KDD2020- Workshop

Hands On: From Autoencoder to Robust Autoencoder

Raghav Chalapathy
Nguyen Lu Dang Khoa
Sanjay Chawla

Hands On

Summary of Dataset used in HandsOn

- **Restaurant**, comprising video background modeling and activity detection consisting of snapshots of restaurant activities.

Dataset	# instances	# anomalies	# features
`restaurant`	200	Unknown (foreground)	19200

Evaluation Methodology

- Predictive performance is measured against the ground truth anomaly labels **using three standard metrics:**

Evaluation Methodology

- Predictive performance is measured against the ground truth anomaly labels **using three standard metrics:**
  - **Area under the precision-recall curve (AUPRC)**
  - **Area under the ROC curve (AUROC))**
  - **Precision at 10 (P@10)**

Evaluation Methodology

- Predictive performance is measured against the ground truth anomaly labels **using three standard metrics:**
  - **Area under the precision-recall curve (AUPRC)**
  - **Area under the ROC curve (AUROC))**
  - **Precision at 10 (P@10)**
- **AUPRC and AUROC** measure ranking performance.

Evaluation Methodology

- Predictive performance is measured against the ground truth anomaly labels **using three standard metrics:**
  - **Area under the precision-recall curve (AUPRC)**
  - **Area under the ROC curve (AUROC))**
  - **Precision at 10 (P@10)**
- **AUPRC and AUROC** measure ranking performance.
- **P@10** measures classification performance.(actual anomalies among top-10 scored instances).

Auto-encoders for anomaly detection.

- Auto encoder with single **hidden layer**

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}} \tag{1}$$

- $\hat{\mathbf{X}} = f(\mathbf{XU})\mathbf{V}$ is **reconstruction error measure.**

Auto-encoders for anomaly detection.

- Auto encoder with single **hidden layer**

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}} \tag{1}$$

- $\hat{\mathbf{X}} = f(\mathbf{X}\mathbf{U})\mathbf{V}$ is **reconstruction error measure.**
  - $\mathbf{U}$: Weights (input $\rightarrow$ hidden),

Auto-encoders for anomaly detection.

- Auto encoder with single **hidden layer**

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}} \tag{1}$$

- $\hat{\mathbf{X}} = f(\mathbf{X}\mathbf{U})\mathbf{V}$ is **reconstruction error measure.**
  - $\mathbf{U}$: Weights (input $\rightarrow$ hidden),
  - $\mathbf{V}$: Weights (hidden $\rightarrow$ output),

- **XU** projects $\mathbf{X}$ into K dimensional space $\mathbf{U} \in \mathbf{R}^{\mathbf{D}\times\mathbf{K}}$, $\mathbf{V} \in \mathbf{R}^{\mathbf{K}\times\mathbf{D}}$.

Auto-encoders for anomaly detection.

- Auto encoder with single **hidden layer**

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}} \tag{1}$$

- $\hat{\mathbf{X}} = f(\mathbf{X}\mathbf{U})\mathbf{V}$ is **reconstruction error measure.**
  - $\mathbf{U}$: Weights (input $\rightarrow$ hidden),
  - $\mathbf{V}$: Weights (hidden $\rightarrow$ output),
  - activation function: $f : \mathbb{R} \rightarrow \mathbb{R}$
- **XU** projects $\mathbf{X}$ into K dimensional space $\mathbf{U} \in \mathbf{R}^{\mathbf{D}\times\mathbf{K}}$, $\mathbf{V} \in \mathbf{R}^{\mathbf{K}\times\mathbf{D}}$.
- Non linear projection: **sigmoid** $f(\cdot) := (1 + \exp(-a))^{-1}$

Code Autoencoder Loss Function (10 Mins)

Robust (convolution) Auto-Encoders [RCAE]

- For activation function $f : \mathbb{R} \to \mathbb{R}$

$$\min_{\mathbf{U},\mathbf{V},\mathbf{N}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda\|\mathbf{N}\|_{\mathbf{1}} \tag{2}$$

