k\|^2_{\mathcal{H}_k}$

Kernel PCA

- Standard PCA: solve eigenproblem $\left(X^T X\right) u = \lambda u$
- Kernel PCA: solve eigenproblem $\left(\phi(X)^T \phi(X)\right) u = \lambda u$
- KPCA: equivalent to solve $\left(\phi(X) \phi(X)^T\right) u := Ku = \lambda u$

Outline

Schema

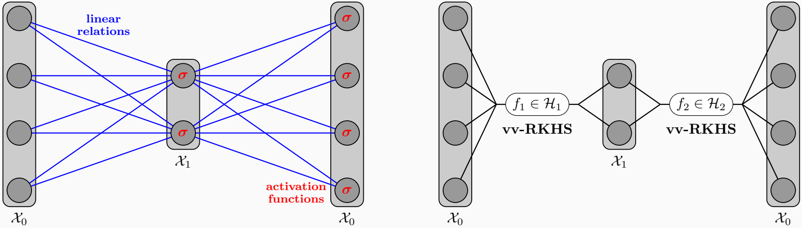


Figure 2: Standard and Kernel 2-layer Autoencoders

Formally

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

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

Formally

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

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

1. Novel algorithm of Representation Learning

Formally

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

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

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

Formally

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

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

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

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

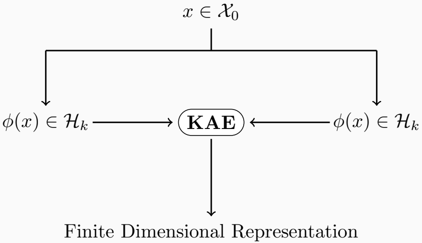


Figure 3: Autoencoding on any $\mathcal{X}_0$

Outline

Concentric Circles

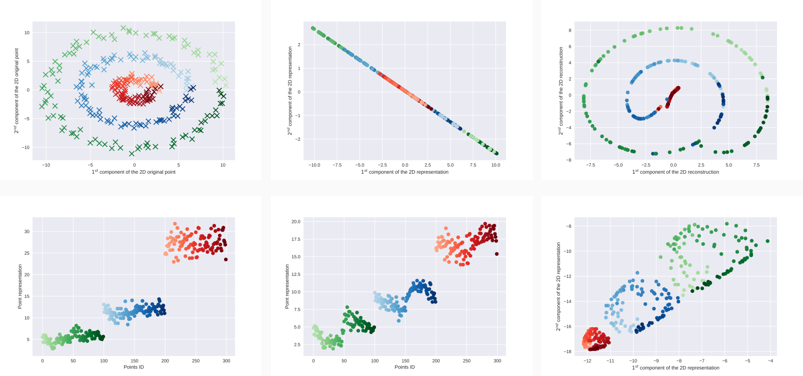


Figure 4: KAE performance on concentric circles

Molecular data (table)

Table 1: NMSEs on Molecular Activity for Different Types of Cancer

	KRR	KPCA 10 + RF	KPCA 50 + RF	K^2AE 10 + RF	K^2AE 50 + RF
Cancer 01	0.02978	0.03035	0.03035	0.03097	**0.02808**
Cancer 02	0.03004	0.02978	0.02978	0.03099	**0.02775**
Cancer 03	0.02878	0.02914	0.02914	0.02989	**0.02709**
Cancer 04	0.03003	0.03074	0.03074	0.03218	**0.02924**
Cancer 05	0.02954	0.02903	0.02903	0.03065	**0.02754**
Cancer 06	0.02914	0.03083	0.03083	0.03134	**0.02838**
Cancer 07	0.03113	0.03207	0.03207	0.03257	**0.03018**
Cancer 08	0.02899	0.02898	0.02898	0.03065	**0.02770**

Molecular data (graph)

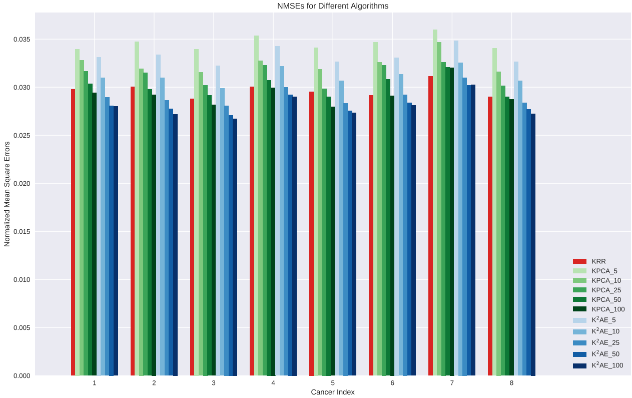


Figure 5: Performance of the different strategies on 8 cancer

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