Assignment 2

Deadline: 08.01.2023

1. What are the embedding techniques used in text mining?
2. What kind of techniques can be used when there are more than one type of outliers. Describe one of them.
3. What is graph mining? Give one of the graph mining technique that is well known in graph mining literature. Describe it.
4. What are the statistical techniques to evaluate the relationship between each input variable and the target variable? Give at least 5 of them and explain them shortly and give formulas.
5. Explain one feature selection and one feature extraction technique, that is not mentioned in our lectures, in details.

Ahmet Tuğkan Ayhan
1901042692

CSE454
Data Mining

Homework 2

Question 1

One-hot Encoding → This embedding technique represents each unique word in the vocabulary as a binary vector. The length of this vector is equal to the size of the vocabulary. and there is a "1" in the position corresponding to the word and "0" everywhere else. This technique is simple to implement and understand but it can be computationally expensive and may lead to overfitting in machine learning models due to its high dimensionality.

Term frequency-inverse document frequency (TF-IDF) → This technique is a weighting method which gives higher weights to words that occur more frequently. The weight of a word in a document is calculated by multiplying its term frequency (the number of times it appears in the text) by the inverse document frequency (IDF), which shows how rare the word is in the text. This technique has the advantage of being able to downweight the importance of common words (such as "a", "the", etc.)

Word2Vec → This technique learns dense vector representations of words, which called "word embeddings". CBOW (continous bag-of-words) version of this technique uses the context of surrounding words to predict the current word. Skip-gram version uses the current word to predict surrounding words. Word2Vec can capture the meanings of words and the relationship between them. (useful for text classification)

GloVe → Global vectors for word representation technique learns dense vector representations of words based on the global statistical properties of the text port. This is done by using co-occurence matrix, which contains the amount of how often each word co-occurs with every other word in the vocab. This technique can also capture the meanings of words and the relationship between them.

FastText → This technique represents words as combination of character n-grams (contiguous sequences of n characters). This allows FastText to handle out of vocabulary words more effectively. Because it predicts them based on the n-grams they contain

Question 2

* We can identify and analyze different types of outliers seperately. For example, we can use a clustering algorithm to identify outliers that are far from the main cluster.

* We can also use a statistical algorithm like Grubbs test to identify outliers that are significantly different from the rest of the data.

* Another technique would be using a combination of univariate and multivariate outlier detection methods.

** Important part is, we need to analyze the characteristics of the different types of outliers we are looking for and choose the techniques according to that.

** Describing one of the techniques:

→ If I am looking for outliers that are significantly different from the majority of data, I would use Standart Deviation or Boxplot algorithm. Standart deviation calculates the standart deviation of a data set and identifies any points that fall more than a certain number of standart deviations from the mean. And boxplot is a graphical representation of the statistical distribution of a data set.

→ If I am looking for outliers that are far from any densely packed clusters of point, I would use DBSCAN. Because it is a density-based algorithm. and it works by identifying clusters of points that are densely packed together.

→ If I am looking for outliers that are isolated in a large number of decision trees, I would choose, Isolation Forest and Robust Random Cut Forest. Isolation Forest works by building a series of decision trees that isolate individual points in the dataset. Points that are isolated in fewer trees are considered to be abnormal. Robust Random Cut Forest is similar to the Isolation Forest but it is designed to be more robust to noisy data.

Question 3

* Graph mining is a subfield of data mining that focuses on the extraction of patterns and graph data.

* Graph mining techniques can be used to perform tasks such as: identifying communities and clusters within a graph, finding central or influential nodes, detecting patterns and trends, predicting future interactions etc.

* Community detection algorithms are well-known in graph literature. One of the community detection algorithm is spectral clustering.

* Spectral clustering is a graph mining technique which uses eigen-vectors of the graph Laplacian matrix to perform dimensionality reduction on the graph, then applies traditional clustering algorithms like k-means to the reduced shape.

