CSE521 DATA MINING FINAL REPORT SUMMER 2021 (30pts)

TEAM NUMBER: 5 Leader **NAME** and **ID:** VENKATA RAVI TEJA TAKKELLA - 113219890

Team members Leader **NAMES** and **IDs:**

1. NAGIREDDY SUSMITHA REDDY - 113271915

2. KRISHNA SHASHANK GORANTA - 113221080

FORMAT OF SUBMISSION

1. The TEAM LEADER submits **ONE PDF** file for the TEAM

2. Name your PDF file as $< TEAMLEADERID >$.pdf

Example: 11384578.pdf

All team members receive the same grade.

FINAL REPORT DESCRIPTION

This is a TEAM REPORT. It must contain **TWO SHORT ESSAYS** on subjects of your CHOICE extending the material we have covered in the course.

You assume that all algorithms and all facts covered in CLASS are KNOWN so there is no need to include them in your Essays.

You can use the material form **Lectures- Presentations** published on course website and Class Lectures not covered in class.

You can also use material from **other sources** depending of your subject choice.

You can also use research papers, research applications and commercial application, books, encyclopedias, overview articles, etc...

You can also choose what is the most interesting for you now and would profit you in the future because of the material you have learned in the course.

You have write a short **motivation** why you decided to choose your subject at the **beginning** of your ESSAY and to make a list of all of your sources at the **end** of the Essay.

There is no special structure/template of essay content that you need to follow. I use the word "ESSAY" for a structured description of the subject content.

The ESSAYS have to contain only **3-4 pages** of **plain text** and contain only **strict subject** content. Your ESSAY do not need to cover all of the information about the SUBJECT you decided to present. You but have to gather your materials and describe the subject in the best, short, and as comprehensible way as you can master.

You must write it all in your own words. I want you to show your own understanding of the material. If you choose to copy some statements from ANY source you must remember to use it as a formal citation, i.e the **strict laws of direct citations** apply to all materials you use.

Here are some possible **GENERAL Categories** of subjects for your ESSAYS.

YOU CAN CREATE OUR OWN SUBJECTS within any of these general categories.

E1 Clustering Methods and Applications

E2 Classification Methods and Applications

E3 NLP methods and Applications

E4 Genetic Algorithms, models and their applications

E5 Neural Networks Deep Learning and other models and their applications

E6 Association Methods and Applications

E7 Web, Sentiment, and Text Mining

E8 Data Warehouse and OLAP

E1 Clustering Methods and Applications

Motivation :

Clustering Analysis is basically a method where we segregate objects which are similar to each other into a group (cluster). The objective of clustering is to make the objects in the same cluster to be more similar than with other clusters. It's one of the main tasks in exploratory data analysis and a common technique in statistical data analysis. It's also very well used in many fields like pattern recognition, bioinformatics, machine learning e.t.c. The motivation behind this essay is to bring out the methods which are not covered in our lecture and are used heavily in industry. This includes other methods which include probability (EM algorithm) and density based clustering (DBSCAN). These are well known techniques which might be useful if anyone wants to delve deeper into clustering analysis. We also tried to explain the importance of Attribute Selection in clustering analysis.

Essay :

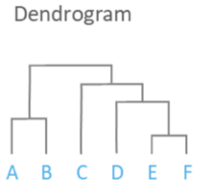
Clustering Analysis is a very useful method in most of the Machine Learning applications and it helps to give a basic visualisation on how the objects classify. But these methods are generally hindered by large scale datasets or the ones with a huge number of features. Different features are useful in clustering the objects and few unimportant features try to hinder the process of clustering. So basically, we can introduce Dimensionality Reduction initially before we start the clustering analysis. The most general method of Dimensionality Reduction is Principal Component Analysis (PCA), but this method in general helps to decrease the number of features but also changes the feature names which might result in understanding to the user. If that's a problem, we can also try to use the entropy method. Entropy is a word from Thermodynamics which basically tries to provide information about the predictability of the data distribution. Basically the data distributions with very less entropy are the easiest to cluster, so we try to greedily remove the features in order to generate a less entropy data distribution. Entropy is calculated based on the below formula where "i" is iterated on each attribute.

$$E = -\sum_{i=1}^{m} [p_i \log(p_i) + (1 - p_i)\log(1 - p_i)].$$

We can now discuss different other techniques of clustering which were not discussed in the class.

