Segment-based Credit Scoring Using Latent Clusters in the Variational Autoencoder

Joint work with: Michael Kampffmeyer, Kjersti Aas, Robert Jenssen

Ph.D. candidate Rogelio Andrade Mancisidor
Machine Learning Group
http://machine-learning.uit.no
UiT The Artic University of Norway

Modelling with Big Data and Machine Learning
Bank of England
26-27 November, 2018

Outline

1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Outline

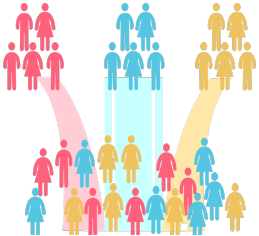
1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Credit Scoring

- Credit scoring estimates the probability of default - used for underwriting and pricing.
- Understanding customers and target populations can improve the accuracy in credit scoring.

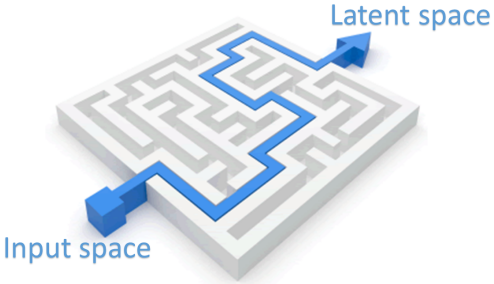
Segmentation

- Two approaches:
  1. Heuristic
  2. Statistical
- Segments can be used for marketing, customer, data, process and **model fit** (classification) purposes.



Feature Learning

- Learns transformations that make it easier to extract useful information.



- We use the Variational Autoencoder (VAE) to learn a data representation, in the latent space, which encapsulates different risk profiles.

Outline

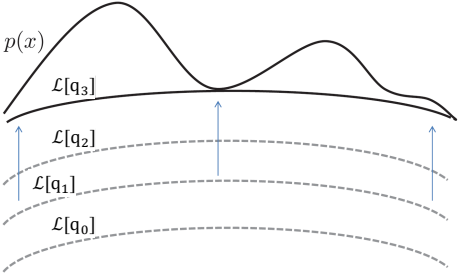
1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Variational Inference - Intuition

- We want to model

$$p(x) = \int p(x, z) dz$$

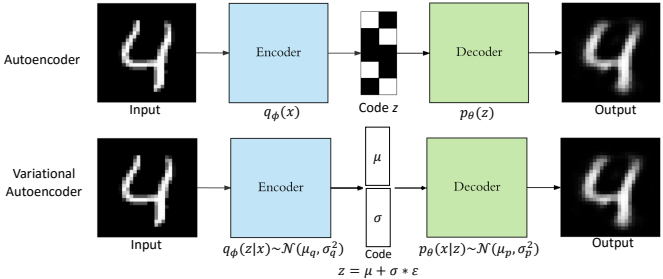
- Variational Inference (IV) optimizes a lower bound on $p(x)$, and approximates the posterior distribution $p(z|x)$ with $q(z|x)$.



Bevan Jones - ML for NLP Reading Group

Autoencoder and Variational Autoencoder

- Autoencoder (AE) and Variational Autoencoder (VAE) architectures



- The lower bound in the VAE has two important terms:

$$\mathcal{L}(x;\theta,\phi) = E_q[log\ \ p_\theta(x|z)] - KL[q_\phi(z|x)||p_\theta(z)]$$
$$= Reconstruction - Regularizer$$

Amortized Inference

- Note that in the mean-field approach we assume:

$$q(z) = \prod_i q_i(z_i)$$

- In contrast, the optimization of the VAE lower bound is wrt $\{\phi, \theta\}$ - weights (and bias) in the feed forward networks.
- This is called amortized inference.
- Hence, VAEs are suitable for large-scale problems.

Outline

1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Feature Learning - Intuition

- The term $E_q[log\ p_\theta(x|z)]$ encourages densities that place their mass on latent variables that can reconstruct the input data.

$$\mathcal{L}(\theta, \phi, x) = E_q[log\ p_\theta(x|z)] - KL[q_\phi(z|x)||p_\theta(z)]$$

- Given that the VAE has converged to q^* (the optimal variational density)
  1. The latent space encapsulates a useful representation of the data - feature learning
  2. Otherwise, the reconstruction would have failed and the algorithm would have not converged to q^*.

