

Cluster Analysis, ANN and Text Mining Project

Start Assignment

Due Dec 1 by 11:59pm **Points** 100 **Submitting** a file upload
Available Oct 27 at 12am - Dec 15 at 11:59pm

Dear Students,

This Assignment4 consists of ANN based classification, cluster analysis and text mining.

The assignment instructions are in:

labs-->data--->project4_clustering_TextMining_ANN folder. I have uploaded the instructions file in both PDF and Word.

Since the referenced dataset is not available please use the following IMDB Dataset:


"imdb_dataset.csv" in files----->labs---->data

For clustering using TFIDF for sentiment analysis this may perhaps be a useful relevant article with python code. You can partition the data for training and validation and do k-means with TFIDF by removing labels (unsupervised) and testing with test partition and compare with true labels in test data:

Additional knowledge reference on K-means clustering by Daniel Foley:

<https://link.medium.com/Ezc6zqqcW5>

Please read this article:

[6.2. Feature extraction — scikit-learn 1.0.2 documentation](https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction)  (https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction)

Summary of article above:

1. find all unique useful words in all documents
2. For each document find the count vector for all the words
3. Then find the tf-idf vector by using inverse document formulae. This gives a relative score of frequencies.
4. Then within the vector normalize data to lie between 0 and 1.
5. Each numerical vector represents a document. Use this numerical vector to find the distance between vectors of documents in k-means.

6. This will give you clusters (groups) of documents. The groups discovered in an unsupervised manner, may now automatically through ML, contain related documents such as biology or chemistry and so on depending on the context!

Since ANN concerns classification topic from last assignment, you may use any dataset for comparative analysis with other classification models.

Further for ANN, make sure you follow the following data pre-processing steps.

IMPORTANT:

Preprocessing the Data (reference: Data Analytics for Business Analytics, by Galit Shmueli et al)

BEGIN of pre-processing instructions for ANN:

When using a logistic activation function (option activation = 'logistic' in scikit-learn), neural networks perform best when the predictors and outcome variable are on a scale of [0, 1]. For this reason, all variables should be scaled to a [0, 1] interval before entering them into the network. For a numerical variable X that takes values in the range $[a, b]$ where $a < b$, we normalize the measurements by subtracting a and dividing by $b - a$. The normalized measurement is then: $X_{\text{norm}} = (X-a)/(b-a)$.

Note that if $[a, b]$ is within the $[0, 1]$ interval, the original scale will be stretched.

If a and b are unknown, we can estimate them from the minimal and maximal values of X in the data. Even if new data exceed this range by a small amount, yielding normalized values slightly lower than 0 or larger than 1, this will not affect the results much.

For binary variables, no adjustment is needed other than creating dummy variables. For categorical variables with m categories, if they are ordinal in nature, a choice of m fractions in $[0, 1]$ should reflect their perceived ordering. For example, if four ordinal categories are equally distant from each other, we can map them to $[0, 0.25, 0.5, 1]$. If the categories are nominal, transforming into $m - 1$ dummies is a good solution.

Another operation that improves the performance of the network is to transform highly skewed predictors. In business applications, there tend to be many highly right-skewed variables (such as income). Taking a log transform of a right-skewed variable (before converting to a $[0, 1]$ scale) will usually spread out the values more symmetrically.

Another common sigmoidal function is the hyperbolic tangent (option activation = 'tanh' in scikit-learn). When using this function, it is usually better to scale predictors to a $[-1, 1]$ scale.

END of data pre-processing instructions for ANN.

Create a Report on all three areas.

The Rubric for this team project is:

1. Clustering:

15% K-Means (5% plot SSE vs # of clusters 10% K-means algorithm)

15% Hierarchical (single, complete and average link (9%), plot (6%))

Subtotal: (30%)

2. Text Mining:

Create count vector and tf-idf vector (normalized vector) 20%

Explain usage 10%

Subtotal (30%)

3. ANN:

Attribute value Binarization 10%

ANN 10%

Accuracy comparison with other classification models 10%

Subtotal (30%)

4. Report with summarized findings (please submit in PDF document format): 10%

Total: 100%

5. Data Preprocessing for ANN: BONUS points 5%

You may use code from tutorials 6, 7 and 8 and write any additional code to accomplish the above goals.

Cheers,

:)

Jagan

Assignment 4

Criteria	Ratings	Pts
Clustering K-Means and Hierarchical		30 pts
Text Mining Count Vector and TF-IDF vector and usage of text mining		30 pts
ANN		30 pts
Report		10 pts
Data Preprocessing for ANN (BONUS Points) Data Preprocessing for ANN (BONUS Points) following the instructions for data preprocessing for ANN as suitable to your application.		5 pts
Total Points: 105		