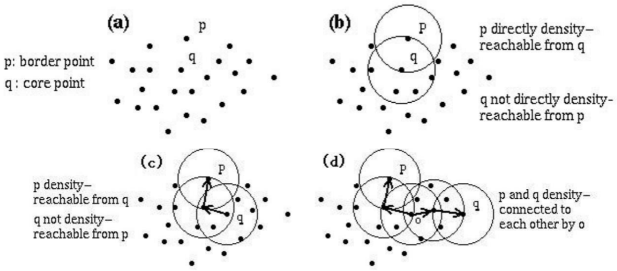
Methods :

- **(Hierarchical Clustering) Connectivity Based Clustering:** This method is discussed in our lecture but is required for the other methods which are discussed below. It basically uses distances between the objects to form clusters. And at different distances, different clusters get formed which can be represented in the form of a dendogram; hence it's also called hierarchical clustering.



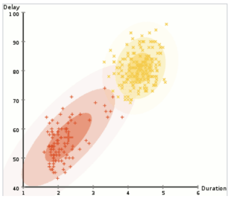
These algorithms not only provide a single partitioning of the data but also provide a hierarchical level of clustering by merging the clusters based on certain distances. This method can also be performed in two different approaches, where the hierarchy moves from top to bottom (divisive) and from bottom to top (agglomerative). Whereas the linkage or the merging of the clusters can also be done in multiple ways. This includes methods like Single-Linkage (by joining the closest pair), Complete-Linkage (joins the farthest pair), Group-average Linkage (average distance between all the objects in the group), Ward's method e.t.c. Basically these methods are a bit sensitive to smaller mistakes which were made during the clustering process. This is due to the noise and this being an unsupervised algorithm, there's no way to rectify this back.

- **(DBSCAN) Density Based Clustering :** This is one of the most popular density based clustering algorithms. This method uses distance between the objects as the metric for clustering but also uses a density parameter which thresholds a minimum number of objects in a specified radius of the cluster. This helps to generally discard the objects which are either the outliers or the noisy data since they don't follow the density criterion. The DBSCAN algorithm can also generate arbitrary shaped clusters, maybe a cluster totally surrounding an inner cluster. This is due to the density threshold and finally reducing the single-linkage issue which is mentioned above in the hierarchical clustering.

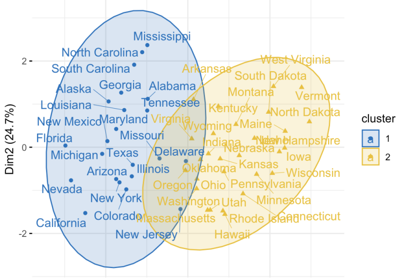

(a)
p
q
p: border point
q : core point
(b)
p
q
p directly density–reachable from q
q not directly density-reachable from p
(c)
p
q
p density–reachable from q
q not density–reachable from p
(d)
p
o
q
p and q density–connected to each other by o

- **(T-SNE) Probability Based Clustering :** T-distributed stochastic neighbour embedding (T-SNE) using the popular technique Linear Discriminant Analysis (LDA) to decrease the dimensionality of each object to 2 or 3 dimensions. Finally after reducing the objects, the objects which are placed together in the new dimensions are clustered together. It basically contains two main parts, firstly each object pairs are provided a probability distribution based on their similarity in the higher dimension. Secondly, a similar probability distribution is provided in the lower 2 or 3 dimension too once the dimensionality reduction is done. It finally tries to minimize the KL divergence between the distributions in both these dimensions. The T-SNE algorithm is entirely dependent on how we define the parameters while providing the probability distributions. A good set of parameters are generally known for providing a well separated cluster in T-SNE.