Robust (convolution) Auto-Encoders [RCAE]

- For activation function $f : \mathbb{R} \to \mathbb{R}$

$$\min_{\mathbf{U},\mathbf{V},\mathbf{N}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \tag{2}$$

- $\hat{X} = f(\mathbf{X}\mathbf{U})\mathbf{V}$ is **reconstruction error measure**

Robust (convolution) Auto-Encoders [RCAE]

- For activation function $f : \mathbb{R} \to \mathbb{R}$

$$\min_{\mathbf{U},\mathbf{V},\mathbf{N}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \tag{2}$$

- $\hat{X} = f(\mathbf{X}\mathbf{U})\mathbf{V}$ is **reconstruction error measure**
- $\lambda, \mu > 0$ are tuning parameters.

Robust (convolution) Auto-Encoders [RCAE]

- For activation function $f : \mathbb{R} \rightarrow \mathbb{R}$

$$\min_{\mathbf{U},\mathbf{V},\mathbf{N}} \|\mathbf{X} - \mathbf{f}(\mathbf{X}\mathbf{U})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \tag{2}$$

- $\hat{X} = f(\mathbf{XU})\mathbf{V}$ is **reconstruction error measure**
- $\lambda, \mu > 0$ are tuning parameters.
- $\mathbf{Z} = \mathbf{f}(\mathbf{XU})$ **latent representation decoded by V weights.**
- $\mathbf{N}$ captures **gross outliers.**

Robust (convolution) Auto-Encoders [RCAE]

- For activation function $f : \mathbb{R} \to \mathbb{R}$

$$\min_{\mathbf{U},\mathbf{V},\mathbf{N}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda\|\mathbf{N}\|_{\mathbf{1}} \tag{2}$$

- $\hat{X} = f(\mathbf{XU})\mathbf{V}$ is **reconstruction error measure**
- $\lambda, \mu > 0$ are tuning parameters.
- $\mathbf{Z} = \mathbf{f}(\mathbf{XU})$ **latent representation decoded by V weights.**
- $\mathbf{N}$ captures **gross outliers.**
- $0 < \lambda < +\infty$, models a standard auto encoder robust to noise.

Training RCAE (1)

- Consider the Objective function of **Robust Autoencoder**.

$$\min_{\mathbf{U,V,N}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \quad (3)$$

Training RCAE (1)

- Consider the Objective function of **Robust Autoencoder**.

$$\min_{\mathbf{U,V,N}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \quad (3)$$

- More generally objective could be rewritten as below.

$$\min_{\theta,\mathbf{N}} \|\mathbf{X} - \hat{\mathbf{X}}(\theta) + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot \mathbf{\Omega}(\theta) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \quad (4)$$

Training RCAE (1)

- Consider the Objective function of **Robust Autoencoder**.

$$\min_{\mathbf{U,V,N}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda\|\mathbf{N}\|_{\mathbf{1}} \quad (3)$$

- More generally objective could be rewritten as below.

$$\min_{\theta,\mathbf{N}} \|\mathbf{X} - \hat{\mathbf{X}}(\theta) + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot \mathbf{\Omega}(\theta) + \lambda\|\mathbf{N}\|_{\mathbf{1}} \quad (4)$$

- where $\hat{\mathbf{X}}(\theta)$ is some generic predictor with parameters θ.
- $\mathbf{\Omega}(\cdot)$: regularization function.

Training RCAE (1)

- Consider the Objective function of **Robust Autoencoder**.

$$\min_{\mathbf{U,V,N}} \|\mathbf{X} - \mathbf{f}(\mathbf{XU})\mathbf{V} + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot (\|\mathbf{U}\|_{\mathbf{F}}^{\mathbf{2}} + \|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \quad (3)$$

- More generally objective could be rewritten as below.

$$\min_{\theta,\mathbf{N}} \|\mathbf{X} - \hat{\mathbf{X}}(\theta) + \mathbf{N}\|_{\mathbf{F}}^{\mathbf{2}} + \frac{\mu}{\mathbf{2}} \cdot \mathbf{\Omega}(\theta) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \quad (4)$$

- where $\hat{\mathbf{X}}(\theta)$ is some generic predictor with parameters θ.
- $\mathbf{\Omega}(\cdot)$: regularization function.
- Equation 7 is **non-convex but unconstrained and sub-differentiable**

Training RCAE (2)

- For differentiable function $\hat{\mathbf{X}}(\theta)$ back-propagation is employed.