Data Transformation

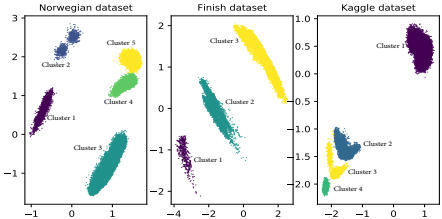
- We can steer latent configurations by transforming the input data.
- Transform the data into weight of evidence (WoE) codes - it is a measure of risk!

$$WoE = log \frac{Pr(x_j \in B_k | y = 1)}{Pr(x_j \in B_k | y = 0)},$$

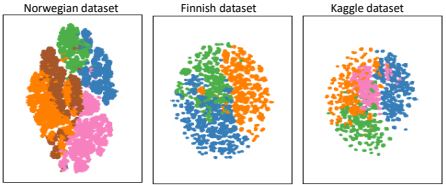
where $y = 1$ means 90+ days past due, else $y = 0$.

Visualization

- VAE data representation, well-defined clusters.



- t-SNE data representation with k-means clusters.



Clusters Risk Profiles

- The default rate (dr) is

$$dr = \frac{\sum_i^n y_i = 1}{n},$$

Cluster ID	Norwegian dataset		
	Default rate	Nr. Customers	Nr. $y = 1$
1	5.30%	2 206	117
2	11.24%	774	87
3	1.39 %	109 969	1539
4	2.89%	11 450	331
5	5.30%	9 106	483

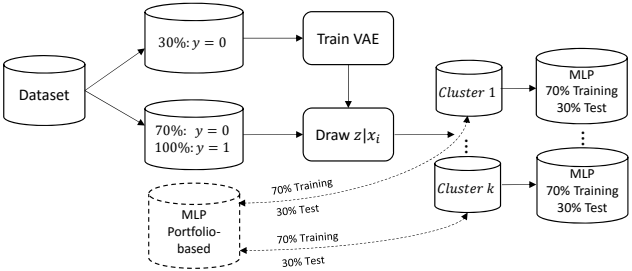
Cluster ID	Finnish dataset			Kaggle dataset		
	Default rate	Nr. Customers	Nr. $y = 1$	Default rate	Nr. Customers	Nr. $y = 1$
1	5.93%	438	26	5.47%	97 434	5 327
2	3.76%	6 067	228	33.13%	6 121	2 028
3	0.91%	75 536	685	55.68%	2 026	1 128
4				63.58%	2 427	1 543

Outline

1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Graphical representation

- Segment-based credit scoring - model development -



Model performance results

- Segment-based vs Portfolio-based approach

Norwegian car loan				
Performance metric	Cluster	Segment-based	Portfolio-based	*p*-value
H-measure	1	0.2774	0.2310	0.0509
	2	0.1665	0.1453	0.4299
	3	0.2174	0.2076	0.0854
	4	0.1760	0.1471	0.0186
	5	0.1302	0.1193	0.0931

Performance metric	Cluster	**Kaggle**			**Finish car loan**		
		Segment-based	Portfolio-based	*p*-value	Segment-based	Portfolio-based	*p*-value
H-measure	1	0.2918	0.2842	0.0410	0.1949	0.2388	0.4141
	2	0.1145	0.1067	0.2230	0.0946	0.1001	0.5870
	3	0.0828	0.0691	0.0883	0.1838	0.1817	0.8305
	4	0.0765	0.0660	0.0626			

Outline

1. Background and Motivation
   - Credit Scoring
   - Segmentation
   - Feature Learning
2. Variational Autoencoders
   - Variational Inference
   - Autoencoders and Variational Autoencoders
   - Variational Autoencoders and Large Datasets
3. Data Representation
   - Intuition
   - Visualization and Clustering Results
4. Segment-based credit scoring
   - Model Development
   - Model Performance
5. Conclusion
   - Main takeaway

Conclusion

- We show that it is possible to steer configurations in the latent space of the VAE.
- The VAE and its feature learning capabilities facilitates customer segmentation, and improve the credit scoring assessment.
- The main advantages of our proposed method are:
  1. Captures non-linear relationships in the data.
  2. Suggests the number of clusters.
  3. Enables visualization for low-dimensional latent spaces.
  4. Generates the latent space for new data and assigns the corresponding cluster.
  5. Improves credit scoring performance.

Thank you for your attention!