- **(E-M) Model-Based Clustering :** This method is similar to k-means clustering but uses probabilities of the objects based on one or more probability distributions. This method contains two steps, firstly we estimate the missing latent variables in the Estimation step and we optimize the parameters to best explain the data in the Maximization step. The method where we use Gaussian probabilistic distributions is called the Gaussian Mixture model. So it basically uses Gaussian probabilistic distributions and we need to estimate the mean and variance of each distribution by the end to attain good clusters. The better thing with this method is we can get overlapped clusters in the initial stages which tend to be more segregated as the parameters are better estimated. This method uses a Maximum Likelihood Estimation for maximizing the parameters.



- **Fuzzy Clustering :** This method is one where each object can be part of multiple clusters rather than just being in a single cluster. For example, an apple can be part of both red and green clusters and not just be in red. There are basically two areas of Fuzzy clustering algorithms, Classical Fuzzy and shape based. Fuzzy C-means algorithm is well known among these where probability is calculated for each object of being part of all the clusters based on the centroids of respective clusters.



Applications :

- Clustering analysis can be used for finding the violations in traffic. It helps to categorize the vehicles based on their movement and all the outliers which tend to be moving in a different direction than they are supposed to can be taken noted as violators of rules.
- Clustering can also be used widely in categorizing the consumers of any institution. It helps to cluster them and provide offers or products based on their interest. It also helps to provide a summary of the revenues and plan ahead for their prospective products.
- Clustering is a well known technique used in NLP to remove the lexical ambiguity which clusters the words which have similar meanings together.
- Cluster Analysis is used widely in finding the Crime hotspots based on the crime reports and the type of crimes and the general victims. This can help the police department to place their officers in abundance near these hotspots.
- It's also widely used in Social Media Platforms to recognize communities among large groups of people.
- Sequence analysis is also widely used in BioInformatics to cluster similar genetic data to infer population structures.
- It also can be used to cluster the stocks of companies into different sectors like Education, Farming e.t.c.
- Fraudulent Transaction Analysis can be implemented using clustering in either financial, telecommunication sectors e.t.c. It can help to categorize the transactions into clusters and find if it got merged into any of the known fraud transactions.

References :

- https://en.wikipedia.org/wiki/Cluster_analysis
- https://en.wikipedia.org/wiki/DBSCAN
- https://www.displayr.com/what-is-dendrogram/
- https://www3.cs.stonybrook.edu/~mueller/teaching/cse564/cluster%20analysis%20CSE%20564%202021.pdf
- https://www3.cs.stonybrook.edu/~cse521/L14Cluster1.pdf
- https://www3.cs.stonybrook.edu/~cse521/L15Cluster2.pdf
- https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
- https://www.statisticshowto.com/fuzzy-clustering/
- https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
- https://www.datanovia.com/en/lessons/fuzzy-clustering-essentials/

E2 Classification Methods and Applications

Motivation :

Classification is a supervised machine learning approach, in which the algorithm learns to categorize the data into a given number of classes. The main objective of a classification problem is to identify the category/class to which a new data will fall under. One of the most common examples, spam detection in emails can be treated as a classification problem. There are many other important applications of classification in various domains such as in customer target marketing, biological data analysis, medical diagnosis, multimedia data analysis etc. We make use of various classification algorithms such as K-Nearest Neighbours, Decision Tree, Support Vector Machine etc to perform a classification. The selection of the classification algorithm depends on the application and nature of the available data set. The main motivation behind this essay is to put together the most common types of classification algorithms such as Logistic Regression, Random Forest, Stochastic Gradient Descent etc and their applications, which are not covered in our lecture and are used heavily in the industry.

Essay :

Classification is the process of predicting the class of given data points. We sometimes refer to classes as targets/ labels or categories. Classification is treated as supervised learning, as the segmentation is done on the basis of a training data set, which encodes knowledge about the structure of the groups in the form of a target variable. We need to provide training data to the classifier, for it to understand how the given input variables are related to the class. And once the classifier is trained accurately, it is used to predict the specific class of the given data set. Predictive accuracy, computational speed, robustness, scalability, and interpretability are the major criterias for the evaluation of classification.

