CSE 454 - Data Mining
Assignment - II

Abdurrahman BULUT
190104222 58

1) What are the embedding system techniques used in text mining?

One of the primary goals of text mining is to derive high-quality information from text. This typically involves structuring the input text, deriving patterns within the structured data, and evaluating and interpreting the output. Text mining often requires the addition of some derived linguistic features and the removal of others, and may involve the insertion of processed text into a database.

In text mining, an embedding is a low-dimensional, continuous vector representation of a word, phrase, or document that captures some of its meaning and context. There are several techniques that can be used to create word embeddings, including:

One-hot encoding: This involves representing a word as a binary vector with a length equal to the number of words in the vocabulary. Each position in the vector corresponds to a word in the vocabulary, and the vector is "hot" (has a value of 1) at the index corresponding to the word it represents, and "cold" (has a value of 0) everywhere else. One-hot encoding is simple, but does not capture any semantic or contextual information about the words.

Count-based methods (Ex: Word2vec): These methods represent a word as a vector of its frequencies within a given context (e.g. a document). For example, the word2vec method represents a word as the sum of the vectors of its neighbors (i.e. the words that appear around it).

Predictive Methods (Ex: GloVe): These methods learn to predict a word based on its context. For example the GloVe method learns to predict a word from the sum of its neighbors' vectors, and the FastText method learns to predict a word from the sum of its subwords' vectors (i.e, n-grams)

Hybrid methods: These methods combine two or more of the above approaches. For example, the ELMo method represents a word as the concatenation of its one-hot encoding, its count-based representation, and a representation learned from its context

~~Term~~

Term frequency-inverse document frequency (TF-IDF): This is a measure of the importance of a word in a document, based on its frequency in the document and the inverse frequency of the word in a corpus.

In text mining, embeddings are often used as input features for machine learning models, such as classifiers, clusterers, and topic models. They can also be used to measure the similarity between words, phrases, or documents, and to visualize the relationships between them.

(Q2) What kind of techniques can be used when there are more than one type of outliers? Describe one of them.

There are several techniques that can be used to handle multiple types of outliers in a dataset. Some of these techniques include:

- Multivariate outlier detection: This involves identifying instances in the dataset that are unusual in more than one feature at a time.
- Clustering-based methods: These methods project/identify outliers as instances that are not part of any cluster.
- Projection-based methods: These methods project the data onto a lower-dimensional space and identify outliers in the projected space.
- Classification-based methods: These methods train a classifiers to predict whether an instance is an outlier or not.
- Ensemble methods: These methods combine the predictions of multiple outlier detection algorithms to identify outliers in the dataset.

LOF

One specific technique that can be used for multivariate outlier detection is the Local Outlier Factor (LOF) algorithm. The LOF algorithm calculates the local density of each instance in the dataset, and identify instances that have a significantly lower density than their neighbors. These instances are considered outliers.

To calculate the local density of an instance, the LOF algorithm first determines the distance between instance and its k nearest neighbors (where k is a user-specified parameter). It then calculates the reachability distance of the instance,

which is the maximum number of the distances between the instance and its k nearest neighbors, and the distance between its k-1 nearest neighbor and its kth nearest neighbor. The local density of the instance is then defined as the inverse of its reachability distance.

Instances with a low local density are considered outliers, as they are surrounded by a small number of other instances. The LOF algorithm can identify multiple types of outliers in the dataset, as it takes into account the densities of instances in multiple dimensions.

Q3) What is graph mining? Give one of the graph mining technique that is well known in graph mining literature. Describe it.

Graph mining is the process of extracting useful and interesting information or pattern from a graph structure, such as a social network or a biological network. Graph mining algorithms can be used to identify patterns, relationships, and trends in the data represented by the graph. This can be useful for various applications, such as finding communities or groups of related individuals in social networks, and predicting the spread of diseases in a population.

One well-known graph mining technique is called "PageRank". PageRank was developed by Google to rank web pages in their search engine and is based on the idea that a web page is important if it is linked to by other important web pages. In the context of graph mining, PageRank can be used to identify the most important nodes in a graph, such as the most influential users in a social network of the most central proteins in a biological network.

To compute PageRank, a transition probability matrix is constructed based on the structure of the graph. The transition probability between two nodes in the graph represents the probability that a random walker will move from one node to the other. The PageRank of each node is then computed as a weighted sum of the PageRank of the nodes that links to it, where the weights represent the transition probabilities. The PageRank values are then normalized so that they sum to 1, and the nodes are ranked in decreasing order of PageRank.

PageRank is just one example of a graph mining technique. There are many other techniques that have been developed for graph mining, such as community detection, centrality measures, and link prediction.