[1] Bach, F., Jenatton, R., Mairal, J., Obozinski, G. Convex Optimization with Sparsity-Inducing Norms.

Training RCAE (2)

- For differentiable function $\hat{\mathbf{X}}(\theta)$ back-propagation is employed.
- We follow **soft thresholding approach** to optimize $\mathbf{N}$.

[1] Bach, F., Jenatton, R., Mairal, J., Obozinski, G. Convex Optimization with Sparsity-Inducing Norms.

Training RCAE (2)

- For differentiable function $\hat{\mathbf{X}}(\theta)$ back-propagation is employed.
- We follow **soft thresholding approach** to optimize $\mathbf{N}$.
- For fixed $\mathbf{N}$,θ,$\mathbf{U}$, $\mathbf{V}$ the objective is:

$$\min_{\theta,\mathbf{N}} \|\mathbf{N} - (\mathbf{X} - \hat{\mathbf{X}}(\theta)) + \lambda\|\mathbf{N}\|_{\mathbf{1}} \tag{5}$$

[1] Bach, F., Jenatton, R., Mairal, J., Obozinski, G. Convex Optimization with Sparsity-Inducing Norms.

Training RCAE (2)

- For differentiable function $\hat{\mathbf{X}}(\theta)$ back-propagation is employed.
- We follow **soft thresholding approach** to optimize $\mathbf{N}$.
- For fixed $\mathbf{N}$,θ,$\mathbf{U}$, $\mathbf{V}$ the objective is:

$$\min_{\theta,\mathbf{N}} \|\mathbf{N} - (\mathbf{X} - \hat{\mathbf{X}}(\theta)) + \lambda \|\mathbf{N}\|_{\mathbf{1}} \tag{5}$$

- Applying **Soft-thresholding**[1] to compute **N**

$$N_{ij} = \begin{cases} (\mathbf{X} - \hat{\mathbf{X}}(\theta))_{ij} - \frac{\lambda}{2} & \text{if } (\mathbf{X} - \hat{\mathbf{X}}(\theta))_{ij} > \frac{\lambda}{2} \\ (\mathbf{X} - \hat{\mathbf{X}}(\theta))_{ij} + \frac{\lambda}{2} & \text{if } (\mathbf{X} - \hat{\mathbf{X}}(\theta))_{ij} < -\frac{\lambda}{2} \\ 0 & \text{else.} \end{cases} \tag{6}$$

[1] Bach, F., Jenatton, R., Mairal, J., Obozinski, G. Convex Optimization with Sparsity-Inducing Norms.

Code Robust Autoencoder Loss Function (10 Mins)

Conclusion

- Extended robust PCA model **to the nonlinear autoencoder setting**.

Conclusion

- Extended robust PCA model **to the nonlinear autoencoder setting**.
- Our approach is **robust, deep and inductive.**

Conclusion

- Extended robust PCA model **to the nonlinear autoencoder setting**.
- Our approach is **robust, deep and inductive.**
- Not oversensitive besides captures **subtle anomalies**.

Conclusion

- Extended robust PCA model **to the nonlinear autoencoder setting**.
- Our approach is **robust, deep and inductive.**
- Not oversensitive besides captures **subtle anomalies**.
- Extend deep autoencoders for **outlier description**.

References

- [1] Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
- [2] Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
- [5]Chalapathy, Raghavendra, Aditya Krishna Menon, and Sanjay Chawla. Robust, Deep and Inductive Anomaly Detection.arXiv:1704.06743 (2017).


&
Q
A