Classification learners are broadly classified into two types namely : lazy learners and eager learners. Lazy learners or instance based classification, which store all the training data in pattern space and wait until a testing data appears before performing the generalisation. K Nearest-Neighbor, Local Regression etc are few examples of lazy classifiers. In contrast, Eager learners construct a classification model based on the given training data before receiving data for classification. Decision tree classifiers, Bayesian classifiers, classification by backpropagation etc are few examples of eager learners. There are also other classification methods such as genetic algorithms, rough sets and fuzzy logic techniques, which have a wide range of applications in the industry.

There are several types of classification algorithms that we can use depending on the application and nature of the dataset that we are working with. The most common algorithms in machine learning are Decision Tree, Logistic Regression, Naive Bayes, KNN, Support Vector Machine, Neural Network, Random Forest etc. Here, we mainly focus on Logistic Regression, Random Forest, Stochastic Gradient Descent classification algorithms and their applications, as the rest of algorithms are already discussed in the lecture.

- **Logistic Regression :**

  It is an approach for predicting the outcome of a categorical dependent variable based on one or more observed variables. The probabilities describing the possible outcomes are modeled as a function of the observed variables using a logistic function. The goal of logistic regression is to find a best-fitting relationship between the dependent variable and a set of independent variables. It is better than other binary classification algorithms like KNN as it quantitatively explains the factors leading to classification.

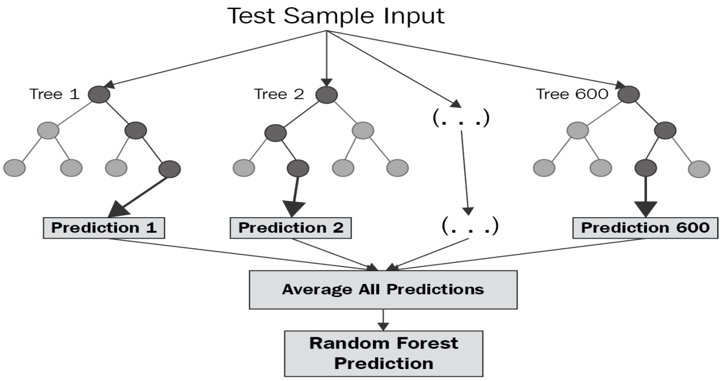
The Logistic regression expression is given by,

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x).

- **Random Forest :**

Random Forest classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation. RF Classifier makes use of sets of decision trees on splits with randomly generated vectors and computes the score as a function of these different components. RF Classifiers are created by either random split selection / input selection. It is a meta-estimator that fits a set of trees on various samples of data, it then uses an average to improve the accuracy in the model's predictive nature.



- **Stochastic Gradient Descent (SGD) :**

Stochastic gradient descent is a very efficient and simple approach to fit the linear dataset models. This refers to calculating the derivative from each training data instance and calculating the update immediately. It is mainly useful when the number of samples is large in number. It supports different loss functions for classification.

The pseudo code for SGD looks like below :

```
$for\ i\ in\ range\ (m):$
    $\theta_j = \theta_j - \alpha\,(\hat{y}^i - y^i)\,x_j^i$
```

where x^i refers to each training example and y^i refers to its label.

The main advantage of this model is its efficiency and ease in implementation. But on the other side, it also requires a number of hyper-parameters and it is also sensitive to feature scaling. It is majorly used in “Internet Of Things”, updating parameters such as weights in neural networks or coefficients in linear regression.

