Data Mining
EX7

Omid Naeej Nejad
610301189

December 8, 2024

1 Exercise 1

1.1 Difference Between Supervised and Unsupervised Methods:

- Supervised Methods:
  These involve learning a mapping from input data to a labeled output. The algorithm uses labeled training data to predict new unseen data. (e.g. regression and classification tasks.)

- Unsupervised Methods:
  These focus on identifying patterns or structures in data without any labeled outcomes. They are typically used for clustering, dimensionality reduction, or anomaly detection.

More Suitable Approach:
Unsupervised methods are more suitable because customer segmentation is about discovering groups based on similarities in behavior or patterns without predefined labels. On the other hand, labeling has a lot of costs, so unsupervised methods are better.

1.2 Data Mining Tasks

Supervised Methods:

- Classification (e.g., predicting customer churn).
- Regression (e.g., predicting customer lifetime value).

Unsupervised Methods:

- Clustering (e.g., grouping customers with similar purchasing behaviors).
- Association rule learning (e.g., identifying products frequently bought together).

1.3

Supervised Methods require labeled data because the algorithm learns to predict or classify based on provided outcomes.
Unsupervised Methods do not require labeled data, as the focus is on exploring patterns within the data.

For this scenario (customer segmentation), **unsupervised methods** (e.g., clustering techniques like k-means) would be appropriate since the aim is to group customers without preexisting labels.

2 Exercise 2

2.1 Differences Between Training, Test, and Validation Sets:

- Training Set: This is the portion of the data used to train the model. The model learns patterns, relationships, and parameters from this dataset. (around 70 - 80% of the dataset records)
- Validation Set: This is used during training to fine-tune the model, such as optimizing hyper-parameters. It helps assess how well the model generalizes to unseen data during the training process. Validation set has an essential role in the cross validation to select the best hyper -parameters. (around 15 - 20% of the dataset records)
- Test Set: This dataset is used after the model is fully trained to evaluate its performance on completely unseen data. It provides an unbiased

estimate of the model's ability to generalize. (around 15 - 20% of the dataset records)

2.2

- Striving for the highest possible accuracy on the **training set** is not ideal because it may lead to **overfitting (High variance model)**, where the model memorizes the training data rather than learning general patterns. Overfitted models perform poorly on unseen data.
- **High accuracy on the validation set is a better indicator of a well-performing model**. However, overly optimizing for the validation set (e.g., by repeatedly adjusting hyperparameters) can lead to validation overfitting. It's important to strike a balance and ensure the model generalizes well.

2.3 Preventing Overfitting:

1. **Regularization:** Add penalties to the model's complexity (e.g., L1/Lasso or L2/Ridge or Elastic Net(L1+L2) regularization) to avoid emphasis on all features.
2. **Simplify the Model:** Use a less complex model (e.g., fewer layers in neural networks, pruning decision trees or simpler model in linear regression).
3. **Cross-Validation:** Use techniques like k-fold cross-validation to assess model performance on multiple subsets of the data.
4. **Dropout (for Neural Networks):** Randomly drop units during training to reduce reliance on specific neurons.
5. **Stop Training (Reduce Epochs):** Stop training when performance on the validation set stops improving. (Optimal Model Complexity)
6. **Increase Training Data:** More data helps the model generalize better. (But it's not possible in many situations.)

By focusing on these techniques, we can create a model that generalizes well without overfitting.

3 Exercise 3

3.1

Assume $N = 1000$:

$$0.04 \times 1000 = 40 \text{ fraudulent samples.}$$

To increase the proportion of fraudulent transactions to 20

$$x + 40 = 0.2 \times (1000 + x)$$

$$x + 40 = 200 + 0.2x$$

$$x = 200$$

Thus, 200 new fraudulent samples need to be added to the dataset. After adding these, the total dataset size becomes:

$$1000 + 200 = 1200,$$

with $240/1200 = 20\%$ fraudulent samples.

3.2

A baseline performance provides a reference point to evaluate the effectiveness of the model.
Without a baseline, it's challenging to determine whether the model offers meaningful improvements compared to a previous model or random guessing (e.g., is the model significantly better than simply predicting all transactions as non-fraudulent?).

3.3

- Absolute Difference: The simple numerical difference between two values. For example, if accuracy improved from 80% to 85%, the absolute difference is $85\% - 80\% = 5\%$.

- Relative Difference: The percentage change relative to the baseline value. Using the same example, the relative difference is:

$$\text{Relative Difference} = \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}} \times 100$$

$$= \frac{85\% - 80\%}{80\%} \times 100 = 6.25\%$$

For reporting model performance, ***Relative difference*** *is more informative for results that involve incremental improvements over a baseline, especially if the baseline value is low.*