* In spectral clustering technique, we first construct the graph Laplacian matrix. Then we compute the eigenvectors of the Laplacian matrix. These eigenvectors can be used to represent the nodes of the graph in a lower dimensional space.

* After computing eigenvectors, we perform dimensionality reduction. This can be done by retaining only the k eigenvectors with the largest eigenvalues.

* And finally, we apply a clustering algorithm to partition the nodes into clusters.

* Spectral clustering algorithm can handle graphs with non-uniformly sized communities, and relatively good against noise. But it can be computationally expensive in large graphs.

Question 4

Linear Regression → Linear regression is a method used to model the linear relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the values of the independent variables.

* $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$

- y is the dependent variable
- $x_1, x_2, \dots x_n$ are the independent variables
- b_0 is intercept and $b_1 \dots b_n$ coefficients

Logistic Regression → Logistic regression is a statistical method used to model the relationship between a binary dependent variable and one or more independent variables. It is a binary classification technique, meaning it is used to predict a binary outcome.

* $p(y) = \dfrac{e^{(b_0 + b_1x_1 + \dots + b_nx_n)}}{1 + e^{(b_0 + b_1x_1 + \dots + b_nx_n)}}$

- $p(y)$ is the probability that the event will occur
- $b_1, b_2, \dots b_n$ coefficients b_0 intercept
- $x_1, x_2, \dots x_n$ independent variables

Chi-square Test → It is a statistical test used to determine if there is significant difference between the observed frequencies of a categoricle variable and the expected frequencies of that variable.

* $X_c^2 = \sum \dfrac{(\text{Observed Value} - \text{Expected Value})^2}{\text{Expected Value}}$

c = degree of freedom
$df = (r-1)(c-1)$

Correlation Analysis → It is a statistical method used to measure the strength and direction of the relationship between two continious variable. The strength of the relationship can be measured using correlation coefficient. Value of -1 indicates strong negative, +1 indicates strong positive relationship.

Pearson' correlation coefficient formula ⇒ $r = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

ANOVA → Analysis of variance is a stasistical method used to compare the means of two or more groups.

* One-way ANOVA → $F = \dfrac{\text{Between-group variance}}{\text{Within-group variance}}$

* Between-group variance = (sum of squares between groups) / (number of groups - 1)
* Within-group variance = (sum of squares within groups) / (total number of data points - number of groups)

Question 5

Chi-square Test → The chi-square test is a feature selection method used to determine whether there is a significant difference between the observed frequencies in a categorical data set and the expected frequencies.

To perform chi-square test, first we need to determine the null hypothesis. This is a statement that there is no significant difference between the observed and expected frequencies

Next, we determine the alternative hypothesis. This is a statement that there is a significant difference between the observed and expected frequencies.

Then we collect the data. The data should be organized into a frequency table that shows the number of observations in each category for both variables

After that, we calculate the expected frequencies. Expected frequencies are calculated using this formula: (row total * column total) / sample size

Next, we calculate the chi-square statistic $\sum \frac{(\text{observed frequency} - \text{expected freq})^2}{\text{expected frequency}}$

After chi-square statistic, we calculate degree of freedom which is number of categories - 1

Before final step, we calculate the p-value. p-value is the probability that the observed frequencies are due to chance. It is calculated by comparing the chi-square statistic to a chi-square distribution table with appropriate degree of freedom.

Finally, we make the decision. If p-value is less than the pre determined level of significance, then the null hypothesis is rejected and the alternative hypothesis is accepted.

Independent Component Analysis → ICA is a feature extraction method that is used to seperate a set of mixed signals into their individual sources. It is based on the assumption that each signal source is statistically independent of others

ICA works by finding the statistical independence of the signal sources, and then seperating them based on this independence. It does this by identifying patterns in the data that are unique to each source. and then using these patterns to reconstruct the individual sources.

There are several different methods for performing ICA (fastICA, infomax, jade). Those methods differ in their approach to finding the statistical independence of the sources, but all aim to identify the sources by finding patterns in the data that are unique to each source.