CS 396 Introduction to the Data Science Pipeline

HW4 – Unsupervised Machine Learning

Q1. Feature Engineering. (15pt)

In this question, we will work on the forest cover type dataset. Information about the dataset can be found at: https://archive.ics.uci.edu/ml/datasets/covertype , and you can obtain the data using the function at

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_covtype.html.

Step 1. We will focus on a binary classifier in this task, so we need to obtain a binary target. Please obtain a binary target value such that the target value is one if and only if the original target value is the most frequent one. For example, if the original target value is [1, 2, 3, 4, 2, 5, 2], the new target value should be [0, 1, 0, 0, 1, 0, 1]

Step 2. Please evaluate the performance of LogisticRegression (for this assignment, please use `saga` as the solver) and DecisionTree classifier using 3-fold cross-validation and ROC-AUC scoring under default parameters. Please compare the performance under the following feature preprocessing:

1) no preprocessing

2) Min-max scaling

Now we get four roc_auc scores. Which model benefits from the scaling? Explain why it is the case.

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing

Q2. Dimensionality Reduction. (30pt)

Following on the previous example. Now compare the performance of logistic regression with the default parameter using 3-fold CV and apply the following dimensionality reduction:

A. perform standard scaling and then Gaussian projection to 15 dimensions

B. perform standard scaling and then Gaussian projection to 40 dimensions

C. perform standard scaling and then PCA to 10 dimensions

D. perform PCA to 10 dimensions and then perform standard scaling.

1) Compare the performance between A, B, and C and explain your observation. (10pt)

2) Compare the performance between C and D and explain your observation. (10pt)

3) We discussed "knee method" during clustering. Use the spirit of this idea to figure out the minimum inherent dimension for this task. (10pt)

Q3. Feature Engineering via TF-IDF. (20pt)

Go to https://www.yelp.com/dataset and download the yelp dataset.

In this task, we would like to use TF-IDF to extract some features from **yelp_academic_dataset_tip.json** and use the content of the tip to predict the type of the business. Follow these three steps to accomplish the task

a. Obtain all business information in Florida **(from yelp_academic_dataset_business.json)** and then add a Boolean column to indicate if this business is a restaurant or not. You can do so by checking if 'Restaurant' appears in the "categories" field. You can drop all businesses without the "categories" attribute. Join this table with the **yelp_academic_dataset_tip.json**. Now you have a table where each row contains: 1) the text of a tip, 2) a Boolean indicating if this review is for a restaurant or not (we will call it "is_restaurant").

b. Use the TFIDF to extract a set of features using default arguments for each tip. Now we want to use logistic regression to build a classification model to predict the is_restaurant attribute using the TFIDF features. We want to use the 3-fold cross-validation method and ROC_AUC as the metric to test the performance. Note that TFIDF produces a sparse matrix but can still be used directly in sklearn. Please print out the mean score of the cross-validation.

Q4. Recommender System. (35pt)

In this question, we will use the MovieLens dataset, which contains user ratings of movies. The dataset is hosted at https://grouplens.org/datasets/movielens/ and we will use the smallest dataset named ml-latest-small.zip under "MovieLens Latest Datasets".

Task 4a (15pt)
Step 1. Understand the format of the dataset and parse it. Randomly reserve 20% of the rating as testing data and build a sparse matrix that contains the rest of the ratings. The value at the i-th row and j-th column should be the rating from user i to movie j.

Step 2. Use non-negative matrix factorization to factor the matrix into two latent factor matrices, namely W and H. The dimension of the latent vector can be set initially to 50. What's the predicted rating for user 1 to movie 2?

Step 3. We would like to recommend the movie with id 1 to other users. Find three users who have not watched this movie but are most likely to watch it.

Task 4b (15pt)

Compute the RMSE score of the model you get when tested on the training set and when tested on the testing set, respectively. RMSE is defined as below. All notations are the same as in the lecture slides except $\hat{v}_{ui}$ is the predicted rating.

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in\mathcal{Z}}(v_{ui} - \hat{v}_{ui})^2}{|\mathcal{Z}|}}$$

Now we will use the knee method to find what would be a good value for the dimension of the latent vector. Run the default NMF but set the dimension increasingly from 1 to 20. Plot a graph how the RMSE score changes as the dimension increases. What value would you use?

Task 4c (5pt)

What would you do to further improve the effectiveness of this recommender system? List two directions.

Related documentation:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

Submission Instructions

Put your code in the template notebook as provided (HW4.ipynb). Please submit two files:

1) The ipynb file. Make sure every question is executed with the printed results. Make sure all explanations are printed.

2) Export your code and results in a separate html document. After you are done with the ipynb file, go to File --> Download/Export As --> HTML in Jupyter notebook.

Submit **two documents** using the following naming convention. Please submit them directly **without zipping them.**

```
HW4.ipynb
HW4.html
```