Göğa Şahin
1710444050

CSE 454 - Data Mining - Assignment 2

1. Solution:

• Word Embeddings: It's a way of representing words as dense, cont-valued vectors in a high-dimensional space, such that the vectors capture the meanings and relationship between words. Used in natural language processing tasks, such as language translation, text classification, information retrieval. Methods: Word2Vec, GloVe.

• Sentence Embeddings: It's a way of representing entire sentences or paragraphs as fixed-length vectors in a high-dimensional space, such that the vectors capture the relationships between words in sentence. There are several way of generating sentence embeddings, including averaging the word embeddings of the words in sentence, using a recurrent neural network to process the seq. and generate a fixed-length vector representation.

• Document Embeddings: Representing entire document as a fixed-length vectors such that the vectors capture the overall meanys and relationship between the words in the document. It can be used for tasks such as text classificaton, informal retrieval and topic modeling.

• Entity Embedding: It represent named entities (such as people, organizations, locations) as vectors, capturing the relationship between different entities and their roles in a given context.

2. Solution:

- Robust statistical methods
- Outlier detection algorithms
- Data transformation
- Anomaly detection models.

Robust Statistical Methods: These methods are designed to be resistant to the influence of outliers, can be used to fit models to data that may have a mix of different types of outliers. One example is RANSAC (Random Sample Consensus) algorithm, which fits the model to a subset of the data that is deemed to be inliers (not outliers) then uses this model to make predictions for the rest of the data.

3. Solution:

Graph mining is a process of extracting the hidden patterns of data from the graphs to get the useful information with retardencies to the actual data given which is represented in the form of graph. It's main motive is to finding the subgraphs which helps in compressing the data and to find hidden patterns

Link prediction:
When large number of nodes and edges occurs in a sinyle graph, it is difficult to find the link. It helps in predicting the edges that will added to the graph for future references.

4. Solution:

a) Correlation: Measures the steyth and direction of the linear relationship between two variables.

Pearson Correlation coefficient:

$$r = \sum(x-\bar{x})(y-\bar{y}) \Big/ \sqrt{\left(\sum(x-\bar{x})^2 \sum (y-\bar{y})^2\right)}$$

$x, y \rightarrow$ variables
$\bar{x}, \bar{y} \rightarrow$ means of x, y

b) Regression: 1 dependent variable (target) $\longleftrightarrow$ 1or more independent variable

$$\hat{y} = b_0 + b_1 x.$$

$\hat{y} \rightarrow$ predicted value of dependent variable
$b_0 \rightarrow$ intercept
$b_1 \rightarrow$ slope of the line
$x \rightarrow$ independent variable

c) Analysis of variance (ANOVA)
Statistical test that is used to determine whether there is a significant difference in the means of two or three groups. One-way ANOVA is used to compare the mean of two groups, while two-way ANOVA is used to compare the means of two or more groups across multiple var.

d) Chi-square test: It is used to determine whether there is a significant association between two categorical variables. It can be used to determine whether the observed frequencies of categories are significantly different from the expected frequencies.

4. Cont.

e) t-test: This is a statistical test that is used to determine whether there is a significant difference between the means of two groups. Independent t-tests and paired t-tests are some types.

5. Solution:

Minimum Redundancy Maximum Relevance (mRMR):
mRMR is a feature selection method that aims to select the most relevant features while minimizing redundancy between the features. It works by selecting the feature that has the highest mutual information with the target variable while minimizing the mutual information between the selected features.

To implement mRMR, we first calculate the mutual information between each feature and the target variable. Then we select the feature with the highest mutual inf. as our first feature. For each subsequent feature, we select the feature with the highest mutual inf. with target, while minimizing the mutual inf. with the already selected features

Singular Value Decomposition (SVD):

SVD is a feature extraction technique that decomposes a matrix into three matrices: U, S and V. The matrix that we want to decompose is usualy a data matrix with dimensions m x n, where m is the number of samples and n is the number of features.
Matrix U → m x m → left singular vectors of data matrix.
Matrix S → m x n → diagonal matrix
Matrix V → n x n → right singular matrix.

SVD is used for dimensionality reduction, by keeping only the top k singular values and corresponding singular vectors. Also used for data imputation, by reconstructing the original data matrix using only the top k singular values and vectors. This can be useful for filling in missing values in the data.