Applications of Classification Models :

- **Customer Target Marketing :** Classification is extremely popular for addressing the problem of customer target marketing. We can use feature variables describing the customer, to predict their buying interests.
- **Medical Disease Diagnosis :** Data Mining methods in the medical field have gained increasing transaction in recent years. Based on the existing data set, Classification models can be used to predict whether a patient may pick up a disease in the future or not.
- **Supervised Event Detection :** Time - series classification methods are very useful in supervised event detection. To detect time stamps of any intrusion activity, to detect unusual events in image sequences etc are some of the applications of supervised event detection.
- **Multimedia Data Analysis :** Multimedia data is a combination of different discrete text, audio, images, videos, animations data etc. Proper analysis of multimedia files using machine learning classification techniques have wide applications in medical diagnosis, video surveillance, text annotation etc.
- **Biological Data Analysis :** Classification methods can be applied for a wide variety of biological data. Gene detection, analysis of mutations in cancer are some of the applications of biological data analysis.
- **Document Categorization and Filtering :** Many applications, such as newswire services, require the classification of large numbers of documents in real time. This application is referred to as document categorization, and is an important area of research in the industry.
- **Social Network Analysis :** Social Network Analysis is the study of social networks to understand their structure and behavior. Data mining based techniques are proving to be useful for analysis of social network data, especially for large datasets that cannot be handled by traditional methods. It has a wide range of applications from product marketing (e.g. viral marketing) to search engines and organizational dynamics (e.g. management).

References :

- https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529
- https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623
- https://ieeexplore.ieee.org/document/6726842?reload=true
- https://www.edureka.co/blog/classification-in-machine-learning/#gradient
- http://www.informatica.si/index.php/informatica/article/viewFile/148/140
- https://machinelearningmastery.com/types-of-classification-in-machine-learning/

E3 NLP methods and Applications

Motivation:

Natural language processing deals with the interaction of computers with human languages. It particularly deals about how to program computers to process and analyze large amounts of natural language data. We see NLP techniques are used in our everyday life, for example: Siri, Google Assistant, Alexa. The evolution of Machine Learning techniques and availability of needed computation and training data, over the past decade, fostered the research in NLP. The main motivation for selecting this topic is to learn about the implementation of NLP techniques such as Sentiment Analysis, Named Entity Recognition (NER), Text Summarization, Aspect Mining. And also to learn further about how NLP uses supervised and unsupervised learning to perform these tasks.

Essay:

Real-world Natural Language Processing (NLP) problems use Statistical Methods and Machine Learning for performing tasks such as speech recognition, syntactic parsing, parts of speech tagging etc. The Commonly used Machine Learning models/techniques in NLP are: Hidden Markov Models, Maximum Entropy Models, Conditional Random Fields, Clustering techniques, Expectation-Maximization algorithm, Support Vector Machines and Active Learning.

These methods are applied to real world NLP problems such as stochastic parsing, information extraction, text segmentation and classification, word sense disambiguation and topic/document clustering. Inference algorithms such as Viterbi, Synchronous Chart Parsing and Beam Search are used as well for some of NLP applications.Some of the NLP methods used in various NLP techniques are discussed below.

NLP Methods:

- **n-gram Models:**

  Smoothed n-gram models are used for Language modeling for automatic speech recognition. N - gram models find the most probable string of words w1; : : :; wn from a set of candidate strings which are compatible with the acoustic data.

  The prediction of an n-gram model is based x_i on , in $x_{i-(n-1)}, \dots, x_{i-1}$ probability terms it is:

  $$P(x_i \mid x_{i-(n-1)}, \dots, x_{i-1})$$

  During the language modelling process an assumption that each word depends only on the last *n* − 1 words is made. This Markov model helps to make an approximation of the true underlying language. And it helps to simplify the problem of estimating the language model from data. The probability of a word, conditioned on previous words can be described as following a categorical distribution. These probability distributions are smoothed by assigning non-zero probabilities to unseen words or *n*-grams using smoothing techniques.

- **Hidden Markov Models:**

  Hidden Markov Model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable or hidden states. Hidden Markov models are used for Part-of-speech tagging to find the most probable tag sequence t1; : : : tn given a word sequence w1; : : : wn.

