UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

AI & ML

(6CS012)

Assignment - III

Student Id	: 238119
Student Name	: Vibek Rana Magar
Group	: L6CG19
Lecturer	: Siman Giri
Tutor	: Durga Pokharel
Word Count	:
Submitted on	:

Table of Contents

Long Question

(5 pt.) You are a Data Scientist at eSewa, Nepal's leading digital payment platform.
• Identify two high - impact areas where unsupervised learning could add significant business value
• For each area:
– Define the problem clearly (e.g., customer behavior clustering, fraud pattern detection).
– Propose a specific unsupervised learning approach (e.g., clustering, anomaly detection, dimensionality reduction) and recommend one or more algorithms (e.g., K-Means, DBSCAN, Autoencoders).
– Briefly explain how the outputs from these models can be integrated into eSewa's products or services to drive business decisions.

Answer:

Unsupervised Learning Applications for eSewa

Customer Segmentation for Personalized Services

Grouping eSewa users is helpful to tailor the app's offerings in terms of demographics or of behavior. Customer segmentation "involves the grouping of customers on the basis of shared characteristics" (geeksforgeeks, 20025). Strategies can be tailored, as specific groups can be targeted more effectively via this segmentation. Certain user clusters (e.g., frequent payers versus casual users) can be discovered through unsupervised clustering (like K-Means or DBSCAN). Through the process of identifying the segments, marketing elements and user interface can be customized by eSewa for each group. Segmentation targets certain promotions or interface features through the enablement of engagement improvement (e.g. special loan offers or simplified dashboards for certain clusters) (Bozkus, 2022) (geeksforgeeks, 20025).

<u>Techniques:</u> Use clustering algorithms such as K-Means (partitioning users into *k* groups), density-based DBSCAN (finding dense subgroups and outliers), or hierarchical clustering to uncover natural user segments (Bozkus, 2022).

<u>Integration</u>: Marketing/UX systems apply multiple clusters. Tailored promotions and personalized UI flows should be sent for each individual segment. Business-loan offers may be seen by heavy-business users, as easy-refill reminders do get to occasional users. This targeted approach "helps businesses to understand their customer base." Businesses can tailor their marketing in accordance (Bozkus, 2022)

Fraud Detection via Anomaly Detection

E sewa needs to detect payment patterns that are unusual or suspicious and differ from normal user behavior, like fraudulent transactions. Anomaly detection is able to find outliers with no labeled fraud. Models that are unsupervised flag deviations. The models learn each user's typical transaction patterns. Normal transactions can indeed be modeled, as well as anomalies scored by, for example, an Isolation Forest (ensemble of random trees) and deep Autoencoder neural networks. The abnormal points can be naturally identified through these methods that do not require pre-labeled fraud data. (Rajeev, 2023)

<u>Methods:</u> Use Isolation Forest (tree-based model that isolates anomalies efficiently) and Autoencoders (neural nets trained on normal data to spot high-reconstruction-error anomalies). Other options include One-Class SVM or Local Outlier Factor. (Bozkus, 2022)

<u>Real-Time Mitigation:</u> Deploy the anomaly detector onto transaction streams that are live. Each and every incoming transaction receives a score which is based on how anomalous it is. Alerts or automatic holds are triggered if transactions are of highly anomalous nature. This "flag[s] potentially fraudulent transactions in real time" as well as "reduces the risk of financial losses". Because of prompt alerts that let eSewa block or review suspicious activity right away, fraud losses get cut and users are then protected. (fraud.com, 2024)

Short Question

Overfitting

(2.5 pt.) In the context of machine learning:
• Define and differentiate between overfitting and underfitting.
• Explain why both are problematic for model performance.
• Illustrate your explanation with simple examples (e.g., overfitting a training dataset, underfitting a complex pattern).

Answer 1

Overfitting occurs in situations when training data is memorized quite effectively as well as a model being complex. It achieves a quite low amount of error on such training set. However, it fails in generalizing since the error on new data is high. Underfitting happens when a model is of insufficient complexity such that it cannot capture the true pattern with elevated error on both the training set and new data.

Key Differences

- Error patterns: Overfitting yields minimal training error; underfitting yields substantial error on both datasets.
- Model complexity: Models that overfit possess many parameters; hence, they display elevated variance. Models that do underfit lack sufficient complexity, meaning that they possess high bias.

Why problematic

- Overfitting: Poor performance occurs on any new data on account of the fact that the model “memorizes” noise as well as outliers in the training set.
- Underfitting: The model misses important patterns, and therefore accurate predictions cannot be made, not even on the training data.

Examples

- Overfitting example: A highly flexible curve is fitted through all training points. Because of noise that it has learned, it fails on new examples though it matches the training data perfectly.
- Underfitting example: Clearly, curved data do use a simple straight line. The training set as well as the test set are each performed badly on, and also the trend is thus missed.

Neural Network
(2.5 pt.) Difference between CNN and RNN:
• Explain the fundamental differences between a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN).
– Provide examples of scenarios where one would be more suitable than the other (e.g., image recognition vs. time-series prediction).
• Briefly discuss common challenges faced during the training of deep learning models (e.g., vanishing gradients, overfitting).
– Provide possible solutions or techniques to address these challenges (e.g., batch normalization, early stopping).

Answer 2

Fundamental Differences

CNNs happen to be feedforward networks which apply learned convolutional filters to fixed-size grid data, such as images, as well as extract spatial features by way of shared weights. RNNs are recurrent networks possessing loops: each output feeds back as input, thereby providing an internal “memory” of prior inputs. RNNs are able to handle many variable-length sequences, plus capture temporal dependencies. Entire inputs are each processed by CNNs in parallel on top of a fixed grid.

Use Cases

- **CNN:** Excels at spatial data. Common applications are image classification, object/facial recognition, and medical image analysis.
- **RNN:** Excels at sequential data. Common applications include time-series prediction, speech recognition, and language modeling plus translation.

Training Challenges and Remedies

Deep networks face vanishing gradients shrinking in early layers and overfitting where the model memorizes training data, slowing learning. Maintaining gradients that are well-scaled does indeed help to stabilize the training and to speed it up through batch normalization (and ReLU activations). Dropout as well as data augmentation are regularization methods which help to prevent overfitting. For the avoidance of over-training, early stopping is additionally a simple way, through halting training with validation error rises.

References

Bozkus, E. (2022, December 24). *Using K-Means, Hierarchical Clustering, and DBSCAN to Group Customers*. Retrieved from Medium: https://eminebozkus.medium.com/using-k-means-hierarchical-clustering-and-dbscan-to-group-customers-2fca47c96eab#:~:text=Using%20K,ultimately%20improve%20their%20bottom%20line

fraud.com. (2024). *Anomaly detection for fraud prevention – Advanced strategies*. Retrieved from fraud: https://www.fraud.com/post/anomaly-detection#:~:text=Anomaly%20detection%20is%20essential%20for,not%20only%20protects%20financial%20assets

geeksforgeeks. (20025, April 08). *Customer Segmentation using Unsupervised Machine Learning in Python*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/customer-segmentation-using-unsupervised-machine-learning-in-python/

Rajeev, A. (2023, March 21). *Unsupervised Anomaly Detection on Credit Card Fraud*. Retrieved from Medium: https://medium.com/@aishwaryarajeev_68039/unsupervised-anomaly-detection-on-credit-card-fraud-3e9c8802fd01#:~:text=Anomaly%20detection%20is%20a%20powerful,intrusions%20or%20industrial%20equipment%20failures