Q4) 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:

Overview of some statistical techniques:

- Regression analysis: This is a statistical method that help us understand how one variable (the dependent variable) is affected by one or more other variables (the independent variables). For example, you might use regression analysis to understand how a person's age, education level and income might affect their likelihood of buying a certain product.

dependent variable: y, independent variables: (x1, x2, x3 --)

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

- Correlation analysis: This technique measures the strength and direction of the relationship between two variables. A strong positive correlation means that as one variable increases, the other variable also tends to increase. A strong negative correlation means that as one variable increases, the other variable tends to decrease. It calculates as:

$$r = \text{sum}((x - \text{mean}(x)) * (y - \text{mean}(y))) / \text{sqrt}(\text{sum}((x - \text{mean}(x))^2) * \text{sum}((y - \text{mean}(y))^2))$$

$$r = \Sigma (x - \bar{x})(y - \bar{y}) / \sqrt{[\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2]}$$

where x and y are the two variables being correlated, and mean() is the mean value of the variable.

- Chi-square test: This is a statistical test that is used to see if there is a relationship between two categorical variables. For example, you might use a chi-square test to see if there is a relationship between a person's political party and their stance on a particular issue.

The formula for the chi-square test statistic is:

$chi^2 = \Sigma (O-E)^2/E$

$chi^2 = sum((observed\ frequency - expected\ frequency)^2 / expected\ frequency)$

where observed frequency is the number of observations in a particular category and expected frequency is the number of observations that would be expected in that categorical if the variables were independent.

- ANOVA: This is a statistical test that is used to see if there is a significant difference between the means of two or more groups. For example, you might use ANOVA to see if there is a significant difference in the average grades of students who have different levels of parental involvement. The formula for calculating the F-statistic, which is used to determine whether there is a significant difference between the means, is:

$F = (mean\ group1 - mean\ group2)^2 / (variance\ group1 / n\ group1 + variance\ group2 / n\ group2)$

where mean group1 and mean group2 are the means of the two groups being compared, variance group 1 and variance group2 are the variances of the two groups, and n group 1 and n group2 are the number of observations in each group

- t-test: This is a statistical test that is used to see if there is a significant difference between the means of two groups. For example, you might use a t-test to see if there is a significant difference in the average exam score of students who studied with a tutor versus those who did not.

The formula:

$t = (\bar{x}1 - \bar{x}2) / [s^2 (1/n1 + 1/n2)^{0.5}]$

$t = (mean\ group1 - mean\ group2) / sqrt(variance\ group1 / n group1 + variance\ group2 / n group2)$

where mean group 1 and mean group 2 are the means of the two groups being compared, variance group1 and variance group2 are the variance of the two groups, and n group 1 and n group2 are the number of observations in each group.

6. • Logistic regression: It is a statistical method that is used to model the probability of a binary outcome. It is written as:

$$p = 1/(1+e^{-(b_0+b_1x_1+b_2x_2+\dots)})$$

where p is the probability of the binary outcome, $x_1, x_2 \dots$ are the independent variables, and $b_0, b_1, b_2 \dots$ are the coefficients that represents the influence of each independent variable on the probability of the outcome.

---

25) Explain one feature selection and one feature extraction technique, that is not mentioned in our lecture, in details.

Feature Selection: Example

• Boruta: Boruta is a feature selection method that uses random forests to identify the importance of each feature in a dataset. It was developed to address the problem of selecting relevant features in the presence of many irrelevant or redundant features.

The basic idea behind Boruta is to create a set of "shadow" features that are randomly generated copies of the original features. The original and shadow features are then fed into a random forest model, and the feature importance scores are calculated using the mean decrease in impurity (MDI) measure. The importance score of each shadow feature is then compared to the importance score of the corresponding original feature. If the importance score of the shadow feature is higher than the importance score of the original feature, it means that the original feature is not important, and it is marked for removal. This process is repeated until all shadow features have been compared to their corresponding original features. At the end, all the original features that were not marked for removal are considered relevant, and the others are considered irrelevant.

One advantage of Boruta is that it handles both continuous and categorical feature, and it can handle high-dimensional datasets with a large number of features. It is also relatively robust to correlated features and noisy data. However, it can be computationally expensive, espetially for large datasets and it may not work well for datasets with highly imbalanced class distributions.

Feature extraction: Example

t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is a non-linear dimensionality reduction technique that is used to visualize high-dimensional datasets. It works by transforming the original data into a lower-dimensional space, typical two or three dimensions, in a way that preserves the local relationships between the points.

The basic idea behind t-SNE is to define a probability distribution over the pairs of points in the original high-dimensional space, such that similar points have a high probability of being chosen, and dissimilar points have a low probability of being chosen. This is done using a Gaussian kernel, which assigns a higher probability to pairs of points that are closer together and a lower probability to pairs of points that are farther apart. The probability distribution is then transformed into a similar distribution in the lower-dimensional space using a technique called gradient descent, which iteratively adjusts the position of the points in the lower-dimensional space to minimize the difference between the two distributions.

t-SNE has several advantages over other dimensionality reduction techniques. It is able to capture the local structure of the data, it is resistant to noise and outliers, and it is able to handle large datasets. However, it can be computationally expensive, especially for large datasets, and results can be sensitive to the choice of hyperparameters. It is designed specifically for visualization.