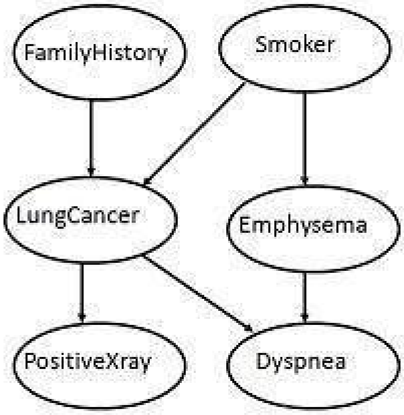
  Let Xn and Yn be discrete-time stochastic processes and n>=1. The pair (Xn , Yn) is said to be hidden markov model if:

  Xn is a Markov process whose behavior is not directly observable ("hidden").

$$\mathbf{P(}\ Yn\epsilon\ A \mid X1 = x1 ,....., Xn = xn\mathbf{)} = \mathbf{P(}\ Yn\epsilon\ A \mid Xn = xn\mathbf{)}$$

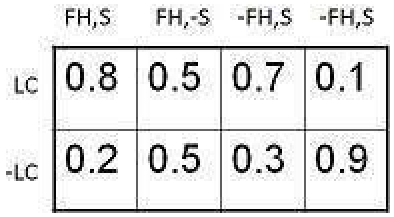
- **Bayesian Classifiers:**
  Bayesian classifiers are the statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classifiers are used to find the most probable sense s for word w in context C (word sense disambiguation).

  Example for Bayesian classification:



:

The conditional probability table for the above graph will be:



	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

- **Probabilistic Models:**
  Probabilistic models are used to find the most probable target language sentence t for a given source language sentence, This process is also called as Machine translation. In these probabilistic models the parameters are derived from the analysis of bilingual text. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation.

- **Probabilistic Grammars:**
  Syntactic parsing is performed using probabilistic grammars by finding the most probable parse tree T given a word sequence w1; : : : ; wn (or tag sequence t1; : : : ; tn). Probabilistic parsing uses dynamic programming algorithms to compute the most likely parse(s) of a given sentence, given a statistical model of the syntactic structure of a language.

Applications of NLP:

- **Sentiment Analysis:** Sentiment analysis is a natural language processing technique which is used to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials. A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. In Advanced cases, emotional states such as enjoyment, anger, disgust, sadness, fear, and surprise can be classified as well.

- **Chatbots & Virtual Assistants:** Chatbots and virtual assistants are used for automatic question answering, these are designed to understand natural language and deliver an appropriate response through natural language generation. These intelligent machines are increasingly present at the frontline of customer support, as they can help teams solve up to 80% of all routine queries and route more complex issues to human agents.

- **Named Entity Recognition (NER):** Named Entity Recognition automatically detects specific information in a text, such as names, companies, places, and more. We can use text extraction for data entry. We can get the information needed and set up an alert to automatically enter this information in the database. Applications of NER also include sifting through incoming support tickets and identifying specific data, like company names, order numbers, and email addresses without needing to open and read every ticket.

- **Machine Translation:** The widely known machine translation tool is Google Translate. This kind of automated translation is useful in many businesses because it facilitates communication and allows companies to reach broader audiences, and also understand foreign documentation in a fast and cost-effective way.

- **Text Summarization:** This technique extracts the most important information from the given text and summarizes it.

- **Auto-Correct:** NLP techniques are used as grammar checking software using auto-correct functions. They detect errors from spelling, grammar, and sentence structure.

- **Speech Recognition:** Speech Recognition is a widely used NLP technique which recognizes human speech and converts it to a written format. Using this technique many actions like web search, setting up events and authenticating a user can be performed.

References:

https://en.wikipedia.org/wiki/N-gram
https://en.wikipedia.org/wiki/Hidden_Markov_model
https://en.wikipedia.org/wiki/Statistical_machine_translation
https://cl.lingfil.uu.se/~nivre/docs/statnlp.pdf
https://nlp.stanford.edu/projects/stat-parsing.shtml